Robustly Learning a Single Neuron via Sharpness

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

Abstract

We study the problem of learning a single neuron with respect to the L_2^2-loss in the presence of adversarial label noise. We give an efficient algorithm that, for a broad family of activations including ReLUs, approximates the optimal L_2^2-error within a constant factor. Notably, our algorithm succeeds under much milder distributional assumptions compared to prior work. The key ingredient enabling our results is a novel connection to local error bounds from optimization theory.

1. Introduction

We study the following learning task: Given labeled examples $(\mathbf{x}, y) \in \mathbb{R}^d \times \mathbb{R}$ from an unknown distribution $\mathcal{D}$, output the best-fitting ReLU (or other nonlinear function) with respect to square loss. This is a fundamental problem in machine learning that has been extensively studied in a number of interrelated contexts over the past two decades, including learning GLMs and neural networks. More specifically, letting $\sigma : \mathbb{R} \mapsto \mathbb{R}$ denote a nonlinear activation, e.g., $\sigma(t) = \mathrm{ReLU}(t) = \max\{0, t\}$, the (population) square loss of a vector $\mathbf{w}$ is defined as the L_2^2 loss of the hypothesis $\sigma(\mathbf{w} \cdot \mathbf{x})$, i.e., $\mathcal{L}_2^{\mathcal{D},\sigma}(\mathbf{w}) \triangleq \mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[(\sigma(\mathbf{w} \cdot \mathbf{x}) - y)^2]$. Our learning problem is then formally defined as follows.

Problem 1.1 (Robustly Learning a Single Neuron)**.** Fix $\epsilon > 0, W > 0$, and a class of distributions $\mathcal{G}$ on $\mathbb{R}^d$. Let $\sigma : \mathbb{R} \mapsto \mathbb{R}$ be an activation and $\mathcal{D}$ a distribution on labeled examples $(\mathbf{x}, y) \in \mathbb{R}^d \times \mathbb{R}$ such that its $\mathbf{x}$-marginal $\mathcal{D}_{\mathbf{x}}$ belongs to $\mathcal{G}$. For some $C \geq 1$, a C-approximate proper learner is given ϵ, W and i.i.d. samples from $\mathcal{D}$ and outputs $\hat{\mathbf{w}} \in \mathbb{R}^d$ such that with high probability it holds $\mathcal{L}_2^{\mathcal{D},\sigma}(\hat{\mathbf{w}}) \leq C\,\mathrm{OPT} + \epsilon$, where $\mathrm{OPT} \triangleq \min_{\|\mathbf{w}\|_2 \leq W} \mathcal{L}_2^{\mathcal{D},\sigma}(\mathbf{w})$ is the minimum attainable square loss. We use $\mathcal{W}^* \triangleq \mathrm{argmin}_{\|\mathbf{w}\|_2 \leq W} \mathcal{L}_2^{\mathcal{D},\sigma}(\mathbf{w})$ to denote the set of square loss minimizers.

Problem 1.1 does not make realizability assumptions on the distribution $\mathcal{D}$. The labels are allowed to be arbitrary and we are interested in the best-fit function with respect to the L_2^2 error. This corresponds to the (distribution-specific) agnostic PAC learning model (Haussler, 1992; Kearns et al., 1994). In this paper, we focus on developing *constant factor* approximate learners, corresponding to the case that C is a universal constant greater than one.

The special case of Problem 1.1 where the labels are consistent with a function in $\mathcal{H} = \{\sigma(\mathbf{w} \cdot \mathbf{x}) : \|\mathbf{w}\|_2 \leq W\}$ was studied in early work (Kalai & Sastry, 2009; Kakade et al., 2011). These papers gave efficient methods that succeed for any distribution on the unit ball and any monotone Lipschitz activation[1]. More recently, (Yehudai & Shamir, 2020) showed that gradient descent on the nonconvex L_2^2 loss succeeds under a natural class of distributions (again in the *realizable* case) but fails in general. In other related work, (Soltanolkotabi, 2017) analyzed the case of ReLUs in the realizable setting under the Gaussian distribution and showed that gradient descent efficiently achieves exact recovery.

The agnostic setting is computationally challenging. First, even for the case that the marginal distribution on the examples is Gaussian, there is strong evidence that any algorithm achieving error $\mathrm{OPT} + \epsilon$ (corresponding to $C = 1$ in Problem 1.1) requires $d^{\mathrm{poly}(1/\epsilon)}$ time (Goel et al., 2019; Diakonikolas et al., 2020b; Goel et al., 2020; Diakonikolas et al., 2021). Second, even if we relax our goal to constant factor approximations, some distributional assumptions are required: known NP-hardness results rule out proper learners achieving *any* constant factor (Síma, 2002; Manurangsi & Reichman, 2018). More recent work (Diakonikolas et al., 2022a) has shown that no polynomial time constant factor *improper* learner exists (under cryptographic assumptions), even if the distribution is bounded. These intractability results motivate the design of constant factor approximate learners — corresponding to $C > 1$ and $C = O(1)$ — that succeed under as mild distributional assumptions as possible.

Prior algorithmic work in the robust setting can be classified in two categories: A line of work (Frei et al., 2020; Diakonikolas et al., 2022b; Awasthi et al., 2022) analyzes

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

Proceedings of the 40th International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023. Copyright 2023 by the author(s).

[1]The results in these works can tolerate zero mean random noise, but do not apply to the adversarial noise setting.

gradient descent-based algorithms on the natural nonconvex L_2^2 objective (possibly with regularization). These works show that *under certain distributional assumptions* gradient descent avoids poor local minima and converges to a good solution. Specifically, (Diakonikolas et al., 2022b) established that gradient descent efficiently converges to a constant factor approximation for a family of well-behaved continuous distributions (including logconcave distributions). The second line of work (Diakonikolas et al., 2020a) proceeds by convexifying the problem, namely constructing a *convex surrogate* whose optimal solution gives a good solution to the initial nonconvex problem. This convex surrogate was analyzed in (Diakonikolas et al., 2020a) for the case of ReLUs and showed that it yields a constant factor approximation for logconcave distributions.

The starting point of our investigation is the observation that *all previous algorithmic works for Problem 1.1 impose fairly stringent distributional assumptions.* These works require all of the following properties from the marginal distribution on examples: (i) anti-concentration, (ii) concentration, and (iii) anti-anti-concentration. Assumption (i) posits that that *every* one-dimensional (or, in some cases, constant-dimensional) projection of the points should not put too much mass in any interval (or "rectangle"). Property (ii) means that every one-dimensional projection should be strongly concentrated around its mean; specifically, prior work required at least exponential concentration. Finally, (iii) requires that the density of every low-dimensional projection is bounded below by a positive constant.

While some concentration appears necessary, prior work required sub-exponential concentration, which rules out the important case of heavy-tailed data. The anticoncentration assumption (i) from prior work rules out possibly lower-dimensional data, while the anti-anti-concentration rules out discrete distributions, which naturally occur in practice.

The preceding discussion raises the following question:

Under what distributional assumptions can we obtain efficient constant factor *learners for Problem 1.1?*

In this paper, we give such an algorithm that succeeds under minimal distributional assumptions. Roughly speaking, our novel assumptions require anti-concentration *only* in the direction of the optimal solution (aka a margin assumption) and allow for heavy-tailed data. Moreover, by removing assumption (iii) altogether, we obtain the first positive results for structured discrete distributions (including, e.g., discrete Gaussians and the uniform distribution over the cube).

In addition to its generality, our algorithm is simple — a mini-batch SGD — and achieves significantly better sample complexity for distributions covered in prior work.

1.1. Overview of Results

We provide a simplified version of our distributional assumptions followed by our main result for ReLU activations.

Distributional Assumptions. We make only the following two distributional assumptions.

Margin-like Condition: There exists $\mathbf{w}^* \in \mathcal{W}^*$ and constants $\gamma, \lambda > 0$ such that

$$\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}} \left[\mathbf{x}\mathbf{x}^\top \mathbb{1}\left\{\mathbf{w}^* \cdot \mathbf{x} \geq \gamma \|\mathbf{w}^*\|_2\right\}\right] \succeq \lambda \mathbf{I} \,. \qquad (1)$$

Concentration: There exists non-increasing $h : \mathbb{R}_+ \to \mathbb{R}_+$ satisfying $h(r) = O(r^{-5})$ such that for any unit vector $\mathbf{u}$ and any $r \geq 1$, it holds $\mathbf{Pr}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[|\mathbf{u}\cdot\mathbf{x}| \geq r] \leq h(r)$.

Before we state our algorithmic result, some comments are in order. Condition (1) is an anti-concentration condition, reminiscent of the classical margin condition for halfspaces. In comparison with prior work, our condition requires anti-concentration *only* in the direction of an optimal solution — as opposed to every direction. Our second condition requires that every univariate projection exhibits some concentration. Our concentration function h can even be inverse polynomial, allowing for heavy-tailed data. In contrast, prior work only considered sub-exponential tails. As we will see, the function h affects the sample complexity of our algorithm.

As we show in Section E of the supplementary material, our distributional assumptions subsume all previous such assumptions considered in the literature and additionally include a range of distributions (including heavy-tailed and discrete distributions) not handled in prior work.

A simplified version of our main result for the special case of ReLU activations is as follows (see Theorem 3.3 for a detailed more general statement):

Theorem 1.2 (Main Algorithmic Result, Informal)**.** *Let $W = O(1)$, $\mathcal{G}$ be a class of marginal distributions satisfying the above distributinal assumptions, and σ be the ReLU activation. There exists a sample-efficient and sample-linear time algorithm that outputs a hypothesis $\hat{\mathbf{w}}$ such that, with high probability, $\mathcal{L}_2^{\mathcal{D},\sigma}(\hat{\mathbf{w}}) = O(\mathrm{OPT}) + \epsilon$. In particular, if the tail function h is subexponential, namely $h(r) = e^{-\Omega(r)}$, then the algorithm has sample complexity $n = \tilde{O}(d\,\mathrm{polylog}(1/\epsilon))$. For heavy-tailed distributions, namely for $h(r) = O(r^{-k})$ for some $k > 4$, the algorithm has sample complexity $n = \tilde{O}(d\,(1/\epsilon)^{2/(k-4)})$. The algorithm's runtime is always $O(nd)$.*

Our algorithm is extremely simple: it amounts to mini-batch SGD on a natural convex surrogate of the problem. As we will explain subsequently, this convex surrogate has been studied before in closely related — yet more restricted — contexts. Our main technical contribution lies in the analysis, which hinges on a new connection to local error bounds

from the theory of optimization. This connection is crucial for us in two ways: First, we leverage it to obtain the first constant-factor approximate learners under much weaker distributional assumptions. Second, even for distributions covered by prior work, the connection allows us to obtain significantly more efficient algorithms.

Finally, we note that our algorithmic result applies to a broad family of monotone activations (Definition 2.1 and Assumption 2.2), and can be adapted to handle non-monotone activations — including GeLU (Hendrycks & Gimpel, 2016) and Swish (Ramachandran et al., 2017) — see Appendix F.

1.2. Technical Contributions

The main algorithmic difficulty in solving Problem 1.1 is its non-convexity. Indeed, the L_2^2 loss is non-convex for nonlinear activations, even without noise. Of course, the presence of adversarial label noise only makes the problem even more challenging. At a high-level, our approach is to convexify the problem via an appropriate *convex surrogate* function (see, e.g., (Bartlett et al., 2006)). In more detail, given a distribution $\mathcal{D}$ on labeled examples $(\mathbf{x}, y)$ and an activation σ, the surrogate $\mathcal{L}_{\mathrm{sur}}^{\mathcal{D},\sigma}(\mathbf{w})$ is defined by $\mathcal{L}_{\mathrm{sur}}^{\mathcal{D},\sigma}(\mathbf{w}) = \mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}\left[\int_0^{\mathbf{w}\cdot\mathbf{x}}(\sigma(r)-y)\,\mathrm{d}r\right]$.

This function is not new. It was first defined in (Auer et al., 1995) and subsequently (implicitly) used in (Kalai & Sastry, 2009; Kakade et al., 2011) for learning GLMs with zero mean noise. More recently, (Diakonikolas et al., 2020a) used this convex surrogate for robustly learning ReLUs under logconcave distributions. Roughly speaking, they showed that – under the logconcavity assumption – a near-optimal solution to the (convex) optimization problem of minimizing $\mathcal{L}_{\mathrm{sur}}^{\mathcal{D},\sigma}(\mathbf{w})$ yields a constant factor approximate learner for Problem 1.1 (for the special case of ReLU activations).

Very roughly speaking, our high-level approach is similar to that of (Diakonikolas et al., 2020a). The main novelty of our contributions lies in two aspects: (1) The *generality* of the distributional assumptions under which we obtain a constant-factor approximation, and (2) the sample and computational complexities of the associated algorithm. Specifically, our analysis yields a constant-factor approximate learner under a vastly more general class of distributions[2] as compared to prior work, and extends to a much broader family of activations beyond ReLUs. Moreover, even if restrict ourselves to, e.g., logconcave distributions, the complexity of our algorithm is exponentially smaller as a function of ϵ — namely, $\mathrm{polylog}(1/\epsilon)$ as opposed to $\Omega(1/\epsilon^2)$. For a more detailed comparison, see Appendix B.

The key technical ingredient enabling our results is the notion of *sharpness* (local error bound) from optimization theory, which we prove holds for our stochastic surrogate problem. Before explaining how this comes up in our setting, we provide an overview from an optimization perspective.

Local Error Bounds and Sharpness. Broadly speaking, given an optimization problem (P) and a "residual" function r that is a measure of error of a candidate solution $\mathbf{w}$ to (P), an error bound certifies that a small residual translates into closeness between the candidate solution and the set of "test" (typically optimal) solutions $\mathcal{W}^*$ to (P). In particular, an error bound certifies an inequality of the form

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

for some parameters $\mu, \nu > 0$, where $\mathrm{dist}(\mathbf{w}, \mathcal{W}^*) = \min_{\mathbf{w}^*\in\mathcal{W}^*}\|\mathbf{w}-\mathbf{w}^*\|_2$ (see, e.g., the survey (Pang, 1997)). When this bound holds *only locally* in some neighborhood of $\mathcal{W}^*$, it is referred to as a *local error bound*.

Local error bounds are well-studied within optimization theory, with the earliest result in this area being attributed to (Hoffman, 1952), which provided local error bounds for systems of linear inequalities. The work of (Hoffman, 1952) was extended to many other optimization problems; 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; Liu et al., 2022) and references therein for a more cotemporary overview. One of the most surprising early results in this area states that for minimizing a convex function f, an inequality of the form

$$f(\mathbf{w}) - \min_{\mathbf{u}} f(\mathbf{u}) \geq (\mu/\nu)\,\mathrm{dist}(\mathbf{w}, \mathcal{W}^*)^\nu \qquad (2)$$

holds generically whenever f is a real analytic or subanalytic function (Łojasiewicz, 1963; 1993). The main downside of this result is that the parameters μ, ν are usually impossible to evaluate and, moreover, even when it is known that, e.g., $\nu = 2$, the parameter μ can be exponentially small in the dimension. Furthermore, local error bounds have primarily been studied in the context of *deterministic* optimization problems, with results for stochastic problems being very rare (Chen & Fukushima, 2005; Liu et al., 2018).

Perhaps the most surprising aspect of our results is that we show that the (stochastic) convex surrogate minimization problem not only satisfies a local error bound (a relaxation of (2) and a much weaker property than strong convexity; see Appendix A) with $\nu = 2$, but we are also able to characterize the parameter μ based on the assumptions about the activation function and the probability distribution over the data. More importantly, for standard activation functions such as ReLU, Swish, and GeLU and for broad classes of distributions (including heavy-tailed and discrete ones), we prove that μ is an *absolute constant*. This is precisely what leads to the error and complexity results achieved in our work.

[2]Recall that without distributional assumptions obtaining *any* constant-factor approximate learner is NP-hard.

Robustly Learning a Single Neuron via Sharpness. Our technical approach can be broken down into the following main ideas. As a surrogate for minimizing the square loss, we first consider the *noise-free* convex surrogate, defined by $\bar{\mathcal{L}}_{\mathrm{sur}}^{\mathcal{D},\sigma}(\mathbf{w};\mathbf{w}^*) = \mathbf{E}_{\mathbf{x}\sim\mathcal{D}_{\mathbf{x}}}\left[\int_0^{\mathbf{w}\cdot\mathbf{x}}(\sigma(r)-\sigma(\mathbf{w}^*\cdot\mathbf{x}))\,\mathrm{d}r\right]$, where $\mathbf{w}^* \in \mathcal{W}^*$ is a square-loss minimizer that satisfies our margin assumption. We keep this $\mathbf{w}^*$ fixed throughout the analysis and simply write $\bar{\mathcal{L}}_{\mathrm{sur}}^{\mathcal{D},\sigma}(\mathbf{w})$ instead of $\bar{\mathcal{L}}_{\mathrm{sur}}^{\mathcal{D},\sigma}(\mathbf{w};\mathbf{w}^*)$. Compared to the convex surrogate $\mathcal{L}_{\mathrm{sur}}^{\mathcal{D},\sigma}(\mathbf{w})$ introduced earlier in the introduction, the noise-free convex surrogate $\bar{\mathcal{L}}_{\mathrm{sur}}^{\mathcal{D},\sigma}(\mathbf{w};\mathbf{w}^*)$ replaces the noisy labels y with $\sigma(\mathbf{w}^*\cdot\mathbf{x})$. Clearly, $\bar{\mathcal{L}}_{\mathrm{sur}}^{\mathcal{D},\sigma}(\mathbf{w};\mathbf{w}^*)$ is a function that cannot be directly optimized, as we lack the knowledge of $\mathbf{w}^*$. On the other hand, the noise-free surrogate relates more directly to the square loss minimization: we prove (Lemma 2.5) that our distributional assumptions suffice for the noise-free surrogate to be sharp on a ball of radius $2\|\mathbf{w}^*\|_2$, $\mathcal{B}(2\|\mathbf{w}^*\|_2)$; this structural result in turn leads to the conclusion that $\mathbf{w}^*$ is its unique minimizer. Hence, we can conclude that minimizing the noise-free surrogate $\bar{\mathcal{L}}_{\mathrm{sur}}^{\mathcal{D},\sigma}(\mathbf{w};\mathbf{w}^*)$ leads to minimizing the L_2^2 loss. Of course, we cannot directly minimize $\bar{\mathcal{L}}_{\mathrm{sur}}^{\mathcal{D},\sigma}(\mathbf{w};\mathbf{w}^*)$, as we do not know $\mathbf{w}^*$.

Had there been no adversarial label noise, we could stop at this conclusion, as there would be no difference between $\mathcal{L}_{\mathrm{sur}}^{\mathcal{D},\sigma}(\mathbf{w})$ and $\bar{\mathcal{L}}_{\mathrm{sur}}^{\mathcal{D},\sigma}(\mathbf{w};\mathbf{w}^*)$ and we could minimize the L_2^2 error to any desired accuracy by minimizing $\mathcal{L}_{\mathrm{sur}}^{\mathcal{D},\sigma}(\mathbf{w})$. This difference between $\mathcal{L}_{\mathrm{sur}}^{\mathcal{D},\sigma}(\mathbf{w})$ and $\bar{\mathcal{L}}_{\mathrm{sur}}^{\mathcal{D},\sigma}(\mathbf{w};\mathbf{w}^*)$ is precisely what causes the L_2^2 error to only be brought down to $O(\mathrm{OPT})+\epsilon$, where the constant in the big-Oh notation depends on the sharpness parameter μ. On the technical side, we prove (Proposition 3.2) that $\mathcal{L}_{\mathrm{sur}}^{\mathcal{D},\sigma}(\mathbf{w})$ is also sharp w.r.t. the same $\mathbf{w}^*$ as $\bar{\mathcal{L}}_{\mathrm{sur}}^{\mathcal{D},\sigma}(\mathbf{w};\mathbf{w}^*)$ and with the sharpness parameter μ of the same order, but *only on a nonconvex subset* of the ball $\mathcal{B}(2\|\mathbf{w}^*\|_2)$, which excludes a neighborhood of $\mathbf{w}^*$. This turns out to be sufficient to relate minimizing $\mathcal{L}_{\mathrm{sur}}^{\mathcal{D},\sigma}(\mathbf{w})$ to minimizing the L_2^2 loss (Theorem 3.1).

What we argued so far is sufficient for ensuring that minimizing the surrogate loss $\mathcal{L}_{\mathrm{sur}}^{\mathcal{D},\sigma}(\mathbf{w})$ leads to the claimed bound on the L_2^2 loss. However, it is not sufficient for obtaining the claimed sample and computational complexities, and there are additional technical hurdles that can only be handled using the specific structural properties of our resulting optimization problem. In particular, using solely smoothness and sharpness of the objective (even if the sharpness held on the entire region over which we are optimizing), would only lead to complexities scaling with $\frac{1}{\epsilon}$, using standard results from stochastic convex optimization. However, the complexity that we get is *exponentially better*, scaling with $\mathrm{polylog}(\frac{1}{\epsilon})$. This is enabled by the refined variance bound for the stochastic gradient estimate (see Corollary D.11), which, unlike in standard stochastic optimization settings (where we get a fixed upper bound), scales with $\mathrm{OPT}+\|\mathbf{w}-\mathbf{w}^*\|_2^2$.[3] This property enables us to construct high-accuracy gradient estimates using mini-batching, which further leads to the improved linear rates within the (nonconvex) region where the surrogate loss is sharp. To complete the argument, we further show that the excluded region on which the sharpness does not hold does not negatively impact the overall complexity, as within it the target approximation guarantee for the L_2^2 loss holds.

1.3. Notation

For $n \in \mathbb{Z}_+$, we denote by $[n]$ the set $\{1,\ldots,n\}$. We use lowercase boldface letters for vectors and uppercase bold letters for matrices. For $\mathbf{x}\in\mathbb{R}^d$ and $i\in[d]$, $\mathbf{x}_i$ denotes the ith coordinate of $\mathbf{x}$, and $\|\mathbf{x}\|_2 := (\sum_{i=1}^d {\mathbf{x}_i}^2)^{1/2}$ denotes the ℓ_2-norm of $\mathbf{x}$. We use $\mathbf{x}\cdot\mathbf{y}$ for the standard inner product of $\mathbf{x},\mathbf{y}\in\mathbb{R}^d$ and $\theta(\mathbf{x},\mathbf{y})$ for the angle between $\mathbf{x},\mathbf{y}$. We use $\mathbb{1}_{\mathcal{E}}$ for the characteristic function of the set/event $\mathcal{E}$, i.e., $\mathbb{1}_{\mathcal{E}}(\mathbf{x})=1$ if $\mathbf{x}\in\mathcal{E}$ and $\mathbb{1}_{\mathcal{E}}(\mathbf{x})=0$ if $\mathbf{x}\notin\mathcal{E}$. We denote by $\mathcal{B}(r)=\{\mathbf{u}:\|\mathbf{u}\|_2\le r\}$ the ℓ_2-ball of radius r. We use the standard asymptotic notation $\widetilde{O}(\cdot)$ and $\widetilde{\Omega}(\cdot)$ to omit polylogarithmic factors in the argument. We write $E\gtrsim F$ for two nonnegative expressions E and F to denote that there exists some universal constant $c>0$ (independent of the variables or parameters on which E and F depend) such that $E\ge c\,F$. We use $\mathbf{E}_{X\sim\mathcal{D}}[X]$ for the expectation of random variable X according to the distribution $\mathcal{D}$ and $\mathbf{Pr}[\mathcal{E}]$ for the probability of event $\mathcal{E}$. For simplicity of exposition, we may omit the distribution when it is clear from the context. For $(\mathbf{x},y)$ distributed according to $\mathcal{D}$, we denote by $\mathcal{D}_{\mathbf{x}}$ the marginal distribution of $\mathbf{x}$.

2. Landscape of Noise-Free Surrogate

We start by defining the class of activations and the distributional assumptions under which our results apply. We then establish our first structural result, showing that these conditions suffice for sharpness of the noise-free surrogate.

2.1. Activations and Distributional Assumptions

The main assumptions used throughout this paper to prove sharpness results are summarized below.

Definition 2.1 (Monotonic Unbounded Activations, (Diakonikolas et al., 2022b))**.** Let $\sigma:\mathbb{R}\mapsto\mathbb{R}$ be nondecreasing and let $\alpha,\beta>0$. We say that σ is (monotonic) (α,β)-unbounded if (i) σ is α-Lipschitz; and (ii) $\sigma'(t)\ge\beta$

[3] Similar variance bound assumptions have been made in the more recent literature on stochastic optimization; see, e.g., Assumption 4.3(c) in (Bottou et al., 2018). We note, however, that our guarantees hold with high probability (compared to the more common expectation guarantees) and that the bulk of of our technical contribution lies in proving that such a variance bound holds, rather than in analyzing SGD under such an assumpton.

for all $t > 0$.

The above class contains a range of popular activations, including the ReLU (which is $(1,1)$-unbounded), and the Leaky ReLU with parameter $0 \le \lambda \le \frac{1}{2}$, i.e., $\sigma(t) = \max\{\lambda t, (1-\lambda)t\}$ (which is is $(1-\lambda, 1-\lambda)$-unbounded).

Our results apply for the following class of activations.

Assumption 2.2 (Controlled Activation)**.** The activation function $\sigma : \mathbb{R} \to \mathbb{R}$ is (α, β)-unbounded, for some positive parameters $\alpha \ge 1, \beta \in (0,1)$, and it holds that $\sigma(0) = 0$.

The assumption on the activation is important both for the convergence analysis of our algorithm and for proving the sharpness property of the surrogate loss.

We can now state our distributional assumptions.

Assumption 2.3 (Margin)**.** There exists $\mathbf{w}^* \in \mathcal{W}^*$ and parameters $\gamma, \lambda \in (0,1]$ such that $\mathbf{E}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}\left[\mathbf{x}\mathbf{x}^\top \mathbb{1}\left\{\mathbf{w}^* \cdot \mathbf{x} \ge \gamma \|\mathbf{w}^*\|_2\right\}\right] \succeq \lambda \mathbf{I}$.

We note that in order to obtain a constant-factor approximate learner, the parameters γ and λ in Assumption 2.3 should be dimension-independent constants.

Assumption 2.4 (Concentration)**.** There exists a non-increasing $h : \mathbb{R}_+ \to \mathbb{R}_+$ satisfying $h(r) \le B r^{-(4+\rho)}$ for some parameters $B \ge 1$ and $1 \ge \rho > 0$, such that for any $\mathbf{u} \in \mathcal{B}(1)$ and any $r \ge 1$, it holds $\mathbf{Pr}[|\mathbf{u} \cdot \mathbf{x}| \ge r] \le h(r)$.

The concentration property enables us to control the moments of $|\mathbf{u} \cdot \mathbf{x}|$, playing an important role when we bound the variance of the gradient of the empirical surrogate loss.

2.2. Key Assumptions Suffice for Sharpness

We now prove that Assumptions 2.2–2.4 suffice to guarantee that the noise-free surrogate loss is sharp. We provide a proof sketch under the simplifying assumption that $\|\mathbf{w}^*\|_2 = 1$. The full proof can be found in Appendix C.1.

Lemma 2.5. *Suppose that Assumptions 2.2–2.4 hold. Then the noise-free surrogate loss $\bar{\mathcal{L}}_{\mathrm{sur}}^{\mathcal{D},\sigma}$ is $\Omega(\lambda^2\gamma\beta\rho/B)$-sharp in the ball $\mathcal{B}(2\|\mathbf{w}^*\|_2)$, i.e., $\forall \mathbf{w} \in \mathcal{B}(2\|\mathbf{w}^*\|_2)$,*

$$\nabla \bar{\mathcal{L}}_{\mathrm{sur}}^{\mathcal{D},\sigma}(\mathbf{w}) \cdot (\mathbf{w} - \mathbf{w}^*) \gtrsim \lambda^2\gamma\beta\rho/B \|\mathbf{w} - \mathbf{w}^*\|_2^2 \,.$$

Proof Sketch of Lemma 2.5. Observe that $\nabla \bar{\mathcal{L}}_{\mathrm{sur}}^{\mathcal{D},\sigma}(\mathbf{w}) = \mathbf{E}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}\left[(\sigma(\mathbf{w}\cdot\mathbf{x}) - \sigma(\mathbf{w}^*\cdot\mathbf{x}))\mathbf{x}\right]$. Using the fact that σ is non-decreasing, it holds that $\nabla \bar{\mathcal{L}}_{\mathrm{sur}}^{\mathcal{D},\sigma}(\mathbf{w}) \cdot (\mathbf{w} - \mathbf{w}^*) = \mathbf{E}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[|\sigma(\mathbf{w}\cdot\mathbf{x}) - \sigma(\mathbf{w}^*\cdot\mathbf{x})||\mathbf{w}\cdot\mathbf{x} - \mathbf{w}^*\cdot\mathbf{x}|]$. Denote $\mathcal{E}_m = \{\mathbf{w}^*\cdot\mathbf{x} \ge \gamma\}$. Using the fact that every term inside the expectation is nonnegative, we can further bound $\nabla \bar{\mathcal{L}}_{\mathrm{sur}}^{\mathcal{D},\sigma}(\mathbf{w}) \cdot (\mathbf{w} - \mathbf{w}^*)$ from below by

$$\nabla \bar{\mathcal{L}}_{\mathrm{sur}}^{\mathcal{D},\sigma}(\mathbf{w}) \cdot (\mathbf{w} - \mathbf{w}^*) \ge \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[|\sigma(\mathbf{w}\cdot\mathbf{x}) - \sigma(\mathbf{w}^*\cdot\mathbf{x})||\mathbf{w}\cdot\mathbf{x} - \mathbf{w}^*\cdot\mathbf{x}|\mathbb{1}_{\mathcal{E}_m}(\mathbf{x})] \,.$$

Since σ is (α,β)-unbounded, we have that $\sigma'(t) \ge \beta$ for all $t \in (0,\infty)$. By the mean value theorem, we can show that for $t_2 \ge t_1 \ge 0$, we have $|\sigma(t_1) - \sigma(t_2)| \ge \beta|t_1 - t_2|$. Additionally, if $t_1 \ge 0$ and $t_2 \le 0$, then $|\sigma(t_1) - \sigma(t_2)| \ge \beta t_1$. Therefore, by combining the above, and denoting the event $\{\mathbf{x} : \mathbf{w}\cdot\mathbf{x} \le 0,\ \mathbf{w}^*\cdot\mathbf{x} \ge \gamma\}$ as $\mathcal{E}_0$, we get

$$\begin{aligned}
&\nabla \bar{\mathcal{L}}_{\mathrm{sur}}^{\mathcal{D},\sigma}(\mathbf{w}) \cdot (\mathbf{w} - \mathbf{w}^*) \\
&\ge \beta \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[(\mathbf{w}\cdot\mathbf{x} - \mathbf{w}^*\cdot\mathbf{x})^2 \mathbb{1}\{\mathbf{w}\cdot\mathbf{x} > 0,\ \mathcal{E}_m(\mathbf{x})\}] \\
&+ \beta \underbrace{\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[|\mathbf{w}^*\cdot\mathbf{x}||\mathbf{w}\cdot\mathbf{x} - \mathbf{w}^*\cdot\mathbf{x}|\mathbb{1}_{\mathcal{E}_0}(\mathbf{x})]}_{(Q)} \,.
\end{aligned} \tag{3}$$

We show that the term (Q) can be bounded below by a quantity that is proportional to $\mathbf{E}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}\left[(\mathbf{w}\cdot\mathbf{x} - \mathbf{w}^*\cdot\mathbf{x})^2 \mathbb{1}_{\mathcal{E}_0}(\mathbf{x})\right]$. To this end, we establish the following claim.

Claim 2.6. For $r_0 \ge 1$, define the event $\mathcal{E}_1 = \mathcal{E}_1(r_0) = \{\mathbf{x} : -2r_0 < \mathbf{w}\cdot\mathbf{x} \le 0, \mathcal{E}_m(\mathbf{x})\}$. It holds $(Q) \ge (\gamma/(3r_0))\, \mathbf{E}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}\left[(\mathbf{w}\cdot\mathbf{x} - \mathbf{w}^*\cdot\mathbf{x})^2 \mathbb{1}_{\mathcal{E}_1}(\mathbf{x})\right]$.

Proof of Claim 2.6. Since $\mathcal{E}_1 \subseteq \mathcal{E}_0$, it holds that $(Q) \ge \mathbf{E}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[|\mathbf{w}^*\cdot\mathbf{x}||\mathbf{w}\cdot\mathbf{x} - \mathbf{w}^*\cdot\mathbf{x}|\mathbb{1}_{\mathcal{E}_1}(\mathbf{x})]$. Restricting $\mathbf{x}$ on the event $\mathcal{E}_1$, it holds that $|\mathbf{w}\cdot\mathbf{x}| \le 2(r_0/\gamma)|\mathbf{w}^*\cdot\mathbf{x}|$. Thus,

$$\mathbf{w}^*\cdot\mathbf{x} - \mathbf{w}\cdot\mathbf{x} = |\mathbf{w}^*\cdot\mathbf{x}| + |\mathbf{w}\cdot\mathbf{x}| \le (1 + 2r_0/\gamma)|\mathbf{w}^*\cdot\mathbf{x}|.$$

By Assumption 2.3 we have that $\gamma \in (0,1]$, therefore we get that $|\mathbf{w}^*\cdot\mathbf{x}| \ge \gamma/(\gamma + 2r_0) \ge \gamma/(3r_0)$, since $r_0 \ge 1$. Taking the expectation of $|\mathbf{w}^*\cdot\mathbf{x}||\mathbf{w}\cdot\mathbf{x} - \mathbf{w}^*\cdot\mathbf{x}|$ with $\mathbf{x}$ restricted on event $\mathcal{E}_1$, we obtain

$$\begin{aligned}
(Q) &\ge \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[|\mathbf{w}^*\cdot\mathbf{x}||\mathbf{w}\cdot\mathbf{x} - \mathbf{w}^*\cdot\mathbf{x}|\mathbb{1}_{\mathcal{E}_1}(\mathbf{x})] \\
&\ge \gamma/(3r_0) \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}\left[(\mathbf{w}\cdot\mathbf{x} - \mathbf{w}^*\cdot\mathbf{x})^2 \mathbb{1}_{\mathcal{E}_1}(\mathbf{x})\right] \,,
\end{aligned}$$

as desired. □

Combining Equation (3) and Claim 2.6, we get that

$$\begin{aligned}
&\nabla \bar{\mathcal{L}}_{\mathrm{sur}}^{\mathcal{D},\sigma}(\mathbf{w}) \cdot (\mathbf{w} - \mathbf{w}^*) \ge \\
&\frac{\beta\gamma}{3r_0} \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[(\mathbf{w}\cdot\mathbf{x} - \mathbf{w}^*\cdot\mathbf{x})^2 \mathbb{1}\{\mathbf{w}\cdot\mathbf{x} > -2r_0,\ \mathcal{E}_m(\mathbf{x})\}],
\end{aligned} \tag{4}$$

where in the last inequality we used the fact that $1 \ge \gamma/(3r_0)$ (since $\gamma \in (0,1]$ and $r_0 \ge 1$). To complete the proof, we need to show that, for an appropriate choice of r_0, the probability of the event $\{\mathbf{x} : \mathbf{w}\cdot\mathbf{x} > -2r_0,\ \mathbf{w}^*\cdot\mathbf{x} > \gamma\}$ is close to the probability of the event $\{\mathbf{x} : \mathbf{w}^*\cdot\mathbf{x} \ge \gamma\}$. Given such a statement, the lemma follows from Assumption 2.3. Formally, we show the following claim.

Claim 2.7. Let $r_0 \geq 1$ such that $h(r_0) \leq \lambda^2\rho/(20B)$. Then, for all $\mathbf{w} \in \mathcal{B}(2\|\mathbf{w}^*\|_2)$, we have that

$$\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[(\mathbf{w}\cdot\mathbf{x} - \mathbf{w}^*\cdot\mathbf{x})^2 \mathbb{1}\{\mathbf{w}\cdot\mathbf{x} > -2r_0,\, \mathcal{E}_m(\mathbf{x})\}] \geq (\lambda/2)\|\mathbf{w}^* - \mathbf{w}\|_2^2 \,.$$

Since $h(r) \leq B/r^{4+\rho}$ and $h(r)$ is decreasing, such an r_0 exists and we can always take $r_0 \geq 1$.

Combining Equation (4) and Claim 2.7, we get:

$$\nabla\bar{\mathcal{L}}_{\mathrm{sur}}^{\mathcal{D},\sigma}(\mathbf{w})\cdot(\mathbf{w}-\mathbf{w}^*) \gtrsim \frac{\gamma\lambda\beta}{r_0}\|\mathbf{w}-\mathbf{w}^*\|_2^2.$$

To complete the proof of Lemma 2.5, it remains to choose r_0 appropriately. By Claim 2.7, we need to select r_0 to be sufficiently large so that $h(r_0) \leq \lambda^2\rho/(20B)$. By Assumption 2.4, we have that $h(r) \leq B/r^{4+\rho}$. Thus, we can choose $r_0 = 5B/(\lambda\rho)$, by our assumptions. □

3. Efficient Constant-Factor Approximation

We now outline our main technical approach, including the algorithm, its analysis, connections between the L_2^2 loss and the two (noisy and noise-free) surrogates, and the role of sharpness. For space constraints, this section contains simplified proofs and proof sketches, while the full technical details are deferred to Appendix C.

3.1. The Landscape of Surrogate Loss

We start this section by showing that the landscape of surrogate loss connects with the error of the true loss.

Theorem 3.1. *Let $\mathcal{D}$ be a distribution supported on $\mathbb{R}^d \times \mathbb{R}$ and let $\sigma : \mathbb{R} \mapsto \mathbb{R}$ be an (α,β)-unbounded activation. Fix $\mathbf{w}^* \in \mathcal{W}^*$ and suppose that $\mathcal{D}_\mathbf{x}$ satisfies Assumptions 2.3 and 2.4 with respect to $\mathbf{w}^*$. Furthermore, let $C > 0$ be a sufficiently small absolute constant and let $\bar{\mu} = C\lambda^2\gamma\beta\rho/B$. Then, for any $\epsilon > 0$ and $\hat{\mathbf{w}} \in \mathcal{B}(2\|\mathbf{w}^*\|_2)$, so that $\mathcal{L}_{\mathrm{sur}}^{\mathcal{D},\sigma}(\hat{\mathbf{w}}) - \inf_{\mathbf{w}\in\mathcal{B}(2\|\mathbf{w}^*\|_2)} \mathcal{L}_{\mathrm{sur}}^{\mathcal{D},\sigma}(\mathbf{w}) \leq \epsilon$, it holds $\mathcal{L}_2^{\mathcal{D},\sigma}(\hat{\mathbf{w}}) \leq O((\alpha B/(\rho\bar{\mu}))^2)(\mathcal{L}_2^{\mathcal{D},\sigma}(\mathbf{w}^*) + \alpha\epsilon)$.*

Proof. For this proof, we assume for ease of presentation that $\mathbf{E}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}\left[\mathbf{x}\mathbf{x}^\top\right] \preceq \mathbf{I}$ and $B, \rho, \alpha = 1$. Denote $\mathcal{K}$ as the set of $\hat{\mathbf{w}}$ such that $\hat{\mathbf{w}} \in \mathcal{B}(2\|\mathbf{w}^*\|_2)$ and $\mathcal{L}_{\mathrm{sur}}^{\mathcal{D},\sigma}(\hat{\mathbf{w}}) - \inf_{\mathbf{w}\in\mathcal{B}(2\|\mathbf{w}^*\|_2)} \mathcal{L}_{\mathrm{sur}}^{\mathcal{D},\sigma}(\mathbf{w}) \leq \epsilon$.

Next observe that the set of minimizers of the loss $\mathcal{L}_{\mathrm{sur}}^{\mathcal{D},\sigma}$ inside the ball $\mathcal{B}(2\|\mathbf{w}^*\|_2)$ is convex. Furthermore, the set $\mathcal{B}(2\|\mathbf{w}^*\|_2)$ is compact. Thus, for any point $\mathbf{w}' \in \mathcal{B}(2\|\mathbf{w}^*\|_2)$ that minimizes $\mathcal{L}_{\mathrm{sur}}^{\mathcal{D},\sigma}$ it will either hold that $\|\nabla\mathcal{L}_{\mathrm{sur}}^{\mathcal{D},\sigma}(\mathbf{w}')\|_2 = 0$ or $\mathbf{w}' \in \partial\mathcal{B}(2\|\mathbf{w}^*\|_2)$. Let $\mathcal{W}_{\mathrm{sur}}^*$ be the set of minimizers of $\mathcal{L}_{\mathrm{sur}}^{\mathcal{D},\sigma}$.

We first show that if there exists a minimizer $\mathbf{w}' \in \mathcal{W}_{\mathrm{sur}}^*$ such that $\mathbf{w}' \in \partial\mathcal{B}(2\|\mathbf{w}^*\|_2)$, then any point $\mathbf{w}$ inside the set $\mathcal{B}(2\|\mathbf{w}^*\|_2)$ gets error proportional to $\mathcal{L}_2^{\mathcal{D},\sigma}(\mathbf{w}^*)$. Observe for such point $\mathbf{w}'$, by the necessary condition of optimality,

$$\nabla\mathcal{L}_{\mathrm{sur}}^{\mathcal{D},\sigma}(\mathbf{w}')\cdot(\mathbf{w}'-\mathbf{w}) \leq 0 \,, \tag{5}$$

for any $\mathbf{w} \in \mathcal{B}(2\|\mathbf{w}^*\|_2)$. Using Proposition 3.2, we get that either $\nabla\mathcal{L}_{\mathrm{sur}}^{\mathcal{D},\sigma}(\mathbf{w}')\cdot(\mathbf{w}'-\mathbf{w}^*) \geq (\bar{\mu}/2)\|\mathbf{w}'-\mathbf{w}^*\|_2^2$ or $\mathbf{w}' \in \{\mathbf{w} : \|\mathbf{w}-\mathbf{w}^*\|_2^2 \leq (20/\bar{\mu}^2)\mathcal{L}_2^{\mathcal{D},\sigma}(\mathbf{w}^*)\}$. But Equation (5) contradicts with $\nabla\mathcal{L}_{\mathrm{sur}}^{\mathcal{D},\sigma}(\mathbf{w}')\cdot(\mathbf{w}'-\mathbf{w}^*) \geq (\bar{\mu}/2)\|\mathbf{w}'-\mathbf{w}^*\|_2^2 > 0$, since $\mathbf{w}' \in \partial\mathcal{B}(2\|\mathbf{w}^*\|_2)$, $\|\mathbf{w}'\|_2 = 2\|\mathbf{w}^*\|_2$; hence $\mathbf{w}' \neq \mathbf{w}^*$. So it must be the case that $\mathbf{w}' \in \{\mathbf{w} : \|\mathbf{w}-\mathbf{w}^*\|_2^2 \leq (20/\bar{\mu}^2)\mathcal{L}_2^{\mathcal{D},\sigma}(\mathbf{w}^*)\}$. Again, we have that $\mathbf{w}' \in \partial\mathcal{B}(2\|\mathbf{w}^*\|_2)$, therefore $\|\mathbf{w}'-\mathbf{w}^*\|_2 \geq \|\mathbf{w}^*\|_2$. Hence, $(20/\bar{\mu}^2)\mathcal{L}_2^{\mathcal{D},\sigma}(\mathbf{w}^*) \geq \|\mathbf{w}^*\|_2^2$. Therefore, for any $\mathbf{w} \in \mathcal{B}(2\|\mathbf{w}^*\|_2)$, we have

$$\mathcal{L}_2^{\mathcal{D},\sigma}(\mathbf{w}) = \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}\left[(\sigma(\mathbf{w}\cdot\mathbf{x}) - y)^2\right] \leq 2\mathcal{L}_2^{\mathcal{D},\sigma}(\mathbf{w}^*) + \|\mathbf{w}-\mathbf{w}^*\|_2^2 = O(1/\bar{\mu}^2)\mathcal{L}_2^{\mathcal{D},\sigma}(\mathbf{w}^*) \,,$$

where we used the fact that $\mathbf{E}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}\left[\mathbf{x}\mathbf{x}^\top\right] \preceq \mathbf{I}$ and that σ is 1-Lipschitz. Since the inequality above holds for any $\mathbf{w} \in \mathcal{B}(2\|\mathbf{w}^*\|_2)$, it will also be true for $\hat{\mathbf{w}} \in \mathcal{K} \subseteq \mathcal{B}(2\|\mathbf{w}^*\|_2)$. It remains to consider the case where the minimizers $\mathcal{W}_{\mathrm{sur}}^*$ are strictly inside the $\mathcal{B}(2\|\mathbf{w}^*\|_2)$. Note that $\mathcal{L}_{\mathrm{sur}}^{\mathcal{D},\sigma}(\mathbf{w})$ is 1-smooth. Therefore, for any $\hat{\mathbf{w}} \in \mathcal{K}$ it holds $\|\nabla\mathcal{L}_{\mathrm{sur}}^{\mathcal{D},\sigma}(\hat{\mathbf{w}})\|_2^2 \leq 2\epsilon$. By Proposition 3.2 (stated and proved below), we get that either $\|\hat{\mathbf{w}}-\mathbf{w}^*\|_2^2 \leq (1/\bar{\mu}^2)\mathcal{L}_2^{\mathcal{D},\sigma}(\mathbf{w}^*)$ or that $\sqrt{2\epsilon} \geq (\bar{\mu}/2)\|\hat{\mathbf{w}}-\mathbf{w}^*\|_2$. Therefore, we obtain that $\|\hat{\mathbf{w}}-\mathbf{w}^*\|_2^2 \leq (1/\bar{\mu}^2)(\mathcal{L}_2^{\mathcal{D},\sigma}(\mathbf{w}^*) + \epsilon)$. □

The proof of Theorem 3.1 required the following proposition which shows that if the current vector $\mathbf{w}$ is sufficiently far away from the true vector $\mathbf{w}^*$, then the gradient of the surrogate loss has a large component in the direction of $\mathbf{w} - \mathbf{w}^*$; in other words, the surrogate loss is sharp.

Proposition 3.2. *Let $\mathcal{D}$ be a distribution supported on $\mathbb{R}^d \times \mathbb{R}$ and let $\sigma : \mathbb{R} \mapsto \mathbb{R}$ be an (α,β)-unbounded activation. Suppose that $\mathcal{D}_\mathbf{x}$ satisfies Assumptions 2.3 and 2.4 and let $C > 0$ be a sufficiently small absolute constant and let $\bar{\mu} = C\lambda^2\gamma\beta\rho/B$. Fix $\mathbf{w}^* \in \mathcal{W}^*$ and let $S = \mathcal{B}(2\|\mathbf{w}^*\|_2) - \{\mathbf{w} : \|\mathbf{w}-\mathbf{w}^*\|_2^2 \leq (20B/(\rho\bar{\mu}^2))\mathcal{L}_2^{\mathcal{D},\sigma}(\mathbf{w}^*)\}$. Then, the surrogate loss $\mathcal{L}_{\mathrm{sur}}^{\mathcal{D},\sigma}$ is $(\bar{\mu}/2)$-sharp in S, i.e.,*

$$\nabla\mathcal{L}_{\mathrm{sur}}^{\mathcal{D},\sigma}(\mathbf{w})\cdot(\mathbf{w}-\mathbf{w}^*) \geq (\bar{\mu}/2)\|\mathbf{w}-\mathbf{w}^*\|_2^2, \quad \forall\mathbf{w} \in S.$$

Proof. For this proof, we assume for ease of presentation that $\mathbf{E}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}\left[\mathbf{x}\mathbf{x}^\top\right] \preceq \mathbf{I}$ and $\kappa, B, \rho, \alpha = 1$. We show that $\nabla\mathcal{L}_{\mathrm{sur}}^{\mathcal{D},\sigma}(\mathbf{w})\cdot(\mathbf{w}-\mathbf{w}^*)$ is bounded away from zero. We decompose the gradient into two parts, i.e., $\nabla\mathcal{L}_{\mathrm{sur}}^{\mathcal{D},\sigma}(\mathbf{w}) = (\nabla\mathcal{L}_{\mathrm{sur}}^{\mathcal{D},\sigma}(\mathbf{w}) - \nabla\mathcal{L}_{\mathrm{sur}}^{\mathcal{D},\sigma}(\mathbf{w}^*)) + \nabla\mathcal{L}_{\mathrm{sur}}^{\mathcal{D},\sigma}(\mathbf{w}^*)$. First, we bound $\nabla\mathcal{L}_{\mathrm{sur}}^{\mathcal{D},\sigma}(\mathbf{w}^*)$ in the direction $\mathbf{w} - \mathbf{w}^*$, which yields

$$\nabla\mathcal{L}_{\mathrm{sur}}^{\mathcal{D},\sigma}(\mathbf{w}^*)\cdot(\mathbf{w}-\mathbf{w}^*) \geq -\sqrt{\mathcal{L}_2^{\mathcal{D},\sigma}(\mathbf{w}^*)}\|\mathbf{w}-\mathbf{w}^*\|_2 \,,$$

Algorithm 1 Stochastic Gradient Descent on Surrogate Loss

Input: Iterations: T, sample access from $\mathcal{D}$, batch size N, step size η, bound M. Initialize: $\mathbf{w}^{(0)} \leftarrow \mathbf{0}$.
for $t = 1$ **to** T **do**
  Draw N samples $\{(\mathbf{x}(j), y(j))\}_{j=1}^N \sim \mathcal{D}$.
  For each $j \in [N]$, $y(j) \leftarrow \mathrm{sign}(y(j)) \min(|y(j)|, M)$.
  $\mathbf{g}^{(t)} \leftarrow \frac{1}{N}\sum_{j=1}^N (\sigma(\mathbf{w}^{(t)} \cdot \mathbf{x}(j)) - y(j))\mathbf{x}(j)$.
  $\mathbf{w}^{(t+1)} \leftarrow \mathbf{w}^{(t)} - \eta \mathbf{g}^{(t)}$.
end for
Output: The weight vector $\mathbf{w}^{(T)}$.

where we used the Cauchy-Schwarz inequality and that $\mathbf{E}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[\mathbf{x}\mathbf{x}^\top] \preceq \mathbf{I}$. It remains to bound the remaining term. Note that $(\nabla \mathcal{L}_{\mathrm{sur}}^{\mathcal{D},\sigma}(\mathbf{w}) - \nabla \mathcal{L}_{\mathrm{sur}}^{\mathcal{D},\sigma}(\mathbf{w}^*)) = \nabla \bar{\mathcal{L}}_{\mathrm{sur}}^{\mathcal{D},\sigma}(\mathbf{w})$. Using the fact that $\bar{\mathcal{L}}_{\mathrm{sur}}^{\mathcal{D},\sigma}(\mathbf{w})$ is $\bar{\mu}$-sharp for any $\mathbf{w} \in S$ from Lemma 2.5, it holds that $\nabla \bar{\mathcal{L}}_{\mathrm{sur}}^{\mathcal{D},\sigma}(\mathbf{w}) \cdot (\mathbf{w} - \mathbf{w}^*) \geq \bar{\mu}\|\mathbf{w} - \mathbf{w}^*\|_2^2$. Combining everything together, we get the claimed result. □

3.2. Fast Rates for L_2^2 Loss Minimization

In this section, we proceed to show that when the surrogate loss is sharp, applying batch Stochastic Gradient Descent (SGD) on the empirical surrogate loss obtains a C-approximate parameter $\hat{\mathbf{w}}$ of the L_2^2 loss in linear time. To be specific, consider the following iteration update

$$\mathbf{w}^{(t+1)} = \operatorname*{argmin}_{\mathbf{w}\in\mathcal{B}(W)} \left\{\mathbf{w}\cdot\mathbf{g}^{(t)} + (1/(2\eta))\|\mathbf{w} - \mathbf{w}^{(t)}\|_2^2\right\}, \quad (6)$$

where η is the step size and $\mathbf{g}^{(t)}$ is the empirical gradient of the surrogate loss, i.e., $\mathbf{g}^{(t)} = \frac{1}{N}\sum_{j=1}^N (\sigma(\mathbf{w}^{(t)} \cdot \mathbf{x}(j)) - y(j))\mathbf{x}(j)$. The algorithm is summarized in Algorithm 1.

We define the helper functions H_2 and H_4 as follows: $H_2(r) \triangleq \max_{\mathbf{u}\in\mathcal{B}(1)} \mathbf{E}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}\left[(\mathbf{u}\cdot\mathbf{x})^2 \mathbb{1}\{|\mathbf{u}\cdot\mathbf{x}| \geq r\}\right]$ and $H_4(r) \triangleq \max_{\mathbf{u}\in\mathcal{B}(1)} \mathbf{E}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}\left[(\mathbf{u}\cdot\mathbf{x})^4 \mathbb{1}\{|\mathbf{u}\cdot\mathbf{x}| \geq r\}\right]$.

Now we state our main theorem.

Theorem 3.3 (Main Algorithmic Result). *Fix $\epsilon, W > 0$ and suppose Assumptions 2.2 to 2.4 hold. Let $\mu := \mu(\lambda, \gamma, \beta, \rho, B)$ be a sufficiently small constant multiple of $\lambda^2\gamma\beta\rho/B$, and let $M = \alpha W H_2^{-1}(\epsilon/(4\alpha^2 W^2))$. Further, choose parameter r_ϵ large enough so that $H_4(r_\epsilon)$ is a sufficiently small constant multiple of ϵ. Then after $T = \widetilde{\Theta}\left((B^2\alpha^2/(\rho^2\mu^2))\log(W/\epsilon)\right)$ iterations with batch size $N = \Omega(dT(r_\epsilon^2 + \alpha^2 M^2))$, Algorithm 1 converges to a point $\mathbf{w}^{(T)}$ such that $\mathcal{L}_2^{\mathcal{D},\sigma}(\mathbf{w}^{(T)}) = O\left((B^2\alpha^2/(\rho^2\mu^2))\right)\mathrm{OPT} + \epsilon$, with probability at least $2/3$.*

As shown in Theorem 3.1, when we find a vector $\hat{\mathbf{w}}$ that minimizes the surrogate loss, then this $\hat{\mathbf{w}}$ is itself a C-approximate solution of Problem 1.1. However, minimizing the surrogate loss can be expensive in sample and computational complexity. Proposition 3.2 says that we can achieve strong-convexity-like rates, as long as we are far away from a minimizer of the L_2^2 loss. Roughly speaking, we show that at each iteration t, it holds $\|\mathbf{w}^{(t+1)} - \mathbf{w}^*\|_2^2 \leq C\|\mathbf{w}^{(t)} - \mathbf{w}^*\|_2^2 + \mathrm{OPT}$, where $0 < C < 1$ is some constant depending on the parameters α, β, μ, ρ, and B. Then $\|\mathbf{w}^{(t)} - \mathbf{w}^*\|_2$ contracts fast as long as $\|\mathbf{w}^{(t)} - \mathbf{w}^*\|_2^2 > (1/(1-C))\mathrm{OPT}$. When this condition fails, we have converged to a point that achieves $O(\mathrm{OPT})$ L_2^2 error.

The following lemma states that we can truncate the labels y to $y' \leq M$, where M is a parameter depending on $\mathcal{D}_\mathbf{x}$. The proof can be found in Appendix D.2.

Lemma 3.4. *Let $M = \alpha W H_2^{-1}(\epsilon/(4\alpha^2 W^2))$ and $y' = \mathrm{sign}(y)\min(|y|, M)$. Then we have that $\mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}\left[(\sigma(\mathbf{w}^*\cdot\mathbf{x}) - y')^2\right] = \mathrm{OPT} + \epsilon$.*

Lemma 3.4 allows us to assume that $|y| \leq M$.

Proof Sketch of Theorem 3.3. For this sketch, we will assume for ease of notation that $B, \rho, \alpha = 1$ and that $\mathbf{E}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}\left[\mathbf{x}\mathbf{x}^\top\right] \preceq \mathbf{I}$. The blueprint of the proof is to show that Algorithm 1 minimizes $\|\mathbf{w} - \mathbf{w}^*\|_2$ efficiently, in terms of both the sample complexity and the iteration complexity. To be specific, we show that at each iteration, $\|\mathbf{w}^{(t+1)} - \mathbf{w}^*\|_2^2 \leq (1-C)\|\mathbf{w}^{(t)} - \mathbf{w}^*\|_2^2 + (\text{small error})$, where $0 < C < 1$. The key technique is to exploit the sharpness property of the surrogate loss, which we have already proved in Proposition 3.2.

To this aim, we study the difference of $\|\mathbf{w}^{(t+1)} - \mathbf{w}^*\|_2^2$ and $\|\mathbf{w}^{(t)} - \mathbf{w}^*\|_2^2$. We remind the reader that for convenience of notation, we denote the empirical gradients as the following $\mathbf{g}^{(t)} = \frac{1}{N}\sum_{j=1}^N (\sigma(\mathbf{w}^{(t)} \cdot \mathbf{x}(j)) - y(j))\mathbf{x}(j)$, $\mathbf{g}^* = \frac{1}{N}\sum_{j=1}^N (\sigma(\mathbf{w}^* \cdot \mathbf{x}(j)) - y(j))\mathbf{x}(j)$. Moreover, we denote the noise-free empirical gradient by $\bar{\mathbf{g}}^{(t)}$, i.e., $\bar{\mathbf{g}}^{(t)} = \mathbf{g}^{(t)} - \mathbf{g}^*$. Plugging in the iteration scheme $\mathbf{w}^{(t+1)} = \mathbf{w}^{(t)} - \eta\mathbf{g}^{(t)}$ while expanding the squared norm, we get

$$\begin{aligned}
&\|\mathbf{w}^{(t+1)} - \mathbf{w}^*\|_2^2 \\
&= \underbrace{\|\mathbf{w}^{(t)} - \mathbf{w}^*\|_2^2 - 2\eta\nabla\mathcal{L}_{\mathrm{sur}}^{\mathcal{D},\sigma}(\mathbf{w}^{(t)}) \cdot (\mathbf{w}^{(t)} - \mathbf{w}^*)}_{Q_1} \\
&\underbrace{-2\eta(\mathbf{g}^{(t)} - \nabla\mathcal{L}_{\mathrm{sur}}^{\mathcal{D},\sigma}(\mathbf{w}^{(t)})) \cdot (\mathbf{w}^{(t)} - \mathbf{w}^*) + \eta^2\|\mathbf{g}^{(t)}\|_2^2}_{Q_2} .
\end{aligned}$$

Observe that we decomposed the right hand side into two parts, the true contribution of the gradient (Q_1) and the estimation error (Q_2). In order to utilize the sharpness property of surrogate loss at the point $\mathbf{w}^{(t)}$, the conditions

$$\begin{aligned}
&\mathbf{w}^{(t)} \in \mathcal{B}(2\|\mathbf{w}^*\|_2) \text{ and} \\
&\mathbf{w}^{(t)} \in \{\mathbf{w} : \|\mathbf{w}^{(t)} - \mathbf{w}^*\|_2^2 \geq 20/\bar{\mu}^2\mathrm{OPT}\} \quad (7)
\end{aligned}$$

need to be satisfied. For the first condition, recall that we initialized $\mathbf{w}^{(0)} = \mathbf{0}$; hence, Equation (7) is valid for $t = 0$. By induction, it suffices to show that assuming

$\mathbf{w}^{(t)} \in \mathcal{B}(2\|\mathbf{w}^*\|_2)$ holds, we have $\|\mathbf{w}^{(t+1)} - \mathbf{w}^*\|_2 \leq (1-C)\|\mathbf{w}^{(t)} - \mathbf{w}^*\|_2$ for some constant $0 < C < 1$. Thus, we assume temporarily that Equation (7) is true at iteration t, and we will show in the remainder of the proof that $\|\mathbf{w}^{(t+1)} - \mathbf{w}^*\|_2 \leq (1-C)\|\mathbf{w}^{(t)} - \mathbf{w}^*\|_2$ until we arrived at some final iteration T. Then, by induction, the first part of Equation (7) is satisfied at each step $t \leq T$. For the second condition, note that if it is violated at some iteration T, then $\|\mathbf{w}^{(T)} - \mathbf{w}^*\|_2^2 = O(\mathrm{OPT})$ implying that this would be the solution we are looking for and the algorithm could be terminated at T. Therefore, whenever $\|\mathbf{w}^{(t)} - \mathbf{w}^*\|_2^2$ is far away from OPT, the prerequisites of Proposition 3.2 are satisfied and $\mathcal{L}_{\mathrm{sur}}^{\mathcal{D},\sigma}$ is sharp.

For the first term (Q_1), using that $\nabla\mathcal{L}_{\mathrm{sur}}^{\mathcal{D},\sigma}(\mathbf{w}^{(t)})$ is μ-sharp by Proposition 3.2, we immediately get a sufficient decrease at each iteration, i.e., $\|\mathbf{w}^{(t+1)} - \mathbf{w}^*\|_2^2 \leq (1-C)\|\mathbf{w}^{(t)} - \mathbf{w}^*\|_2^2$. Namely, applying Proposition 3.2, we get

$$\begin{aligned}(Q_1) &= \|\mathbf{w}^{(t)} - \mathbf{w}^*\|_2^2 - 2\eta\nabla\mathcal{L}_{\mathrm{sur}}^{\mathcal{D},\sigma}(\mathbf{w}^{(t)}) \cdot (\mathbf{w}^{(t)} - \mathbf{w}^*) \\ &\leq (1 - 2\eta\mu)\|\mathbf{w}^{(t)} - \mathbf{w}^*\|_2^2 \,,\end{aligned}$$

where $\mu = C\lambda^2\gamma\beta$ for some sufficiently small constant C.

Now it suffices to show that (Q_2) can be bounded above by $C'\|\mathbf{w}^{(t)} - \mathbf{w}^*\|_2^2$, where C' is a parameter depending on η and μ that can be made comparatively small. Formally, we show the following claim.

Claim 3.5. Suppose $\eta \leq 1$. Fix $r_\epsilon \geq 1$ such that $H_4(r_\epsilon)$ is a sufficiently small constant multiple of ϵ. Choosing N to be a sufficiently large constant multiple of $(d/\delta)(r_\epsilon^2 + M^2)$, then we have with probability at least $1-\delta$

$$(Q_2) \leq ((3/2)\eta\mu + 8\eta^2)\|\mathbf{w}^{(t)} - \mathbf{w}^*\|_2^2 + (8\eta/\mu)(\mathrm{OPT} + \epsilon) \,.$$

Proof. Observe that by the inequality $\mathbf{x}\cdot\mathbf{y} \leq (\mu/2)\|\mathbf{x}\|_2^2 + (1/(2\mu))\|\mathbf{y}\|_2^2$ applied to the inner product $(\mathbf{g}^{(t)} - \nabla\mathcal{L}_{\mathrm{sur}}^{\mathcal{D},\sigma}(\mathbf{w}^{(t)})) \cdot (\mathbf{w}^{(t)} - \mathbf{w}^*)$, we get

$$\begin{aligned}(Q_2) \leq \frac{\eta}{\mu}\|\mathbf{g}^{(t)} - \nabla\mathcal{L}_{\mathrm{sur}}^{\mathcal{D},\sigma}(\mathbf{w}^{(t)})\|_2^2 + \eta\mu\|\mathbf{w}^{(t)} - \mathbf{w}^*\|_2^2 \\ + 2\eta^2\|\bar{\mathbf{g}}^{(t)}\|_2^2 + 2\eta^2\|\mathbf{g}^*\|_2^2 \,,\end{aligned}$$

where μ is the sharpness parameter and we used the definition that $\bar{\mathbf{g}}^{(t)} = \mathbf{g}^{(t)} - \mathbf{g}^*$ in the first inequality.

Note that $\|\mathbf{g}^{(t)} - \nabla\mathcal{L}_{\mathrm{sur}}^{\mathcal{D},\sigma}(\mathbf{w}^{(t)})\|_2^2 \leq 2\|\bar{\mathbf{g}}^{(t)} - \nabla\bar{\mathcal{L}}_{\mathrm{sur}}^{\mathcal{D},\sigma}(\mathbf{w}^{(t)})\|_2^2 + 2\|\mathbf{g}^* - \nabla\mathcal{L}_{\mathrm{sur}}^{\mathcal{D},\sigma}(\mathbf{w}^*)\|_2^2$, since we have $\bar{\mathcal{L}}_{\mathrm{sur}}^{\mathcal{D},\sigma}(\mathbf{w}^{(t)}) = \mathcal{L}_{\mathrm{sur}}^{\mathcal{D},\sigma}(\mathbf{w}^{(t)}) - \mathcal{L}_{\mathrm{sur}}^{\mathcal{D},\sigma}(\mathbf{w}^*)$. Thus, it holds

$$\begin{aligned}(Q_2) &\leq \eta\mu\|\mathbf{w}^{(t)} - \mathbf{w}^*\|_2^2 + 2\eta^2\|\bar{\mathbf{g}}^{(t)}\|_2^2 + 2\eta^2\|\mathbf{g}^*\|_2^2 + \\ &(2\eta/\mu)\left(\|\bar{\mathbf{g}}^{(t)} - \nabla\bar{\mathcal{L}}_{\mathrm{sur}}^{\mathcal{D},\sigma}(\mathbf{w}^{(t)})\|_2^2 + \|\mathbf{g}^* - \nabla\mathcal{L}_{\mathrm{sur}}^{\mathcal{D},\sigma}(\mathbf{w}^*)\|_2^2\right) \,.\end{aligned} \tag{8}$$

Furthermore, using standard concentration tools, it can be shown that when $N \geq Cd(r_\epsilon^2 + M^2)/\delta$ where C is a sufficiently large absolute constant, with probability at least $1-\delta$, it holds

$$\begin{aligned}\|\bar{\mathbf{g}}^{(t)} - \nabla\bar{\mathcal{L}}_{\mathrm{sur}}^{\mathcal{D},\sigma}(\mathbf{w}^{(t)})\|_2^2 \leq (\mu^2/4)\|\mathbf{w}^{(t)} - \mathbf{w}^*\|_2^2, \\ \|\bar{\mathbf{g}}^{(t)}\|_2^2 \leq 4\|\mathbf{w}^{(t)} - \mathbf{w}^*\|_2^2,\end{aligned}$$

and $\|\mathbf{g}^* - \nabla\mathcal{L}_{\mathrm{sur}}^{\mathcal{D},\sigma}(\mathbf{w}^*)\|_2^2 \leq \mathrm{OPT} + \epsilon$, $\|\mathbf{g}^*\|_2^2 \leq 2\mathrm{OPT} + \epsilon$. It remains to plug these bounds back into Equation (8). □

Combining the upper bounds on (Q_1) and (Q_2) and choosing $\eta = \mu/32$, we have:

$$\begin{aligned}\|\mathbf{w}^{(t+1)} - \mathbf{w}^*\|_2^2 \leq \left(1 - \mu^2/128\right)\|\mathbf{w}^{(t)} - \mathbf{w}^*\|_2^2 \\ + (1/4)\left(\mathrm{OPT} + \epsilon\right) \,.\end{aligned} \tag{9}$$

When $\|\mathbf{w}^{(t)} - \mathbf{w}^*\|_2^2 \geq (64/\mu^2)(\mathrm{OPT} + \epsilon)$, in other words when $\mathbf{w}^{(t)}$ is still away from the minimizer $\mathbf{w}^*$, it further holds with probability $1-\delta$:

$$\|\mathbf{w}^{(t+1)} - \mathbf{w}^*\|_2^2 \leq (1 - \mu^2/256)\|\mathbf{w}^{(t)} - \mathbf{w}^*\|_2^2, \tag{10}$$

which proves the sufficient decrease of $\|\mathbf{w}^{(t)} - \mathbf{w}^*\|_2^2$ that we proposed at the beginning.

Let T be the first iteration such that $\mathbf{w}^{(T)}$ satisfies $\|\mathbf{w}^{(T)} - \mathbf{w}^*\|_2^2 \leq (64/\mu^2)(\mathrm{OPT} + \epsilon)$. Recall that we need Equation (7) for every $t \leq T$ to be satisfied to implement sharpness. The first condition is satisfied naturally for $\|\mathbf{w}^{(t+1)} - \mathbf{w}^*\|_2^2 \leq \|\mathbf{w}^*\|_2^2$ as a consequence of Equation (10) (recall that $\mathbf{w}^{(0)} = 0$). For the second condition, when $t+1 \leq T$, we have $\|\mathbf{w}^{(t+1)} - \mathbf{w}^*\|_2^2 \geq (64/\mu^2)(\mathrm{OPT} + \epsilon)$, hence the second condition also holds.

When $t \leq T$, the contraction of $\|\mathbf{w}^{(t)} - \mathbf{w}^*\|_2^2$ indicates a linear convergence of SGD. Since $\mathbf{w}^{(0)} = 0$, $\|\mathbf{w}^*\|_2 \leq W$, it holds $\|\mathbf{w}^{(t)} - \mathbf{w}^*\|_2^2 \leq (1 - \mu^2/256)^t\|\mathbf{w}^{(0)} - \mathbf{w}^*\|_2^2 \leq \exp(-t\mu^2/256)W^2$. Thus, to generate $\mathbf{w}^{(T)}$ such that $\|\mathbf{w}^{(T)} - \mathbf{w}^*\|_2^2 \leq (64/\mu^2)(\mathrm{OPT} + \epsilon)$, it suffices to run Algorithm 1 for $T = \widetilde{\Theta}((1/\mu^2)\log(W/\epsilon))$ iterations. Recall that at each step t the contraction $\|\mathbf{w}^{(t+1)} - \mathbf{w}^*\|_2^2 \leq (1 - \mu^2/256)\|\mathbf{w}^{(t)} - \mathbf{w}^*\|_2^2$ holds with probability $1-\delta$, thus the union bound implies $\|\mathbf{w}^{(T)} - \mathbf{w}^*\|_2^2 \leq (64/\mu^2)(\mathrm{OPT} + \epsilon)$ holds with probability $1 - T\delta$. Moreover, as $\mathcal{L}_2^{\mathcal{D},\sigma}(\mathbf{w}^{(T)}) \lesssim \|\mathbf{w}^{(T)} - \mathbf{w}^*\|_2^2$, if $\|\mathbf{w}^{(T)} - \mathbf{w}^*\|_2^2 \leq (64/\mu^2)(\mathrm{OPT} + \epsilon)$, then $\mathcal{L}_2^{\mathcal{D},\sigma}(\mathbf{w}^{(T)}) = O(1/\mu^2)(\mathrm{OPT} + \epsilon)$. Letting $\delta = 1/(3T)$ completes the proof. □

4. Conclusion

We provided an efficient constant-factor approximate learner for the problem of agnostically learning a single neuron over structured classes of distributions. Notably, our algorithmic result applies under much milder distributional assumptions as compared to prior work. Our results are obtained by leveraging a sharpness property (a local error bound) from optimization theory that we prove holds for the considered

problems. This property is crucial both to establishing a constant factor approximation and to obtaining improved sample complexity and runtime. An interesting direction for future work is to explore whether sharpness can be leveraged to obtain positive results for other related learning problems.

5. Acknowledgements

PW was supported in part by NSF Award CCF-2007757. NZ was supported in part by NSF award DMS-2023239. ID was supported by NSF Medium Award CCF-2107079, NSF Award CCF-1652862 (CAREER), a Sloan Research Fellowship, and a DARPA Learning with Less Labels (LwLL) grant. JD was supported by NSF Award CCF-2007757 and by the Office of Naval Research under contract number N00014-22-1-2348. JD thanks Alexandre D'Aspremont and Jérôme Bolte for a useful discussion on local error bounds.

References

Auer, P., Herbster, M., and Warmuth, M. K. Exponentially many local minima for single neurons. In *Advances in Neural Information Processing Systems 8, NIPS*, pp. 316–322. MIT Press, 1995.

Awasthi, P., Tang, A., and Vijayaraghavan, A. Agnostic learning of general relu activation using gradient descent. *CoRR*, abs/2208.02711, 2022.

Bartlett, P., Jordan, M., and McAuliffe, J. Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, 101(473):138–156, 2006.

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.

Bottou, L., Curtis, F. E., and Nocedal, J. Optimization methods for large-scale machine learning. *Siam Review*, 60(2):223–311, 2018.

Chen, X. and Fukushima, M. Expected residual minimization method for stochastic linear complementarity problems. *Mathematics of Operations Research*, 30(4):1022–1038, 2005.

De, A., Diakonikolas, I., Feldman, V., and Servedio, R. A. Nearly optimal solutions for the chow parameters problem and low-weight approximation of halfspaces. *J. ACM*, 61(2):11:1–11:36, 2014.

De, A., Diakonikolas, I., and Servedio, R. A. The inverse shapley value problem. *Games Econ. Behav.*, 105:122–147, 2017.

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., 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., 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, 2022b.

Diakonikolas, I., Pavlou, C., Peebles, J., and Stewart, A. Efficient approximation algorithms for the inverse semi-value problem. In *21st International Conference on Autonomous Agents and Multiagent Systems, AAMAS 2022*, pp. 354–362. International Foundation for Autonomous Agents and Multiagent Systems (IFAAMAS), 2022c.

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

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

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

Hendrycks, D. and Gimpel, K. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*, 2016.

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

Kakade, S., 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 - The 22nd Conference on Learning Theory*, 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.

Karmakar, S. and Mukherjee, A. Provable training of a ReLU gate with an iterative non-gradient algorithm. *Neural Networks*, 151:264–275, 2022.

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 non-convex compound stochastic programs with applications to risk measure minimization. *Mathematics of Operations Research*, 2022.

Liu, M., Zhang, X., Zhang, L., Jin, R., and Yang, T. Fast rates of erm and stochastic approximation: Adaptive to error bound conditions. *Advances in Neural Information Processing Systems*, 31, 2018.

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.

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

Pang, J.-S. Error bounds in mathematical programming. *Mathematical Programming*, 79(1):299–332, 1997.

Ramachandran, P., Zoph, B., and Le, Q. V. Searching for activation functions. *arXiv preprint arXiv:1710.05941*, 2017.

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.

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

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

Supplementary Material

Organization The supplementary material is organized as follows: In Appendix A, we provide some remarks on the sharpness property we have been using throughout the paper. In Appendix B, we provide additional detailed comparison with prior work. In Appendix C and Appendix D, we present the full contents of Section 2 and Section 3 respectively, providing supplementary lemmas and completing the omitted proofs in the main body. Appendix E shows that there are many natural distributions satisfying Assumption 2.3 and Assumption 2.4. Finally, in Appendix F, we show that our results extend to certain non-monotonic distributions, including GeLUs (Hendrycks & Gimpel, 2016) and Swish (Ramachandran et al., 2017).

Additional Notation Some additional notation we use here is listed below. Given a distribution $\mathcal{D}$ on $\mathbb{R}^d \times \mathbb{R}$, we use $\{(\mathbf{x}(j), y(j))\}_{j=1}^N$ to denote N i.i.d. samples from $\mathcal{D}$. We slightly abuse the notation and denote by $\mathbf{e}_i$ the ith standard basis vector in $\mathbb{R}^d$. The notation $[\cdot]_+$ is used for the positive part of the argument, i.e., $[\cdot]_+ = \max\{\cdot, 0\}$. For a vector $\mathbf{x} = (\mathbf{x}_1, \cdots, \mathbf{x}_n)$, $[\cdot]_+$ is applied element-wise: $[\mathbf{x}]_+ := ([\mathbf{x}_1]_+, \cdots, [\mathbf{x}_n]_+)$. For nonnegative expressions E, F we write $E \gg F$ to denote $E \geq C\,F$, where $C > 0$ is a *sufficiently large* universal constant (independent of the parameters of E and F). The notation $\ll$ is defined similarly.

A. Remarks about Sharpness

We recall the formal definition of sharpness, already mentioned in the introduction.

Definition A.1 (Sharpness)**.** Given a function $f : \mathcal{C} \mapsto \mathbb{R}$ where $\mathcal{C} \subseteq \mathbb{R}^d$, suppose the set of its minimizers $\mathcal{Z}^* = \operatorname{argmin}_{\mathbf{z}\in\mathcal{C}} f(\mathbf{z})$ is closed and not empty. Let $f^* = \min_{\mathbf{z}\in\mathcal{C}} f(\mathbf{z})$. We say that f is μ-sharp, for some $\mu > 0$, if the following inequality holds:

$$f(\mathbf{z}) - f^* \geq \frac{\mu}{2}\operatorname{dist}(\mathbf{z}, \mathcal{Z}^*)^2, \forall \mathbf{z} \in \mathbb{R}^d,$$

where $\operatorname{dist}(\mathbf{z}, \mathcal{Z}^*) = \min_{\mathbf{z}^*\in\mathcal{Z}^*} \|\mathbf{z} - \mathbf{z}^*\|_2$.

Remark A.2. We will slightly abuse the name of sharpness to refer to sharpness-like properties. For example, if a function satisfies

$$\nabla f(\mathbf{z}) \cdot (\mathbf{z} - \mathbf{z}^*) \geq \mu\|\mathbf{z} - \mathbf{z}^*\|_2^2, \tag{11}$$

for some $\mathbf{z}^* \in \mathcal{Z}^*$, then we say f is μ-sharp. This is due to the fact that when f is a convex function, it holds $f(\mathbf{z}) - f^* \leq \nabla f(\mathbf{z}) \cdot (\mathbf{z} - \mathbf{z}^*)$, hence Definition A.1 implies Equation (11). Thus, Equation (11) can be viewed as a milder property of sharpness.

Compared to strong convexity, sharpness is a milder condition. Indeed, for any μ-strongly-convex function f, if $\mathbf{z}^* \in \operatorname{argmin}_{\mathbf{z}\in\mathbb{R}^d} f(\mathbf{z})$ then $f(\mathbf{z}) - f^* \geq \nabla f(\mathbf{z}^*) \cdot (\mathbf{z} - \mathbf{z}^*) + \frac{\mu}{2}\|\mathbf{z} - \mathbf{z}^*\|_2^2 \geq \frac{\mu}{2}\|\mathbf{z} - \mathbf{z}^*\|_2^2$; therefore, f is μ-sharp. However, the opposite does not hold in general. For example, consider $f : \mathbb{R} \mapsto \mathbb{R}$ defined by $f(z) = z^2$ if $z \geq 0$ and $f(z) = 0$ otherwise, whose set of minimizers on $\mathbb{R}$ is $\mathcal{Z}^* = (-\infty, 0]$ and $f^* = 0$. Thus, if $z \geq 0$, then $f(z) - f^* \geq \operatorname{dist}(z, \mathcal{Z}^*)^2 = z^2$ and if $z < 0$, we have $f(z) - f^* = 0 = \operatorname{dist}(z, \mathcal{Z}^*)^2$. Therefore, f is 2-sharp but it is not strongly convex.

B. Additional Comparison to Prior Work

Here we summarize and provide additional technical comparison to prior work that did not appear in the main body, due to space limitations.

Comparison with Frei et al. (2020). The work of Frei et al. (2020) studies the problem of learning ReLU (and other nonlinear) activations and shows that gradient descent on the L_2^2 loss converges to a point achieving error $K\sqrt{\mathrm{OPT}}$. The parameter K depends on the maximum norm of the points $\mathbf{x}$, and can depend on the dimension d. Specifically, even for the basic case that the marginal distribution on examples is the standard normal distribution, the parameter K scales (polynomially) with d. That is, Frei et al. (2020) does not provide constant factor approximate learners in this setting.

Comparison with Diakonikolas et al. (2020a). The work of Diakonikolas et al. (2020a) studies the problem of learning ReLU activations using the same surrogate loss we consider in this work. Our work differs from Diakonikolas et al. (2020a) in two key aspects. The first aspect concerns the generality and strengh of results; the second aspect concerns the techniques.

In terms of the results themselves, the algorithm given in Diakonikolas et al. (2020a) is restricted to the case of ReLUs (while we handle a broader family of activations). More importantly, the distributional assumptions of Diakonikolas et al. (2020a) are much stricter than ours, — focusing on logconcave distributions — whereas we handle broader classes of distributions, including heavy tailed and discrete distributions, not covered by any prior work (see also Appendix E). Informally, what allows us to handle broader classes of distributions is our focus on proving the sharpness property (as opposed to strong convexity), which is a much milder property. Further, we show that it suffices for this property to hold only in a small region (ball of radius $2\|\mathbf{w}^*\|_2$) and for the (impossible to evaluate) noise-free surrogate loss. Another remark is that Diakonikolas et al. (2020a) assume that the (corrupted) labels are bounded, not fully capturing the agnostic setting. By contrast, our analysis can handle unbounded labels, i.e., we do not make further assumptions about the noise. Finally, even if we restrict our focus to the class of logconcave distributions, our algorithm has sample complexity scaling with $\mathrm{polylog}(1/\epsilon)$, as opposed to $1/\epsilon^2$ in Diakonikolas et al. (2020a).

The second and more important difference lies in the techniques that are used in each work. Diakonikolas et al. (2020a) optimizes the surrogate loss directly and shows that finding a point with a small gradient of the surrogate loss leads to the small L_2^2 error. More specifically, the requirement in Diakonikolas et al. (2020a) is that the gradient is sufficiently small so that the optimality gap of the surrogate loss is of the order ϵ. This statement is similar to the result we show in Theorem 3.1. Crucially, while we utilize the gradients of the surrogate loss in the algorithm and in the analysis, we never impose a requirement that the optimality gap of the surrogate loss is of the order ϵ. Instead, we show that as long as the gradient is larger than order-$\sqrt{\mathrm{OPT}+\epsilon}$, sharpness holds and linear convergence rate applies. On the other hand, when the gradient is of the order $\sqrt{\mathrm{OPT}+\epsilon}$ or smaller, we argue that the candidate solution that the algorithm maintains is already an $(O(\mathrm{OPT})+\epsilon)$-approximate solution in terms of the L_2^2 error. This approach further enables us to be agnostic in the value of OPT. Notably, if $\epsilon \ll \mathrm{OPT}$ and we were to require that the algorithm finds a solution with either the gradient of the order $\sqrt{\epsilon}$ or the optimality gap ϵ, we would need to optimize the surrogate loss within a region where the sharpness does not necessarily hold. Without sharpness, only sublinear rates of convergence apply, and the number of iterations increases to order-$\frac{1}{\epsilon}$. Thus, leveraging the structural properties that we prove in this work is crucial to obtaining the exponential improvements in sample and computational complexities.

Finally, Diakonikolas et al. (2020a) requires the surrogate loss to be strongly convex to connect the small gradient condition with the small L_2^2 error. This makes the argument rather straightforward, compared with what is used in our work. For the sake of discussion, assume that $\mathcal{L}_{\mathrm{sur}}^{\mathcal{D},\sigma}$ is 1- strongly convex and the distribution is isotropic. Furthermore, denote by $\mathbf{w}^*$ the minimizer of the L_2^2 loss and by $\mathbf{w}'$ the minimizer of $\mathcal{L}_{\mathrm{sur}}^{\mathcal{D},\sigma}$. The property that $\mathcal{L}_{\mathrm{sur}}^{\mathcal{D},\sigma}$ is strongly convex implies that $\|\nabla\mathcal{L}_{\mathrm{sur}}^{\mathcal{D},\sigma}(\mathbf{w}^*)-\nabla\mathcal{L}_{\mathrm{sur}}^{\mathcal{D},\sigma}(\mathbf{w}')\|_2^2 \geq \|\mathbf{w}^*-\mathbf{w}'\|_2^2$; furthermore, it can be shown that $\|\nabla\mathcal{L}_{\mathrm{sur}}^{\mathcal{D},\sigma}(\mathbf{w}^*)\|_2^2 \leq \mathcal{L}_2^{\mathcal{D},\sigma}(\mathbf{w}^*)$. Therefore, because $\nabla\mathcal{L}_{\mathrm{sur}}^{\mathcal{D},\sigma}(\mathbf{w}')=\mathbf{0}$, it immediately follows that $\mathcal{L}_2^{\mathcal{D},\sigma}(\mathbf{w}') \lesssim \mathcal{L}_2^{\mathcal{D},\sigma}(\mathbf{w}^*)$. Our work leverages a much weaker property than strong convexity — sharpness — as summarized in Proposition 3.2. This weaker property turns out to be sufficient to ensure that the noise cannot make the gradient field guide us far away from the optimal solution.

Comparison with Diakonikolas et al. (2022b). The work of Diakonikolas et al. (2022b) studies the problem of ReLU (and other unbounded activations) regression with agnostic noise. They show that for a class of well-behaved distributions (see Definition E.1) gradient descent on the L_2^2 loss converges to a point achieving $O(\mathrm{OPT})+\epsilon$ error. Moreover, the sample and computational complexities of their algorithm are similar to those achieved in our work (for the class of well-behaved distributions). On the other hand, the distributional assumptions used in Diakonikolas et al. (2022b) are quite strong. Specifically, the "well-behaved" assumption requires that the marginal distribution have sub-exponential concentration and anti-anti-concentration in every lower dimensional subspace; that is, the probability density function is lower bounded by a positive constant at every point. The latter assumption does not allow for several discrete distributions, like discrete Gaussians or uniform on the cube, that is handled in our work. Moreover, our work can additionally handle distributions with much weaker concentration properties.

Comparison with Karmakar & Mukherjee (2022). In a weaker noise model, the work of Karmakar & Mukherjee (2022) considered a similar-looking — though crucially different — condition for robust ReLU regression, namely that:

$$\lambda_{\min}\left(\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_{\mathbf{x}}}\left[\mathbf{x}\mathbf{x}^\top \mathbb{1}\{\mathbf{w}^*\cdot\mathbf{x}\geq 2\theta^*\}\right]\right)=\lambda_1>0, \tag{12}$$

where θ^* is the largest possible absolute value of the noise; in other words, $\theta^*=\sup_{(\mathbf{x},y)\sim\mathcal{D}}|y-\sigma(\mathbf{w}^*\cdot\mathbf{x})|$. It is worth noting that Equation (12) cannot be easily satisfied, as the noise in the agnostic model is not bounded. But even if the

noise was bounded, this condition would give slack for a small number of distributions. For instance in the uniform on the hypercube, if $\theta^* > 1/2$, then the minimum eigenvalue is zero. Furthermore, the algorithm in that work converges to a point that achieves $O(\theta^*)$ error, instead of $O(\mathrm{OPT})$ error. In contrast, we make no assumptions about the boundedness of the noise, and obtain near-optimal error in more general settings.

Additional Related Work As mentioned in the introduction, the convex surrogate we leverage was first defined in (Auer et al., 1995) and then implicitly used in (Kalai & Sastry, 2009; Kakade et al., 2011) for learning GLMs. In addition to these and the aforementioned works, it is worth mentioning that the same convex surrogate has been useful in the context of learning linear separators from limited information (De et al., 2014) and in related game-theoretic settings (De et al., 2017; Diakonikolas et al., 2022c).

C. Full Version of Section 2

Discussion about the Parameters in Assumptions 2.2 to 2.4 If an activation σ is (α', β')-bounded, then it is also (α, β)-bounded for $\alpha \geq \alpha'$ and $\beta \leq \beta'$. This justifies the convention $\alpha \geq 1$ and $\beta \leq 1$ in Assumption 2.2. If $\sigma(0) \neq 0$, we can generate new labels y' by subtracting $\sigma(0)$ from y and consider the activation $\sigma_0(t) = \sigma(t) - \sigma(0)$. Similar reasoning justifies the conventions $\lambda, \gamma \in (0, 1]$ in Assumption 2.3 and $B \geq 1$, $\rho \leq 1$ in Assumption 2.4.

C.1. Proof of Lemma 2.5

For convenience, we restate the lemma followed by its detailed proof.

Lemma C.1. *Suppose that Assumptions 2.2–2.4 hold. Then the noise-free surrogate loss $\bar{\mathcal{L}}_{\mathrm{sur}}^{\mathcal{D},\sigma}$ is $\Omega(\lambda^2\gamma\beta\rho/B)$-sharp in the ball $\mathcal{B}(2\|\mathbf{w}^*\|_2)$, i.e., $\forall \mathbf{w} \in \mathcal{B}(2\|\mathbf{w}^*\|_2)$ we have*

$$\nabla \bar{\mathcal{L}}_{\mathrm{sur}}^{\mathcal{D},\sigma}(\mathbf{w}) \cdot (\mathbf{w} - \mathbf{w}^*) \gtrsim \lambda^2\gamma\beta\rho/B \|\mathbf{w} - \mathbf{w}^*\|_2^2 \,.$$

Proof. By definition, we can write $\nabla \bar{\mathcal{L}}_{\mathrm{sur}}^{\mathcal{D},\sigma}(\mathbf{w}) = \mathbf{E}_{\mathbf{x}\sim\mathcal{D}_{\mathbf{x}}}\left[(\sigma(\mathbf{w}\cdot\mathbf{x}) - \sigma(\mathbf{w}^*\cdot\mathbf{x}))\mathbf{x}\right]$. Therefore, the inner product $\nabla \bar{\mathcal{L}}_{\mathrm{sur}}^{\mathcal{D},\sigma}(\mathbf{w}) \cdot (\mathbf{w} - \mathbf{w}^*)$ can be written as

$$\begin{aligned}
&\nabla \bar{\mathcal{L}}_{\mathrm{sur}}^{\mathcal{D},\sigma}(\mathbf{w}) \cdot (\mathbf{w} - \mathbf{w}^*) \\
&= \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_{\mathbf{x}}}\left[(\sigma(\mathbf{w}\cdot\mathbf{x}) - \sigma(\mathbf{w}^*\cdot\mathbf{x}))(\mathbf{w}\cdot\mathbf{x} - \mathbf{w}^*\cdot\mathbf{x})\right] \\
&= \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_{\mathbf{x}}}\left[|\sigma(\mathbf{w}\cdot\mathbf{x}) - \sigma(\mathbf{w}^*\cdot\mathbf{x})||\mathbf{w}\cdot\mathbf{x} - \mathbf{w}^*\cdot\mathbf{x}|\right] \\
&\geq \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_{\mathbf{x}}}\left[|\sigma(\mathbf{w}\cdot\mathbf{x}) - \sigma(\mathbf{w}^*\cdot\mathbf{x})||\mathbf{w}\cdot\mathbf{x} - \mathbf{w}^*\cdot\mathbf{x}|\mathbb{1}\{\mathbf{w}^*\cdot\mathbf{x} \geq \gamma\|\mathbf{w}^*\|_2\}\right] \,,
\end{aligned}$$

where the second equality is due to the non-decreasing property of σ, and the inequality is due to the fact that every term inside the expectation is nonnegative. Since σ is (α, β)-unbounded, we have that $\sigma'(t) \geq \beta$ for all $t \in [0, \infty)$. By the mean value theorem, for $t_2 \geq t_1 \geq 0$, we have $\sigma(t_1) - \sigma(t_2) = \sigma'(\xi)(t_1 - t_2)$ for some $\xi \in (t_1, t_2)$. Thus, we obtain that $|\sigma(t_1) - \sigma(t_2)| \geq \beta|t_1 - t_2|$. Additionally, if $t_1 \geq 0$ and $t_2 \leq 0$, then $|\sigma(t_1) - \sigma(t_2)| = |\sigma(t_1) - \sigma(0)| + |\sigma(0) - \sigma(t_2)| \geq |\sigma(t_1) - \sigma(0)| \geq \beta t_1$. Therefore, by combining the above, we get

$$\begin{aligned}
\nabla \bar{\mathcal{L}}_{\mathrm{sur}}^{\mathcal{D},\sigma}(\mathbf{w}) \cdot (\mathbf{w} - \mathbf{w}^*) \geq \beta &\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_{\mathbf{x}}}\left[(\mathbf{w}\cdot\mathbf{x} - \mathbf{w}^*\cdot\mathbf{x})^2\mathbb{1}\{\mathbf{w}\cdot\mathbf{x} > 0,\, \mathbf{w}^*\cdot\mathbf{x} > \gamma\|\mathbf{w}^*\|_2\}\right] \\
&+ \beta \underbrace{\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_{\mathbf{x}}}\left[|\mathbf{w}^*\cdot\mathbf{x}||\mathbf{w}\cdot\mathbf{x} - \mathbf{w}^*\cdot\mathbf{x}|\mathbb{1}\{\mathbf{w}\cdot\mathbf{x} \leq 0,\, \mathbf{w}^*\cdot\mathbf{x} \geq \gamma\|\mathbf{w}^*\|_2\}\right]}_{(Q)} \,. \qquad (13)
\end{aligned}$$

Denote $\mathcal{E}_0 = \{\mathbf{x} : \mathbf{w}\cdot\mathbf{x} \leq 0,\, \mathbf{w}^*\cdot\mathbf{x} \geq \gamma\|\mathbf{w}^*\|_2\}$. We show that the term (Q) can be bounded below by a quantity that is proportional to $\mathbf{E}_{\mathbf{x}\sim\mathcal{D}_{\mathbf{x}}}\left[(\mathbf{w}\cdot\mathbf{x} - \mathbf{w}^*\cdot\mathbf{x})^2\mathbb{1}\{\mathbf{w}\cdot\mathbf{x} \leq 0,\, \mathbf{w}^*\cdot\mathbf{x} > \gamma\|\mathbf{w}^*\|_2\}\right]$. To this end, we establish the following claim.

Claim C.2. For $r_0 \geq 1$, define the event $\mathcal{E}_1 = \mathcal{E}_1(r_0) = \{\mathbf{x} : -2r_0\|\mathbf{w}^*\|_2 < \mathbf{w}\cdot\mathbf{x} \leq 0, \mathbf{w}^*\cdot\mathbf{x} \geq \gamma\|\mathbf{w}^*\|_2\}$. It holds $(Q) \geq (\gamma/(3r_0))\, \mathbf{E}_{\mathbf{x}\sim\mathcal{D}_{\mathbf{x}}}\left[(\mathbf{w}\cdot\mathbf{x} - \mathbf{w}^*\cdot\mathbf{x})^2\mathbb{1}_{\mathcal{E}_1}(\mathbf{x})\right]$.

Proof of Claim C.2. Since $\mathcal{E}_1 \subseteq \mathcal{E}_0$, it holds that $(Q) \geq \mathbf{E}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[|\mathbf{w}^*\cdot\mathbf{x}||\mathbf{w}\cdot\mathbf{x}-\mathbf{w}^*\cdot\mathbf{x}|\mathbb{1}_{\mathcal{E}_1}(\mathbf{x})]$. Restricting $\mathbf{x}$ on the event $\mathcal{E}_1$, it holds that $|\mathbf{w}\cdot\mathbf{x}| \leq 2(r_0/\gamma)|\mathbf{w}^*\cdot\mathbf{x}|$. Therefore, we get

$$\mathbf{w}^*\cdot\mathbf{x}-\mathbf{w}\cdot\mathbf{x} = |\mathbf{w}^*\cdot\mathbf{x}| + |\mathbf{w}\cdot\mathbf{x}| \leq (1+2r_0/\gamma)|\mathbf{w}^*\cdot\mathbf{x}|.$$

By Assumption 2.3 we have that $\gamma \in (0,1]$, therefore we get that $|\mathbf{w}^*\cdot\mathbf{x}| \geq \gamma/(\gamma+2r_0) \geq \gamma/(3r_0)$, since $r_0 \geq 1$. Taking the expectation of $|\mathbf{w}^*\cdot\mathbf{x}||\mathbf{w}\cdot\mathbf{x}-\mathbf{w}^*\cdot\mathbf{x}|$ with $\mathbf{x}$ restricted on event $\mathcal{E}_1$, we obtain

$$(Q) \geq \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[|\mathbf{w}^*\cdot\mathbf{x}||\mathbf{w}\cdot\mathbf{x}-\mathbf{w}^*\cdot\mathbf{x}|\mathbb{1}_{\mathcal{E}_1}(\mathbf{x})] \geq \gamma/(3r_0)\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}\left[(\mathbf{w}\cdot\mathbf{x}-\mathbf{w}^*\cdot\mathbf{x})^2\mathbb{1}_{\mathcal{E}_1}(\mathbf{x})\right] ,$$

as desired. □

Combining Equation (13) and Claim C.2, we get that

$$\begin{aligned}
&\nabla\bar{\mathcal{L}}^{\mathcal{D},\sigma}_{\mathrm{sur}}(\mathbf{w})\cdot(\mathbf{w}-\mathbf{w}^*)\\
&\geq \beta\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}\left[(\mathbf{w}\cdot\mathbf{x}-\mathbf{w}^*\cdot\mathbf{x})^2\mathbb{1}\{\mathbf{w}\cdot\mathbf{x}>0, \mathbf{w}^*\cdot\mathbf{x}\geq\gamma\|\mathbf{w}^*\|_2\}\right]\\
&\quad+\frac{\beta\gamma}{3r_0}\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}\left[(\mathbf{w}\cdot\mathbf{x}-\mathbf{w}^*\cdot\mathbf{x})^2\mathbb{1}\{-2r_0\|\mathbf{w}^*\|_2<\mathbf{w}\cdot\mathbf{x}\leq 0,\ \mathbf{w}^*\cdot\mathbf{x}\geq\gamma\|\mathbf{w}^*\|_2\}\right]\\
&\geq \frac{\beta\gamma}{3r_0}\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}\left[(\mathbf{w}\cdot\mathbf{x}-\mathbf{w}^*\cdot\mathbf{x})^2\mathbb{1}\{\mathbf{w}\cdot\mathbf{x}>-2r_0\|\mathbf{w}^*\|_2,\ \mathbf{w}^*\cdot\mathbf{x}>\gamma\|\mathbf{w}^*\|_2\}\right] ,
\end{aligned} \tag{14}$$

where in the last inequality we used the fact that $1 \geq \gamma/(3r_0)$ (since $\gamma \in (0,1]$ and $r_0 \geq 1$). To complete the proof, we need to show that, for an appropriate choice of r_0, the probability of the event $\{\mathbf{x} : \mathbf{w}\cdot\mathbf{x} > -2r_0\|\mathbf{w}^*\|_2,\ \mathbf{w}^*\cdot\mathbf{x} > \gamma\|\mathbf{w}^*\|_2\}$ is close to the probability of the event $\{\mathbf{x} : \mathbf{w}^*\cdot\mathbf{x} \geq \gamma\|\mathbf{w}^*\|_2\}$. Given such a statement, the lemma follows from Assumption 2.3.

Formally, we show the following claim.

Claim C.3. Let $r_0 \geq 1$ such that $h(r_0) \leq \lambda^2\rho/(20B)$. Then, for all $\mathbf{w} \in \mathcal{B}(2\|\mathbf{w}^*\|_2)$, we have that

$$\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}\left[(\mathbf{w}\cdot\mathbf{x}-\mathbf{w}^*\cdot\mathbf{x})^2\mathbb{1}\{\mathbf{w}\cdot\mathbf{x}>-2r_0\|\mathbf{w}^*\|_2,\ \mathbf{w}^*\cdot\mathbf{x}>\gamma\|\mathbf{w}^*\|\}\right] \geq \frac{\lambda}{2}\|\mathbf{w}^*-\mathbf{w}\|_2^2 .$$

Since $h(r) \leq B/r^{4+\rho}$ and $h(r)$ is decreasing, such an r_0 exists and we can always make $r_0 \geq 1$.

Proof of Claim C.3. By Assumption 2.3, we have that $\mathbf{E}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}\left[(\mathbf{w}^*\cdot\mathbf{x})^2\mathbb{1}\{\mathbf{w}^*\cdot\mathbf{x}\geq\gamma\|\mathbf{w}^*\|_2\}\right] \geq \lambda\|\mathbf{w}^*\|_2^2$. Let $\mathcal{E}_2 = \{\mathbf{w}\cdot\mathbf{x} \leq -2r_0\|\mathbf{w}^*\|_2, \mathbf{w}^*\cdot\mathbf{x} \geq \gamma\|\mathbf{w}^*\|_2\}$. We have that

$$\begin{aligned}
&\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}\left[(\mathbf{w}\cdot\mathbf{x}-\mathbf{w}^*\cdot\mathbf{x})^2\mathbb{1}\{\mathbf{w}\cdot\mathbf{x}>-2r_0\|\mathbf{w}^*\|_2,\ \mathbf{w}^*\cdot\mathbf{x}>\gamma\|\mathbf{w}^*\|\}\right]\\
&= \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}\left[(\mathbf{w}\cdot\mathbf{x}-\mathbf{w}^*\cdot\mathbf{x})^2\mathbb{1}\{\mathbf{w}^*\cdot\mathbf{x}\geq\gamma\|\mathbf{w}^*\|_2\}\right] - \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}\left[(\mathbf{w}\cdot\mathbf{x}-\mathbf{w}^*\cdot\mathbf{x})^2\mathbb{1}_{\mathcal{E}_2}(\mathbf{x})\right]\\
&\geq \lambda\|\mathbf{w}^*-\mathbf{w}\|_2^2 - \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}\left[(\mathbf{w}\cdot\mathbf{x}-\mathbf{w}^*\cdot\mathbf{x})^2\mathbb{1}_{\mathcal{E}_2}(\mathbf{x})\right] .
\end{aligned}$$

By the Cauchy-Schwarz inequality, we get

$$\begin{aligned}
\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}\left[(\mathbf{w}\cdot\mathbf{x}-\mathbf{w}^*\cdot\mathbf{x})^2\mathbb{1}_{\mathcal{E}_2}(\mathbf{x})\right] &\leq \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}\left[(\mathbf{w}\cdot\mathbf{x}-\mathbf{w}^*\cdot\mathbf{x})^2\mathbb{1}\{\mathbf{w}\cdot\mathbf{x}\leq-2r_0\|\mathbf{w}^*\|_2\}\right]\\
&\leq \|\mathbf{w}-\mathbf{w}^*\|_2^2\max_{\mathbf{u}\in\mathcal{B}(1)}\sqrt{\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[(\mathbf{u}\cdot\mathbf{x})^4]}\sqrt{\mathbf{Pr}\left[\mathbf{w}\cdot\mathbf{x}\leq-2r_0\|\mathbf{w}^*\|_2\}\right]} .
\end{aligned}$$

Since $\mathbf{w} \in \mathcal{B}(2\|\mathbf{w}^*\|_2)$, it holds that $\mathbf{w}/(2\|\mathbf{w}^*\|_2) \in \mathcal{B}(1)$. Thus, from the concentration properties of $\mathcal{D}_\mathbf{x}$, it follows that $\mathbf{Pr}\left[\mathbf{w}\cdot\mathbf{x} \leq -2r_0\|\mathbf{w}^*\|_2\}\right] \leq h(r_0)$. It remains to bound $\max_{\mathbf{u}\in\mathcal{B}(1)}\mathbf{E}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}\left[(\mathbf{u}\cdot\mathbf{x})^4\right]$. It is not hard to see that for distributions satisfying the concentration property of Assumption 2.4, $\max_{\mathbf{u}\in\mathcal{B}(1)}\mathbf{E}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}\left[(\mathbf{u}\cdot\mathbf{x})^4\right]$ as well as $\max_{\mathbf{u}\in\mathcal{B}(1)}\mathbf{E}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}\left[(\mathbf{u}\cdot\mathbf{x})^2\right]$ are at most $5B/\rho$. The proof of the following simple fact can be found in Appendix C.2.

Fact C.4. Let $\mathcal{D}_{\mathbf{x}}$ be a distribution satisfying Assumption 2.4. Then $\max_{\mathbf{u}\in\mathcal{B}(1)} \mathbf{E}_{\mathbf{x}\sim\mathcal{D}_{\mathbf{x}}}\left[(\mathbf{u}\cdot\mathbf{x})^i\right] \le 5B/\rho$ for $i = 2, 4$.

Although only the bound on the 4th order moment is needed here, the upper bound on $\max_{\mathbf{u}\in\mathcal{B}(1)} \mathbf{E}_{\mathbf{x}\sim\mathcal{D}_{\mathbf{x}}}\left[(\mathbf{u}\cdot\mathbf{x})^2\right]$ will also be used in later sections.

Therefor, by our choice of r_0, we have $h(r_0) \le \frac{\lambda^2\rho}{20B}$, hence $\max_{\mathbf{u}\in\mathcal{B}(1)} \mathbf{E}_{\mathbf{x}\sim\mathcal{D}_{\mathbf{x}}}\left[(\mathbf{u}\cdot\mathbf{x})^4\right] h(r_0) \le \lambda^2/4$. Therefore,

$$\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_{\mathbf{x}}}\left[(\mathbf{w}\cdot\mathbf{x}-\mathbf{w}^*\cdot\mathbf{x})^2\mathbb{1}_{\mathcal{E}_2}(\mathbf{x})\right] \le (\lambda/2)\|\mathbf{w}-\mathbf{w}^*\|_2^2 \,,$$

completing the proof of Claim C.3. □

Combining Equation (14) and Claim C.3, we get:

$$\nabla\bar{\mathcal{L}}_{\text{sur}}^{\mathcal{D},\sigma}(\mathbf{w})\cdot(\mathbf{w}-\mathbf{w}^*) \gtrsim \frac{\gamma\lambda\beta}{r_0}\|\mathbf{w}-\mathbf{w}^*\|_2^2.$$

To complete the proof, it remains to choose r_0 appropriately. By Claim C.3, we need to select r_0 to be sufficiently large so that $h(r_0) \le \lambda^2\rho/(20B)$. By Assumption 2.4, we have that $h(r) \le B/r^{4+\rho}$. Thus, we can choose $r_0 = 5B/(\lambda\rho)$, which is at least 1 by our assumptions. This completes the proof of the lemma.

□

C.2. Proof of Fact C.4

We restate and prove the following fact.

Fact C.5. Let $\mathcal{D}_{\mathbf{x}}$ be a distribution satisfying Assumption 2.4. Then $\max_{\mathbf{u}\in\mathcal{B}(1)} \mathbf{E}_{\mathbf{x}\sim\mathcal{D}_{\mathbf{x}}}\left[(\mathbf{u}\cdot\mathbf{x})^i\right] \le 5B/\rho$ for $i = 2, 4$.

Proof. Let $i = 2$ or 4. By Assumption 2.4, for any unit vector $\mathbf{u}$, we have

$$\begin{aligned}\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_{\mathbf{x}}}\left[(\mathbf{u}\cdot\mathbf{x})^i\right] &= \int_0^\infty \mathbf{Pr}\left[(\mathbf{u}\cdot\mathbf{x})^i \ge t\right]\mathrm{d}t \\ &= \int_0^\infty is^{i-1}\mathbf{Pr}\left[|\mathbf{u}\cdot\mathbf{x}| \ge s\right]\mathrm{d}s \\ &\le \int_0^\infty is^{i-1}\min\{1, h(s)\}\,\mathrm{d}s.\end{aligned}$$

By Assumption 2.4 we have $h(s) \le B/s^{4+\rho}$ for some $1 \ge \rho > 0$ and $B \ge 1$, thus it further holds

$$\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_{\mathbf{x}}}\left[(\mathbf{u}\cdot\mathbf{x})^i\right] \le \int_0^1 is^{i-1}\,\mathrm{d}s + \int_1^\infty is^{i-1}h(s)\,\mathrm{d}s \le 1 + B\int_1^\infty is^{i-1}\frac{1}{s^{4+\rho}}\,\mathrm{d}s \le \frac{5B}{\rho}.$$

□

D. Full Version of Section 3

D.1. The Landscape of Surrogate Loss

Theorem D.1 (Landscape of Surrogate Loss)**.** *Let $\bar{\mu} \in (0, 1]$ and $\alpha, \kappa \ge 1$. Let $\mathcal{D}$ be a distribution supported on $\mathbb{R}^d \times \mathbb{R}$ and let $\sigma : \mathbb{R} \mapsto \mathbb{R}$ be an (α, β)-unbounded activation for some $\beta > 0$. Furthermore, assume that the maximum eigenvalue of the matrix $\mathbf{E}_{\mathbf{x}\sim\mathcal{D}_{\mathbf{x}}}[\mathbf{x}\mathbf{x}^\top]$ is κ. Further, fix $\mathbf{w}^* \in \mathcal{W}^*$ and suppose $\bar{\mathcal{L}}_{\text{sur}}^{\mathcal{D},\sigma}$ is $\bar{\mu}$-sharp with respect to $\mathbf{w}^*$ in a subset $S_1 \subseteq \mathbb{R}^d$. Let $S_2 = \{\mathbf{w} : \mathcal{L}_2^{\mathcal{D},\sigma}(\mathbf{w}) \ge (4\alpha\kappa/\bar{\mu})^2\mathcal{L}_2^{\mathcal{D},\sigma}(\mathbf{w}^*)\}$. Then for any $\mathbf{w} \in S_1 \cap S_2$, we have*

$$\|\nabla\mathcal{L}_{\text{sur}}^{\mathcal{D},\sigma}(\mathbf{w})\|_2 \le \alpha\sqrt{\kappa}\|\mathbf{w}-\mathbf{w}^*\|_2 + \sqrt{\kappa\mathcal{L}_2^{\mathcal{D},\sigma}(\mathbf{w}^*)}\,,$$

and

$$\|\nabla\mathcal{L}_{\text{sur}}^{\mathcal{D},\sigma}(\mathbf{w})\|_2 \ge \frac{\bar{\mu}}{4\alpha\sqrt{\kappa}}\sqrt{\mathcal{L}_2^{\mathcal{D},\sigma}(\mathbf{w})}\,.$$

If we can assume that the set S_1 of Theorem D.1 is convex and that there is no local minima in the boundary of S_1, then by running any convex-optimization algorithm in the feasible set S_1, we guarantee that we converge either to a local minimum which has zero gradient or to a point inside the set $(S_2)^c$ where the true loss is sufficiently small. The next corollary shows that this is indeed the case for a distribution that satisfies Assumptions 2.3 and 2.4.

Corollary D.2. *Let $\mathcal{D}$ be a distribution supported on $\mathbb{R}^d \times \mathbb{R}$ and let $\sigma : \mathbb{R} \mapsto \mathbb{R}$ be an (α, β)-unbounded activation. Fix $\mathbf{w}^* \in \mathcal{W}^*$ and suppose that $\mathcal{D}_{\mathbf{x}}$ satisfies Assumptions 2.3 and 2.4 with respect to $\mathbf{w}^*$. Furthermore, let $C > 0$ be a sufficiently small absolute constant and let $\bar{\mu} = C\lambda^2\gamma\beta\rho/B$. Then, for any $\epsilon > 0$ and $\hat{\mathbf{w}} \in \mathcal{B}(2\|\mathbf{w}^*\|_2)$, so that $\mathcal{L}_{\mathrm{sur}}^{\mathcal{D},\sigma}(\hat{\mathbf{w}}) - \inf_{\mathbf{w}\in\mathcal{B}(2\|\mathbf{w}^*\|_2)} \mathcal{L}_{\mathrm{sur}}^{\mathcal{D},\sigma}(\mathbf{w}) \le \epsilon$, it holds*

$$\mathcal{L}_2^{\mathcal{D},\sigma}(\hat{\mathbf{w}}) \le O((\alpha B/(\rho\bar{\mu}))^2)(\mathcal{L}_2^{\mathcal{D},\sigma}(\mathbf{w}^*) + \alpha\epsilon) \ .$$

Proof of Corollary D.2. Denote $\mathcal{K}$ as the set of $\hat{\mathbf{w}}$ such that $\hat{\mathbf{w}} \in \mathcal{B}(2\|\mathbf{w}^*\|_2)$ and $\mathcal{L}_{\mathrm{sur}}^{\mathcal{D},\sigma}(\hat{\mathbf{w}}) - \inf_{\mathbf{w}\in\mathcal{B}(2\|\mathbf{w}^*\|_2)} \mathcal{L}_{\mathrm{sur}}^{\mathcal{D},\sigma}(\mathbf{w}) \le \epsilon$. First, note that as claimed in Fact C.4, $\mathbf{E}_{\mathbf{x}\sim\mathcal{D}_{\mathbf{x}}}[\mathbf{x}\mathbf{x}^\top] \preceq (5B/\rho)\mathbf{I}$ for any unit vector $\mathbf{u}$ when Assumption 2.4 holds.

Next, observe that the set of minimizers of the loss $\mathcal{L}_{\mathrm{sur}}^{\mathcal{D},\sigma}$ inside the ball $\mathcal{B}(2\|\mathbf{w}^*\|_2)$ is convex. Furthermore, the set $\mathcal{B}(2\|\mathbf{w}^*\|_2)$ is compact. Thus, for any point $\mathbf{w}' \in \mathcal{B}(2\|\mathbf{w}^*\|_2)$ that minimizes $\mathcal{L}_{\mathrm{sur}}^{\mathcal{D},\sigma}$ it will either hold that $\|\nabla\mathcal{L}_{\mathrm{sur}}^{\mathcal{D},\sigma}(\mathbf{w}')\|_2 = 0$ or $\mathbf{w}' \in \partial\mathcal{B}(2\|\mathbf{w}^*\|_2)$. Let $\mathcal{W}_{\mathrm{sur}}^*$ be the set of minimizers of $\mathcal{L}_{\mathrm{sur}}^{\mathcal{D},\sigma}$.

We first show that if there exists a minimizer $\mathbf{w}' \in \mathcal{W}_{\mathrm{sur}}^*$ such that $\mathbf{w}' \in \partial\mathcal{B}(2\|\mathbf{w}^*\|_2)$, then any point $\mathbf{w}$ inside the set $\mathcal{B}(2\|\mathbf{w}^*\|_2)$ gets error proportional to $\mathcal{L}_2^{\mathcal{D},\sigma}(\mathbf{w}^*)$. Observe for such point $\hat{\mathbf{w}}$, by the necessary condition of optimality, it should hold

$$\nabla\mathcal{L}_{\mathrm{sur}}^{\mathcal{D},\sigma}(\mathbf{w}') \cdot (\mathbf{w}' - \mathbf{w}) \le 0 \ , \tag{15}$$

for any $\mathbf{w} \in \mathcal{B}(2\|\mathbf{w}^*\|_2)$. Using Corollary D.4, we get that either $\nabla\mathcal{L}_{\mathrm{sur}}^{\mathcal{D},\sigma}(\mathbf{w}') \cdot (\mathbf{w}' - \mathbf{w}^*) \ge (\bar{\mu}/2)\|\mathbf{w}' - \mathbf{w}^*\|_2^2$ or $\mathbf{w}' \in \{\mathbf{w} : \|\mathbf{w} - \mathbf{w}^*\|_2^2 \le (20B/(\bar{\mu}^2\rho))\mathcal{L}_2^{\mathcal{D},\sigma}(\mathbf{w}^*)\}$. But Equation (15) contradicts with $\nabla\mathcal{L}_{\mathrm{sur}}^{\mathcal{D},\sigma}(\mathbf{w}') \cdot (\mathbf{w}' - \mathbf{w}^*) \ge (\bar{\mu}/2)\|\mathbf{w}' - \mathbf{w}^*\|_2^2 > 0$ since $\mathbf{w}' \in \partial\mathcal{B}(2\|\mathbf{w}^*\|_2)$, $\|\mathbf{w}'\|_2 = 2\|\mathbf{w}^*\|_2$ hence $\mathbf{w}' \ne \mathbf{w}^*$. So it must be the case that $\mathbf{w}' \in \{\mathbf{w} : \|\mathbf{w} - \mathbf{w}^*\|_2^2 \le (20B/(\bar{\mu}^2\rho))\mathcal{L}_2^{\mathcal{D},\sigma}(\mathbf{w}^*)\}$. Again, we have that $\mathbf{w}' \in \partial\mathcal{B}(2\|\mathbf{w}^*\|_2)$, therefore $\|\mathbf{w}' - \mathbf{w}^*\|_2 \ge \|\mathbf{w}^*\|_2$. Hence, $(20B/(\bar{\mu}^2\rho))\mathcal{L}_2^{\mathcal{D},\sigma}(\mathbf{w}^*) \ge \|\mathbf{w}^*\|_2^2 \ge (1/9)\|\mathbf{w} - \mathbf{w}^*\|_2^2$ for any $\mathbf{w} \in \mathcal{B}(2\|\mathbf{w}^*\|_2)$. Therefore, for any $\mathbf{w} \in \mathcal{B}(2\|\mathbf{w}^*\|_2)$, we have

$$\begin{aligned}
\mathcal{L}_2^{\mathcal{D},\sigma}(\mathbf{w}) &= \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}\left[(\sigma(\mathbf{w}\cdot\mathbf{x}) - y)^2\right] \\
&\le 2\mathcal{L}_2^{\mathcal{D},\sigma}(\mathbf{w}^*) + 2\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_{\mathbf{x}}}\left[(\sigma(\mathbf{w}\cdot\mathbf{x}) - \sigma(\mathbf{w}^*\cdot\mathbf{x}))^2\right] \\
&\le 2\mathcal{L}_2^{\mathcal{D},\sigma}(\mathbf{w}^*) + 10B\alpha^2/\rho\|\mathbf{w} - \mathbf{w}^*\|_2^2 \\
&\le O(B^2\alpha^2/(\bar{\mu}^2\rho^2))\mathcal{L}_2^{\mathcal{D},\sigma}(\mathbf{w}^*) \ ,
\end{aligned} \tag{16}$$

where in the second inequality we used the fact that $\mathbf{E}_{\mathbf{x}\sim\mathcal{D}_{\mathbf{x}}}\left[\mathbf{x}\mathbf{x}^\top\right] \preceq (5B/\rho)\mathbf{I}$ and σ is α-Lipschitz. Since the inequality above holds for any $\mathbf{w} \in \mathcal{B}(2\|\mathbf{w}^*\|_2)$, it will also be true for $\hat{\mathbf{w}} \in \mathcal{K} \subseteq \mathcal{B}(2\|\mathbf{w}^*\|_2)$.

It remains to consider the case where the minimizers $\mathcal{W}_{\mathrm{sur}}^*$ are strictly inside the $\mathcal{B}(2\|\mathbf{w}^*\|_2)$. Note that $\mathcal{L}_{\mathrm{sur}}^{\mathcal{D},\sigma}(\mathbf{w})$ is α-smooth. Therefore, we get that for any $\hat{\mathbf{w}} \in \mathcal{K}$, it holds $\|\nabla\mathcal{L}_{\mathrm{sur}}^{\mathcal{D},\sigma}(\hat{\mathbf{w}})\|_2^2 \le 2\alpha\epsilon$. By applying Corollary D.4, we get that either $\|\hat{\mathbf{w}} - \mathbf{w}^*\|_2^2 \le (20B/(\bar{\mu}^2\rho))\mathcal{L}_2^{\mathcal{D},\sigma}(\mathbf{w}^*)$ or that $\sqrt{2\alpha\epsilon} \ge (\bar{\mu}/2)\|\hat{\mathbf{w}} - \mathbf{w}^*\|_2$. Therefore we get that, $\|\hat{\mathbf{w}} - \mathbf{w}^*\|_2^2 \le (20B/(\bar{\mu}^2\rho))(\mathcal{L}_2^{\mathcal{D},\sigma}(\mathbf{w}^*) + \alpha\epsilon)$. Then the result follows from Equation (16).

□

To prove Theorem D.1, we need the following proposition which shows that if the current vector $\mathbf{w}$ is sufficiently far away from the true vector $\mathbf{w}^*$, then the gradient of the surrogate loss has a large component in the direction of $\mathbf{w} - \mathbf{w}^*$, in other words, the surrogate loss is sharp.

Proposition D.3. *Let $\mathcal{D}$ be a distribution supported on $\mathbb{R}^d \times \mathbb{R}$ and let $\sigma : \mathbb{R} \mapsto \mathbb{R}$ be an (α, β)-unbounded activation. Furthermore, assume that the maximum eigenvalue of the matrix $\mathbf{E}_{\mathbf{x}\sim\mathcal{D}_{\mathbf{x}}}[\mathbf{x}\mathbf{x}^\top]$ is $\kappa > 0$. Fix $\mathbf{w}^* \in \mathcal{W}^*$ and suppose $\mathcal{L}_{\mathrm{sur}}^{\mathcal{D},\sigma}$ is $\bar{\mu}$-sharp for some $\bar{\mu} > 0$ with respect to $\mathbf{w}^*$ in a nonempty subset $S_1 \subseteq \mathbb{R}^d$. Further, let $S_2 = \{\mathbf{w} : \|\mathbf{w} - \mathbf{w}^*\|_2^2 \ge 4(\kappa/\bar{\mu}^2)\mathcal{L}_2^{\mathcal{D},\sigma}(\mathbf{w}^*)\}$. Then for any $\mathbf{w} \in S_1 \cap S_2$, we have*

$$\nabla\mathcal{L}_{\mathrm{sur}}^{\mathcal{D},\sigma}(\mathbf{w}) \cdot (\mathbf{w} - \mathbf{w}^*) \ge (\bar{\mu}/2)\|\mathbf{w} - \mathbf{w}^*\|_2^2 \ .$$

Proof of Proposition D.3. We show that $\nabla \mathcal{L}_{\text{sur}}^{\mathcal{D},\sigma}(\mathbf{w}) \cdot (\mathbf{w} - \mathbf{w}^*)$ is bounded sufficiently far away from zero. We decompose the gradient into the noise-free part and the noisy, i.e., $\nabla \mathcal{L}_{\text{sur}}^{\mathcal{D},\sigma}(\mathbf{w}) = \nabla \bar{\mathcal{L}}_{\text{sur}}^{\mathcal{D},\sigma}(\mathbf{w}) + \nabla \mathcal{L}_{\text{sur}}^{\mathcal{D},\sigma}(\mathbf{w}^*)$. First, we bound the noisy term in the direction $\mathbf{w} - \mathbf{w}^*$, which yields

$$\begin{aligned}
&\nabla \mathcal{L}_{\text{sur}}^{\mathcal{D},\sigma}(\mathbf{w}^*) \cdot (\mathbf{w} - \mathbf{w}^*) \\
&\geq -\mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}[|\sigma(\mathbf{w}^* \cdot \mathbf{x}) - y||\mathbf{w} \cdot \mathbf{x} - \mathbf{w}^* \cdot \mathbf{x}|] \\
&\geq -\sqrt{\mathcal{L}_2^{\mathcal{D},\sigma}(\mathbf{w}^*)}\|\mathbf{w} - \mathbf{w}^*\|_2 \sqrt{\kappa} \,,
\end{aligned}$$

where we used the Cauchy-Schwarz inequality and that $\mathbf{E}_{\mathbf{x}\sim\mathcal{D}_{\mathbf{x}}}[\mathbf{x}\mathbf{x}^\top] \preceq \kappa \mathbf{I}$. Next, we bound the contribution of $\nabla \bar{\mathcal{L}}_{\text{sur}}^{\mathcal{D},\sigma}(\mathbf{w})$ in the direction $\mathbf{w} - \mathbf{w}^*$. Using the fact that $\bar{\mathcal{L}}_{\text{sur}}^{\mathcal{D},\sigma}(\mathbf{w})$ is $\bar{\mu}$-sharp for any $\mathbf{w} \in S_1$, it holds that

$$\nabla \bar{\mathcal{L}}_{\text{sur}}^{\mathcal{D},\sigma}(\mathbf{w}) \cdot (\mathbf{w} - \mathbf{w}^*) \geq \bar{\mu}\|\mathbf{w} - \mathbf{w}^*\|_2^2 \,.$$

Combining everything together we have that

$$\begin{aligned}
&\nabla \mathcal{L}_{\text{sur}}^{\mathcal{D},\sigma}(\mathbf{w}) \cdot (\mathbf{w} - \mathbf{w}^*) \\
&\geq \bar{\mu}\|\mathbf{w} - \mathbf{w}^*\|_2 \left(\|\mathbf{w} - \mathbf{w}^*\|_2 - (\sqrt{\kappa}/\bar{\mu})\sqrt{\mathcal{L}_2^{\mathcal{D},\sigma}(\mathbf{w}^*)} \right) \,.
\end{aligned}$$

The proof is completed by taking any $\mathbf{w} \in S_1 \cap S_2$, where $\|\mathbf{w} - \mathbf{w}^*\|_2 \geq (2\sqrt{\kappa}/\bar{\mu})\sqrt{\mathcal{L}_2^{\mathcal{D},\sigma}(\mathbf{w}^*)}$, and therefore

$$\nabla \mathcal{L}_{\text{sur}}^{\mathcal{D},\sigma}(\mathbf{w}) \cdot (\mathbf{w} - \mathbf{w}^*) \geq (\bar{\mu}/2)\|\mathbf{w} - \mathbf{w}^*\|_2^2 \,.$$

□

Corollary D.4. *Let $\mathcal{D}$ be a distribution supported on $\mathbb{R}^d \times \mathbb{R}$ and let $\sigma : \mathbb{R} \mapsto \mathbb{R}$ be an (α, β)-unbounded activation. Suppose that $\mathcal{D}_{\mathbf{x}}$ satisfies Assumptions 2.3 and 2.4 and let $C > 0$ be a sufficiently small absolute constant and let $\bar{\mu} = C\lambda^2\gamma\beta\rho/B$. Fix $\mathbf{w}^* \in \mathcal{W}^*$ and let $S = \mathcal{B}(2\|\mathbf{w}^*\|_2) - \{\mathbf{w} : \|\mathbf{w} - \mathbf{w}^*\|_2^2 \leq \frac{20B}{\bar{\mu}^2\rho}\mathcal{L}_2^{\mathcal{D},\sigma}(\mathbf{w}^*)\}$. Then, the surrogate loss $\mathcal{L}_{\text{sur}}^{\mathcal{D},\sigma}$ is $\bar{\mu}$-sharp in S, i.e.,*

$$\nabla \mathcal{L}_{\text{sur}}^{\mathcal{D},\sigma}(\mathbf{w}) \cdot (\mathbf{w} - \mathbf{w}^*) \geq (\bar{\mu}/2)\|\mathbf{w} - \mathbf{w}^*\|_2^2, \;\; \forall \mathbf{w} \in S.$$

Proof of Corollary D.4. Note that $\max_{\mathbf{u}\in\mathcal{B}(1)} \mathbf{E}_{\mathbf{x}\sim\mathcal{D}_{\mathbf{x}}}\left[(\mathbf{u} \cdot \mathbf{x})^2\right] = \kappa \leq 5B/\rho$ as proven in Fact C.4. Then combining Proposition D.3 and Lemma 2.5 we get the desired result. □

Proof of Theorem D.1. Using Proposition D.3, we get that for any $\mathbf{w} \in S' \cap S_1$, where $S' = \{\mathbf{w} : \|\mathbf{w} - \mathbf{w}^*\|_2^2 \geq 4(\kappa/\bar{\mu}^2)\mathcal{L}_2^{\mathcal{D},\sigma}(\mathbf{w}^*)\}$, we have that $\nabla \mathcal{L}_{\text{sur}}^{\mathcal{D},\sigma}(\mathbf{w}) \cdot (\mathbf{w} - \mathbf{w}^*) \geq (\bar{\mu}/2)\|\mathbf{w} - \mathbf{w}^*\|_2^2$. Note that

$$\begin{aligned}
\mathcal{L}_2^{\mathcal{D},\sigma}(\mathbf{w}) &= \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}\left[(\sigma(\mathbf{w} \cdot \mathbf{x}) - y)^2\right] \\
&\leq 2 \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_{\mathbf{x}}}\left[(\sigma(\mathbf{w} \cdot \mathbf{x}) - \sigma(\mathbf{w}^* \cdot \mathbf{x}))^2\right] + 2 \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}\left[(\sigma(\mathbf{w}^* \cdot \mathbf{x}) - y)^2\right] \\
&\leq 2\alpha^2\kappa\|\mathbf{w} - \mathbf{w}^*\|_2^2 + 2\mathcal{L}_2^{\mathcal{D},\sigma}(\mathbf{w}^*) \leq 2\alpha^2\kappa\|\mathbf{w} - \mathbf{w}^*\|_2^2 + (1/2)\mathcal{L}_2^{\mathcal{D},\sigma}(\mathbf{w}) \,,
\end{aligned}$$

where we used that $\mathcal{L}_2^{\mathcal{D},\sigma}(\mathbf{w}) \geq 4\mathcal{L}_2^{\mathcal{D},\sigma}(\mathbf{w}^*)$. Hence, it holds $4\alpha^2\kappa\|\mathbf{w} - \mathbf{w}^*\|_2^2 \geq \mathcal{L}_2^{\mathcal{D},\sigma}(\mathbf{w})$. Therefore, when $\mathbf{w} \in S_2$, it holds that $\|\mathbf{w} - \mathbf{w}^*\|_2^2 \geq (4\alpha\kappa/\bar{\mu})^2\mathcal{L}_2^{\mathcal{D},\sigma}(\mathbf{w})$, hence $S_2 \subseteq S'$.

Now observe that for any unit vector $\mathbf{v} \in \mathbb{R}^d$, it holds $\|\nabla \mathcal{L}_{\text{sur}}^{\mathcal{D},\sigma}(\mathbf{w})\|_2 \geq \mathbf{v} \cdot \nabla \mathcal{L}_{\text{sur}}^{\mathcal{D},\sigma}(\mathbf{w})$. Therefore, for any $\mathbf{w} \in S_1 \cap S_2 \subseteq S_1 \cap S'$, we have

$$\|\nabla \mathcal{L}_{\text{sur}}^{\mathcal{D},\sigma}(\mathbf{w})\|_2 \geq \nabla \mathcal{L}_{\text{sur}}^{\mathcal{D},\sigma}(\mathbf{w}) \cdot \left(\frac{\mathbf{w} - \mathbf{w}^*}{\|\mathbf{w} - \mathbf{w}^*\|_2} \right) \geq (\bar{\mu}/2)\|\mathbf{w} - \mathbf{w}^*\|_2 \geq \frac{\bar{\mu}}{4\alpha\sqrt{\kappa}}\sqrt{\mathcal{L}_2^{\mathcal{D},\sigma}(\mathbf{w})} \,.$$

We now show that the gradient is also bounded from above. By definition, we have

$$\begin{aligned}\|\nabla \mathcal{L}_{\text{sur}}^{\mathcal{D},\sigma}(\mathbf{w})\|_2 &= \left\| \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}} [(\sigma(\mathbf{w}\cdot\mathbf{x}) - y)\mathbf{x}] \right\|_2 \\ &\le \left\| \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_{\mathbf{x}}} [(\sigma(\mathbf{w}\cdot\mathbf{x}) - \sigma(\mathbf{w}^*\cdot\mathbf{x}))\mathbf{x}] \right\|_2 + \left\| \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}} [(\sigma(\mathbf{w}^*\cdot\mathbf{x}) - y)\mathbf{x}] \right\|_2 \\ &\le \max_{\|\mathbf{u}\|_2\le 1} \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_{\mathbf{x}}} [|\sigma(\mathbf{w}\cdot\mathbf{x}) - \sigma(\mathbf{w}^*\cdot\mathbf{x})||\mathbf{u}\cdot\mathbf{x}|] + \max_{\|\mathbf{v}\|_2\le 1} \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}} [|\sigma(\mathbf{w}^*\cdot\mathbf{x}) - y||\mathbf{v}\cdot\mathbf{x}|]\end{aligned}$$

Applying Cauchy-Schwarz to the inequality above, we further get

$$\begin{aligned}\|\nabla \mathcal{L}_{\text{sur}}^{\mathcal{D},\sigma}(\mathbf{w})\|_2 &\le \max_{\|\mathbf{u}\|_2\le 1} \sqrt{\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_{\mathbf{x}}} [|\sigma(\mathbf{w}\cdot\mathbf{x}) - \sigma(\mathbf{w}^*\cdot\mathbf{x})|^2] \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_{\mathbf{x}}} [|\mathbf{u}\cdot\mathbf{x}|^2]} \\ &\quad + \max_{\|\mathbf{v}\|_2\le 1} \sqrt{\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}} [|\sigma(\mathbf{w}^*\cdot\mathbf{x}) - y|^2] \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_{\mathbf{x}}} [|\mathbf{v}\cdot\mathbf{x}|^2]} \\ &\le \alpha\sqrt{\kappa}\|\mathbf{w} - \mathbf{w}^*\|_2 + \sqrt{\kappa \mathcal{L}_2^{\mathcal{D},\sigma}(\mathbf{w}^*)}\,,\end{aligned}$$

where in the last inequality we used the fact that σ is α-Lipschitz and that the maximum eigenvalue of $\mathbf{E}_{\mathbf{x}\sim\mathcal{D}_{\mathbf{x}}}[\mathbf{x}\mathbf{x}^\top]$ is κ. □

D.2. Fast Rates for Surrogate Loss

In this section, we proceed to show that when the surrogate loss is sharp, then applying batch Stochastic Gradient Descent (SGD) on the empirical surrogate loss obtains a C-approximate parameter $\hat{\mathbf{w}}$ of the L_2^2 loss in linear time. To be specific, consider the following iteration update

$$\mathbf{w}^{(t+1)} = \mathop{\operatorname{argmin}}_{\mathbf{w}\in\mathcal{B}(W)} \left\{ \mathbf{w}\cdot\mathbf{g}^{(t)} + \frac{1}{2\eta}\|\mathbf{w} - \mathbf{w}^{(t)}\|_2^2 \right\}, \tag{17}$$

where η is the step size and $\mathbf{g}^{(t)}$ is the empirical gradient of the surrogate loss:

$$\mathbf{g}^{(t)} = \frac{1}{N}\sum_{j=1}^{N} (\sigma(\mathbf{w}^{(t)}\cdot\mathbf{x}(j)) - y(j))\mathbf{x}(j). \tag{18}$$

The algorithm is summarized in Algorithm 2.

Algorithm 2 Stochastic Gradient Descent on Surrogate Loss

Input: Iterations: T, sample access from $\mathcal{D}$, batch size N, step size η, bound M.
Initialize $\mathbf{w}^{(0)} \leftarrow \mathbf{0}$.
for $t = 1$ **to** T **do**
  Draw N samples $\{(\mathbf{x}(j), y(j))\}_{j=1}^N \sim \mathcal{D}$.
  For each $j \in [N]$, $y(j) \leftarrow \operatorname{sign}(y(j))\min(|y(j)|, M)$.
  Calculate

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

  $\mathbf{w}^{(t+1)} \leftarrow \mathbf{w}^{(t)} - \eta\mathbf{g}^{(t)}$.
end for
Output: The weight vector $\mathbf{w}^{(T)}$.

Further, for simplicity of notation, we use $\bar{\mathbf{g}}^{(t)}$ to denote the empirical gradient of the noise-free surrogate loss:

$$\bar{\mathbf{g}}^{(t)} = \frac{1}{N}\sum_{j=1}^{N} (\sigma(\mathbf{w}^{(t)}\cdot\mathbf{x}(j)) - \sigma(\mathbf{w}^*\cdot\mathbf{x}(j)))\mathbf{x}(j). \tag{19}$$

In addition, we define the following helper functions H_2 and H_4.

Definition D.5. Let $\mathcal{D}_{\mathbf{x}}$ be a distribution on supported on $\mathbb{R}^d$ that satisfies Assumption 2.4 we define non-negative non-increasing functions H_2 and H_4 as follows:

$$H_2(r) \triangleq \max_{\mathbf{u}\in\mathcal{B}(1)} \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_{\mathbf{x}}} \left[(\mathbf{u}\cdot\mathbf{x})^2 \mathbb{1}\{|\mathbf{u}\cdot\mathbf{x}| \geq r\}\right],$$
$$H_4(r) \triangleq \max_{\mathbf{u}\in\mathcal{B}(1)} \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_{\mathbf{x}}} \left[(\mathbf{u}\cdot\mathbf{x})^4 \mathbb{1}\{|\mathbf{u}\cdot\mathbf{x}| \geq r\}\right].$$

Remark D.6. In particular, when $r = 0$, $H_2(0)$ and $H_4(0)$ bounds from above the second and fourth moments. Recall that in Fact C.4, it is proved that $H_2(0), H_4(0) \leq 5B/\rho$.

Now we state our main theorem.

Theorem D.7 (Main Algorithmic Result)**.** *Fix $\epsilon > 0$ and $W > 0$ and suppose Assumptions 2.2 to 2.4 hold. Let $\mu := \mu(\lambda, \gamma, \beta, \rho, B)$ be a sufficiently small constant multiple of $\lambda^2\gamma\beta\rho/B$, and let $M = \alpha W H_2^{-1}(\frac{\epsilon}{4\alpha^2 W^2})$. Further, choose parameter r_ϵ large enough so that $H_4(r_\epsilon)$ is a sufficiently small constant multiple of ϵ. Then after*

$$T = \widetilde{\Theta}\left(\frac{B^2\alpha^2}{\rho^2\mu^2}\log\left(\frac{W}{\epsilon}\right)\right)$$

iterations with batch size $N = \Omega(dT(r_\epsilon^2 + \alpha^2 M^2))$, Algorithm 2 converges to a point $\mathbf{w}^{(T)}$ such that

$$\mathcal{L}_2^{\mathcal{D},\sigma}(\mathbf{w}^{(T)}) = O\left(\frac{B^2\alpha^2}{\rho^2\mu^2}\mathrm{OPT}\right) + \epsilon\,,$$

with probability at least $2/3$.

We now provide a brief overview of the proof. As follows from Corollary D.2, when we find a vector $\hat{\mathbf{w}}$ that minimizes the surrogate loss, then this $\hat{\mathbf{w}}$ is itself a C-approximate solution of Problem 1.1. However, minimizing the surrogate loss can be expensive in computational and sample complexity. Corollary D.4 says that we can achieve strong-convexity-like rates as long as we are far away from a minimizer of the L_2^2 loss, i.e., when $\|\mathbf{w} - \mathbf{w}^*\|_2^2 \geq O(\mathrm{OPT})$. Roughly speaking, we would like to show that at each iteration t, it holds $||\mathbf{w}^{(t+1)} - \mathbf{w}^*||_2^2 \leq C||\mathbf{w}^{(t)} - \mathbf{w}^*||_2^2$ where $0 < C < 1$ is some constant depending on the parameters α, β, μ, ρ and B. Then since the distance from $\mathbf{w}^{(t)}$ to $\mathbf{w}^*$ contracts fast, we are able to get the linear convergence rate of the algorithm. To this end, we prove that under a sufficiently large batch size, the empirical gradient of the surrogate loss $\mathbf{g}^{(t)}$ approximates $\nabla\mathcal{L}_{\mathrm{sur}}^{\mathcal{D},\sigma}(\mathbf{w}^{(t)})$ with a small error. Thus, $||\mathbf{w}^{(t+1)} - \mathbf{w}^*||_2^2$ can be written as

$$||\mathbf{w}^{(t+1)} - \mathbf{w}^*||_2^2 = ||\mathbf{w}^{(t)} - \mathbf{w}^*||_2^2 - 2\eta\nabla\mathcal{L}_{\mathrm{sur}}^{\mathcal{D},\sigma}(\mathbf{w}^{(t)})\cdot(\mathbf{w}^{(t)} - \mathbf{w}^*) + (\mathrm{error}).$$

We then apply the sharpness property of the surrogate (Proposition D.3) to the inner product $\nabla\mathcal{L}_{\mathrm{sur}}^{\mathcal{D},\sigma}(\mathbf{w}^{(t)})\cdot(\mathbf{w}^{(t)} - \mathbf{w}^*)$, which as a result leads to $||\mathbf{w}^{(t+1)} - \mathbf{w}^*||_2^2 \leq (1 - 2\eta\mu)||\mathbf{w}^{(t)} - \mathbf{w}^*||_2^2 + (\mathrm{error})$. By choosing the parameters η and the batch size N carefully, one can show that

$$||\mathbf{w}^{(t+1)} - \mathbf{w}^*||_2^2 \leq (1 - C)||\mathbf{w}^{(t)} - \mathbf{w}^*||_2^2 + C'(\mathrm{OPT} + \epsilon),$$

indicating a fast contraction $||\mathbf{w}^{(t+1)} - \mathbf{w}^*||_2^2 \leq (1 - C/2)||\mathbf{w}^{(t)} - \mathbf{w}^*||_2^2$ whenever $C'(\mathrm{OPT} + \epsilon) \leq (C/2)||\mathbf{w}^{(t)} - \mathbf{w}^*||_2^2$.

To prove the theorem, we provide some supplementary lemmata. The following lemma states that we can truncate the labels y to $y' \leq M$, where M is a parameter determined by distribution $\mathcal{D}_{\mathbf{x}}$.

Lemma D.8. *Define $y' = \mathrm{sign}(y)\min(|y|, M)$ for $M = \alpha W H_2^{-1}(\frac{\epsilon}{4\alpha^2 W^2})$, then:*

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

meaning that we can consider y' instead of y and assume $|y| \leq M$ without loss of generality, where H_2 was defined in Definition D.5.

Proof of Lemma D.8. Fix $M > 0$, and denote $P : \mathbb{R} \to \mathbb{R}$ the operator that projects the points of $\mathbb{R}$ onto the interval $[-M, M]$, i.e., $P(t) = \mathrm{sign}(t)\min(|t|, M)$. To prove the aforementioned claim, we split the expectation into two events: the first event is when $|\mathbf{w}^* \cdot \mathbf{x}| \leq (M/\alpha)$ and the second when the latter is not true. Observe that in the first case, $P(\sigma(\mathbf{w}^* \cdot \mathbf{x})) = \sigma(\mathbf{w}^* \cdot \mathbf{x})$, hence, using the fact that P is non-expansive, we get

$$\begin{aligned}
&\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}\left[(\sigma(\mathbf{w}^* \cdot \mathbf{x}) - P(y))^2 \mathbb{1}\{|\mathbf{w}^* \cdot \mathbf{x}| \leq (M/\alpha)\}\right] \\
&= \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}\left[(P(\sigma(\mathbf{w}^* \cdot \mathbf{x})) - P(y))^2 \mathbb{1}\{|\mathbf{w}^* \cdot \mathbf{x}| \leq (M/\alpha)\}\right] \\
&\leq \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}\left[(\sigma(\mathbf{w}^* \cdot \mathbf{x} - y)^2 \mathbb{1}\{|\mathbf{w}^* \cdot \mathbf{x}| \leq (M/\alpha)\}\right] \\
&\leq \mathrm{OPT}\ .
\end{aligned}$$

It remains to bound the error in the event that $|\mathbf{w}^* \cdot \mathbf{x}| > (M/\alpha)$. In this event $\alpha|\mathbf{w}^* \cdot \mathbf{x}| \geq |P(y)|$, and so we have

$$\begin{aligned}
\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}\left[(\sigma(\mathbf{w}^* \cdot \mathbf{x}) - P(y))^2 \mathbb{1}\{|\mathbf{w}^* \cdot \mathbf{x}| > (M/\alpha)\}\right] &\leq 4\alpha^2 \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}\left[(\mathbf{w}^* \cdot \mathbf{x})^2 \mathbb{1}\{|\mathbf{w}^* \cdot \mathbf{x}| > (M/\alpha)\}\right] \\
&\leq 4\alpha^2 \|\mathbf{w}^*\|_2^2 H_2(M/(\alpha W)) \leq \epsilon\ ,
\end{aligned}$$

where in the first inequality we used the standard inequality $(a+b)^2 \leq 2(a^2+b^2)$ and that σ is α-Lipschitz hence $|\sigma(\mathbf{w}^* \cdot \mathbf{x})| = |\sigma(\mathbf{w}^* \cdot \mathbf{x}) - \sigma(0)| \leq \alpha|\mathbf{w}^* \cdot \mathbf{x}|$. □

Next, we show that the difference between the empirical gradients and the population gradients of the surrogate loss can be made small by choosing a large batch size N. Specifically, we have:

Lemma D.9. *Suppose N samples $\{(\mathbf{x}(j), y(j))\}_{j=1}^N$ are drawn from $\mathcal{D}$ independently and suppose Assumptions 2.2 to 2.4 hold. Let $\mathbf{g}^*$ be the empirical gradient of $\mathcal{L}_{\mathrm{sur}}^{\mathcal{D},\sigma}$ at $\mathbf{w}^*$ and let $\bar{\mathbf{g}}^t$ be the empirical gradient of $\bar{\mathcal{L}}_{\mathrm{sur}}^{\mathcal{D},\sigma}(\mathbf{w}^t)$, i.e.,*

$$\begin{aligned}
\mathbf{g}^* &= \frac{1}{N}\sum_{j=1}^N (\sigma(\mathbf{w}^* \cdot \mathbf{x}(j)) - y(j))\mathbf{x}(j), \\
\bar{\mathbf{g}}^{(t)} &= \frac{1}{N}\sum_{j=1}^N (\sigma(\mathbf{w}^{(t)} \cdot \mathbf{x}(j)) - \sigma(\mathbf{w}^* \cdot \mathbf{x}(j)))\mathbf{x}(j).
\end{aligned}$$

Moreover, let $H_4(r)$ be defined as in Definition D.5. Then for a fixed positive real number r_ϵ satisfying $H_4(r_\epsilon) \lesssim \epsilon$ and $r_\epsilon \geq 1$, we have the following bounds holds with probability at least $1-\delta$:

$$\|\mathbf{g}^* - \nabla\mathcal{L}_{\mathrm{sur}}^{\mathcal{D},\sigma}(\mathbf{w}^*)\|_2 \lesssim \sqrt{\frac{d(r_\epsilon^2 \mathrm{OPT} + \alpha^2 M^2 \epsilon)}{\delta N}}, \tag{20}$$

and similarly:

$$\|\bar{\mathbf{g}}^{(t)} - \nabla\bar{\mathcal{L}}_{\mathrm{sur}}^{\mathcal{D},\sigma}(\mathbf{w}^{(t)})\|_2 \lesssim \sqrt{\frac{\alpha^2 dB}{\delta\rho N}}\|\mathbf{w}^{(t)} - \mathbf{w}^*\|_2. \tag{21}$$

Proof of Lemma D.9. The proof follows from a direct application of Markov's inequality and a careful bound on the variance term using the tail-bound assumptions. To be specific, by Markov's Inequality, for any $\xi > 0$ it holds:

$$\mathbf{Pr}\left[\|\mathbf{g}^* - \nabla\mathcal{L}_{\mathrm{sur}}^{\mathcal{D},\sigma}(\mathbf{w}^*)\|_2 \geq \xi\right] = \mathbf{Pr}\left[\|\mathbf{g}^* - \nabla\mathcal{L}_{\mathrm{sur}}^{\mathcal{D},\sigma}(\mathbf{w}^*)\|_2^2 \geq \xi^2\right] \leq \frac{1}{\xi^2}\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}\left[\|\mathbf{g}^* - \nabla\mathcal{L}_{\mathrm{sur}}^{\mathcal{D},\sigma}(\mathbf{w}^*)\|_2^2\right].$$

Now for the variance term $\mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}\left[||\mathbf{g}^* - \nabla\mathcal{L}_{\mathrm{sur}}^{\mathcal{D},\sigma}(\mathbf{w}^*)||_2^2\right]$, recall that each sample $\mathbf{x}(j)$ and $y(j)$ are i.i.d., therefore, we

can bound it in the following way

$$
\begin{aligned}
&\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}\left[\|\mathbf{g}^* - \nabla\mathcal{L}_{\text{sur}}^{\mathcal{D},\sigma}(\mathbf{w}^*)\|_2^2\right] \\
&= \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}\left[\frac{1}{N^2}\left\|\sum_{j=1}^{N}\left((\sigma(\mathbf{w}^*\cdot\mathbf{x}(j)) - y(j))\mathbf{x}(j) - \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}\left[(\sigma(\mathbf{w}^*\cdot\mathbf{x}(j)) - y(j))\mathbf{x}(j)\right]\right)\right\|_2^2\right] \\
&= \frac{1}{N}\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}\left[\left\|(\sigma(\mathbf{w}^*\cdot\mathbf{x}) - y)\mathbf{x} - \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}\left[(\sigma(\mathbf{w}^*\cdot\mathbf{x}) - y)\mathbf{x}\right]\right\|_2^2\right] \\
&\leq \frac{1}{N}\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}\left[\|(\sigma(\mathbf{w}^*\cdot\mathbf{x}) - y)\mathbf{x}\|_2^2\right], \qquad (22)
\end{aligned}
$$

where in the second equation we used that for any mean-zero independent random variables $\mathbf{z}_j$, we have $\mathbf{E}[\|\sum_j \mathbf{z}_j\|_2^2] = \sum_j \mathbf{E}[\|\mathbf{z}_j\|_2^2]$, and in the final inequality we used that for any random variable X, it holds $\mathbf{E}[\|X - \mathbf{E}[X]\|_2^2] \leq \mathbf{E}[\|X\|_2^2]$.

Next, we show that $\mathbf{E}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}\left[\|(\sigma(\mathbf{w}^*\cdot\mathbf{x}) - y)\mathbf{x}\|_2^2\right]$ can be bounded above in terms of H_2 and H_4.

Claim D.10. $\mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}\left[\|(\sigma(\mathbf{w}^*\cdot\mathbf{x}) - y)\mathbf{x}\|_2^2\right] \lesssim d(r_\epsilon^2\mathrm{OPT} + \alpha^2 M^2 H_4(r_\epsilon))$.

Proof of Claim D.10. To prove the claim, note that $\|\mathbf{x}\|_2^2 = \sum_{i=1}^d |\mathbf{x}_i|^2$, therefore by linearity of expectation it holds

$$
\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}\left[\|(\sigma(\mathbf{w}^*\cdot\mathbf{x}) - y)\mathbf{x}\|_2^2\right] = \sum_{i=1}^{d}\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}\left[(\sigma(\mathbf{w}^*\cdot\mathbf{x}) - y)^2\mathbf{x}_i^2\right].
$$

Thus, the goal is to bound $\mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}\left[(\sigma(\mathbf{w}^*\cdot\mathbf{x}) - y)^2\mathbf{x}_i^2\right]$ effectively for each entry i. Deploying the intuition that the probability of $|\mathbf{x}_i| = |\mathbf{e}_i\cdot\mathbf{x}|$ being very large is tiny since we have $\mathbf{Pr}\left[|\mathbf{e}_i\cdot\mathbf{x}| > r\right] \leq h(r)$ and $h(r) \leq Br^{-(4+\rho)}$ by the Assumption 2.4, we fix some large r_ϵ and bound the expectation by looking separately at the events that $|\mathbf{x}_i| \leq r_\epsilon$ and $|\mathbf{x}_i| > r_\epsilon$, i.e.,

$$
\begin{aligned}
\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}\left[(\sigma(\mathbf{w}^*\cdot\mathbf{x}) - y)^2\mathbf{x}_i^2\right] &= \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}\left[(\sigma(\mathbf{w}^*\cdot\mathbf{x}) - y)^2\mathbf{x}_i^2\mathbb{1}\{|\mathbf{x}_i| \leq r_\epsilon\}\right] \\
&\quad + \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}\left[(\sigma(\mathbf{w}^*\cdot\mathbf{x}) - y)^2\mathbf{x}_i^2\mathbb{1}\{|\mathbf{x}_i| > r_\epsilon\}\right]. \qquad (23)
\end{aligned}
$$

Note when conditioned on the event $|\mathbf{x}_i| \leq r_\epsilon$ the bound follows easily as:

$$
\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}\left[(\sigma(\mathbf{w}^*\cdot\mathbf{x}) - y)^2\mathbf{x}_i^2\mathbb{1}\{|\mathbf{x}_i| \leq r_\epsilon\}\right] \leq r_\epsilon^2\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}\left[(\sigma(\mathbf{w}^*\cdot\mathbf{x}) - y)^2\right] = r_\epsilon^2\mathrm{OPT}. \qquad (24)
$$

When considering $|\mathbf{x}_i| > r_\epsilon$, notice that σ is α-Lipschitz and that $\sigma(0) = 0$, as well as that we assumed $|y| \leq M$ due to Lemma D.8, therefore, denoting $\mathbf{u}_{\mathbf{w}^*} = \mathbf{w}^*/\|\mathbf{w}^*\|_2$, it holds:

$$
\begin{aligned}
\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}\left[(\sigma(\mathbf{w}^*\cdot\mathbf{x}) - y)^2\mathbf{x}_i^2\mathbb{1}\{|\mathbf{x}_i| > r_\epsilon\}\right] &\leq 2\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}\left[((\sigma(\mathbf{w}^*\cdot\mathbf{x}))^2 + y^2)\mathbf{x}_i^2\mathbb{1}\{|\mathbf{x}_i| > r_\epsilon\}\right] \\
&\leq 2\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}\left[(\alpha^2(\mathbf{w}^*\cdot\mathbf{x})^2 + M^2)\mathbf{x}_i^2\mathbb{1}\{|\mathbf{x}_i| > r_\epsilon\}\right] \\
&\leq 2\alpha^2\|\mathbf{w}^*\|_2^2\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}\left[(\mathbf{u}_{\mathbf{w}^*}\cdot\mathbf{x})^2\mathbf{x}_i^2\mathbb{1}\{|\mathbf{x}_i| > r_\epsilon\}\right] + 2M^2H_2(r_\epsilon),
\end{aligned}
$$

where in the last inequality we used Definition D.5. For the first term above, note that $\mathbf{u}_{\mathbf{w}^*}$ is also a unit vector, so by Assumption 2.4 the probability mass of $|\mathbf{u}_{\mathbf{w}^*}\cdot\mathbf{x}| > r_\epsilon$ is also small, thus, we can show that $\mathbf{E}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}\left[(\mathbf{u}_{\mathbf{w}^*}\cdot\mathbf{x})^2\mathbf{x}_i^2\mathbb{1}\{|\mathbf{x}_i| > r_\epsilon\}\right]$ is dominated by $r_\epsilon^2\,\mathbf{E}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}\left[\mathbf{x}_i^2\mathbb{1}\{|\mathbf{x}_i| > r_\epsilon\}\right]$, which can then be bounded above by H_2 and H_4. In detail, we split the expectation by conditioning on the events that $|\mathbf{u}_{\mathbf{w}^*}\cdot\mathbf{x}| > r_\epsilon$ and $|\mathbf{u}_{\mathbf{w}^*}\cdot\mathbf{x}| \leq r_\epsilon$, then noticing that $\mathbb{1}\{|\mathbf{x}_i| > r_\epsilon, |\mathbf{u}_{\mathbf{w}^*}\cdot\mathbf{x}| \leq$

$r_\epsilon\} \le \mathbb{1}\{|\mathbf{x}_i| \ge r_\epsilon\}$, we get:

$$\begin{aligned}
\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}\left[(\mathbf{u}_{\mathbf{w}^*}\cdot\mathbf{x})^2\mathbf{x}_i^2\mathbb{1}\{|\mathbf{x}_i|>r_\epsilon\}\right] &\le \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}\left[r_\epsilon^2\mathbf{x}_i^2\mathbb{1}\{|\mathbf{x}_i|>r_\epsilon, |\mathbf{u}_{\mathbf{w}^*}\cdot\mathbf{x}|\le r_\epsilon\}\right] \\
&\quad + \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}\left[(\mathbf{u}_{\mathbf{w}^*}\cdot\mathbf{x})^2\mathbf{x}_i^2\mathbb{1}\{|\mathbf{x}_i|>r_\epsilon, |\mathbf{u}_{\mathbf{w}^*}\cdot\mathbf{x}|> r_\epsilon\}\right] \\
&\le \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}\left[r_\epsilon^2\mathbf{x}_i^2\mathbb{1}\{|\mathbf{x}_i|>r_\epsilon\}\right] \\
&\quad + \sqrt{\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}\left[(\mathbf{u}_{\mathbf{w}^*}\cdot\mathbf{x})^4\mathbb{1}\{|\mathbf{u}_{\mathbf{w}^*}\cdot\mathbf{x}|>r_\epsilon\}\right]\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}\left[\mathbf{x}_i^4\mathbb{1}\{|\mathbf{x}_i|>r_\epsilon\}\right]} \\
&\le r_\epsilon^2 H_2(r_\epsilon) + H_4(r_\epsilon), \qquad (25)
\end{aligned}$$

where the second inequality comes from Cauchy-Schwarz and in the last inequality we applied $H_4(r_\epsilon) \ge \mathbf{E}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}\left[(\mathbf{u}\cdot\mathbf{x})^4\mathbb{1}\{|\mathbf{u}\cdot\mathbf{x}|\ge r_\epsilon\}\right]$ for any $\mathbf{u}\in\mathcal{B}(1)$ by Definition D.5. Now plugging Equation (25) to the bound we get for $\mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}\left[\sigma(\mathbf{w}^*\cdot\mathbf{x})-y)^2\mathbf{x}_i^2\mathbb{1}\{|\mathbf{x}_i|\ge r_\epsilon\}\right]$, we have:

$$\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}\left[(\sigma(\mathbf{w}^*\cdot\mathbf{x})-y)^2\mathbf{x}_i^2\mathbb{1}\{|\mathbf{x}_i|\ge r_\epsilon\}\right] \le 2\alpha^2\|\mathbf{w}^*\|_2^2(r_\epsilon^2 H_2(r_\epsilon)+H_4(r_\epsilon)) + 2M^2H_2(r_\epsilon).$$

Further recall that by definition:

$$H_4(r) = \max_{\mathbf{u}\in\mathcal{B}(1)}\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}\left[(\mathbf{u}\cdot\mathbf{x})^4\mathbb{1}\{|\mathbf{u}\cdot\mathbf{x}|\ge r\}\right] \ge \max_{\mathbf{u}\in\mathcal{B}(1)} r^2\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}\left[(\mathbf{u}\cdot\mathbf{x})^2\mathbb{1}\{|\mathbf{u}\cdot\mathbf{x}|\ge r\}\right] = r^2H_2(r),$$

hence $H_4(r)\ge H_2(r)$ when $r\ge 1$. Then applying these facts along with the fact that $\|\mathbf{w}^*\|_2\le M$ simplifies the inequality above to the following:

$$\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}\left[(\sigma(\mathbf{w}^*\cdot\mathbf{x})-y)^2\mathbf{x}_i^2\mathbb{1}\{|\mathbf{x}_i|\ge r_\epsilon\}\right] \lesssim \alpha^2M^2H_4(r_\epsilon). \qquad (26)$$

Combining Equation (26) and Equation (24) with Equation (23), we get:

$$\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}\left[(\sigma(\mathbf{w}^*\cdot\mathbf{x})-y)^2\|\mathbf{x}\|_2^2\right] \lesssim d(r_\epsilon^2\mathrm{OPT}+\alpha^2M^2H_4(r_\epsilon)),$$

proving the desired claim. □

Plugging Claim D.10 above back to Equation (22), we immediately get:

$$\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}\left[\|\mathbf{g}^*-\nabla\mathcal{L}_{\mathrm{sur}}^{\mathcal{D},\sigma}(\mathbf{w}^*)\|_2^2\right] \lesssim \frac{d}{N}(r_\epsilon^2\mathrm{OPT}+\alpha^2M^2\epsilon),$$

given that $H_4(r_\epsilon)\lesssim\epsilon$. Then choosing $\xi\gtrsim\sqrt{\frac{d}{\delta N}(r_\epsilon^2\mathrm{OPT}+\alpha^2M^2\epsilon)}$, we get Equation (20):

$$\mathbf{Pr}\left[\|\mathbf{g}^*-\nabla\mathcal{L}_{\mathrm{sur}}^{\mathcal{D},\sigma}(\mathbf{w}^*)\|_2 \gtrsim \sqrt{\frac{d}{\delta N}(r_\epsilon^2+\alpha^2M^2)\mathrm{OPT}}\right] \le \delta.$$

For Equation (21), we repeat the steps when proving Equation (20). Using Markov inequality again, we have

$$\begin{aligned}
\mathbf{Pr}\left[\|\bar{\mathbf{g}}^{(t)}-\nabla\bar{\mathcal{L}}_{\mathrm{sur}}^{\mathcal{D},\sigma}(\mathbf{w}^{(t)})\|_2\ge\zeta\right] &= \mathbf{Pr}\left[\|\bar{\mathbf{g}}^{(t)}-\nabla\bar{\mathcal{L}}_{\mathrm{sur}}^{\mathcal{D},\sigma}(\mathbf{w}^{(t)})\|_2^2\ge\zeta^2\right] \\
&\le \frac{1}{\zeta^2}\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}\left[\|\bar{\mathbf{g}}^{(t)}-\nabla\bar{\mathcal{L}}_{\mathrm{sur}}^{\mathcal{D},\sigma}(\mathbf{w}^{(t)})\|_2^2\right].
\end{aligned}$$

The goal is to bound the expectation of the squared norm. Notice that $(\mathbf{x}(j),y(j))\sim\mathcal{D}$ are i.i.d. samples, therefore, it holds:

$$\begin{aligned}
&\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}\left[\|\bar{\mathbf{g}}^{(t)}-\nabla\bar{\mathcal{L}}_{\mathrm{sur}}^{\mathcal{D},\sigma}(\mathbf{w}^{(t)})\|_2^2\right] \\
&= \frac{1}{N^2}\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}\left[\left\|\sum_{j=1}^N\left((\sigma(\mathbf{w}^{(t)}\cdot\mathbf{x}(j))-\sigma(\mathbf{w}^*\cdot\mathbf{x}(j)))\mathbf{x}(j)-\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}\left[(\sigma(\mathbf{w}^{(t)}\cdot\mathbf{x}(j))-\sigma(\mathbf{w}^*\cdot\mathbf{x}(j)))\mathbf{x}(j)\right]\right)\right\|_2^2\right] \\
&= \frac{1}{N}\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}\left[\|(\sigma(\mathbf{w}^{(t)}\cdot\mathbf{x})-\sigma(\mathbf{w}^*\cdot\mathbf{x}))\mathbf{x}-\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}\left[(\sigma(\mathbf{w}^{(t)}\cdot\mathbf{x})-\sigma(\mathbf{w}^*\cdot\mathbf{x}))\mathbf{x}\right]\|_2^2\right],
\end{aligned}$$

because for any i.i.d. zero-mean random variables $\mathbf{z}(j)$ it holds $\mathbf{E}[||\sum_j \mathbf{z}(j)||_2^2] = \sum_j \mathbf{E}[||\mathbf{z}(j)||_2^2]$. Note that $\mathbf{E}[||\mathbf{z} - \mathbf{E}[\mathbf{z}]||_2^2] \leq \mathbf{E}[||\mathbf{z}||_2^2]$, therefore, we can further bound the variance of $\bar{\mathbf{g}}^t - \nabla\bar{\mathcal{L}}_{\text{sur}}^{\mathcal{D},\sigma}(\mathbf{w}^t)$ as:

$$\begin{aligned}
\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}\left[\|\bar{\mathbf{g}}^{(t)} - \nabla\bar{\mathcal{L}}_{\text{sur}}^{\mathcal{D},\sigma}(\mathbf{w}^{(t)})\|_2^2\right] &\leq \frac{1}{N}\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}\left[(\sigma(\mathbf{w}^{(t)}\cdot\mathbf{x}) - \sigma(\mathbf{w}^*\cdot\mathbf{x}))^2\|\mathbf{x}\|_2^2\right] \\
&= \frac{1}{N}\sum_{i=1}^d \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}\left[(\sigma(\mathbf{w}^{(t)}\cdot\mathbf{x}) - \sigma(\mathbf{w}^*\cdot\mathbf{x}))^2\mathbf{x}_i^2\right] \\
&\leq \frac{\alpha^2}{N}\sum_{i=1}^d \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}\left[(\mathbf{w}^{(t)}\cdot\mathbf{x} - \mathbf{w}^*\cdot\mathbf{x})^2\mathbf{x}_i^2\right],
\end{aligned} \tag{27}$$

where in the last inequality we used $|\sigma(\mathbf{w}^{(t)}\cdot\mathbf{x}) - \sigma(\mathbf{w}^*\cdot\mathbf{x})| \leq \alpha|\mathbf{w}^{(t)}\cdot\mathbf{x} - \mathbf{w}^*\cdot\mathbf{x}|$, as σ is α-Lipschitz.

It remains to bound $\mathbf{E}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}\left[(\mathbf{w}^{(t)}\cdot\mathbf{x} - \mathbf{w}^*\cdot\mathbf{x})^2\mathbf{x}_i^2\right]$. Denote $\mathbf{u}_{\mathbf{w}^{(t)}} = (\mathbf{w}^{(t)} - \mathbf{w}^*)/\|\mathbf{w}^{(t)} - \mathbf{w}^*\|$, which is a unit vector. Abstracting $\|\mathbf{w}^t - \mathbf{w}^*\|_2$ from the expectation then applying Cauchy-Schwarz, we get:

$$\begin{aligned}
\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}\left[(\mathbf{w}^{(t)}\cdot\mathbf{x} - \mathbf{w}^*\cdot\mathbf{x})^2\mathbf{x}_i^2\right] &= \|\mathbf{w}^{(t)} - \mathbf{w}^*\|_2^2 \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}\left[(\mathbf{u}_{\mathbf{w}^{(t)}}\cdot\mathbf{x})^2\mathbf{x}_i^2\right] \\
&\leq \|\mathbf{w}^{(t)} - \mathbf{w}^*\|_2^2\sqrt{\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}\left[(\mathbf{u}_{\mathbf{w}^{(t)}}\cdot\mathbf{x})^4\right]\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}\left[\mathbf{x}_i^4\right]} \\
&\leq \|\mathbf{w}^{(t)} - \mathbf{w}^*\|_2^2 H_4(0),
\end{aligned}$$

where the last inequality comes from $H_4(0) = \max_{\mathbf{u}\in\mathcal{B}(1)}\mathbf{E}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}\left[(\mathbf{u}\cdot\mathbf{x})^4\right]$, which holds by definition. Further from Fact C.4, $H_4(0) \leq 5B/\rho$, thus, we get

$$\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}\left[(\mathbf{w}^{(t)}\cdot\mathbf{x} - \mathbf{w}^*\cdot\mathbf{x})^2\mathbf{x}_i^2\right] \leq \frac{5B}{\rho}\|\mathbf{w}^{(t)} - \mathbf{w}^*\|_2^2. \tag{28}$$

To sum up, plugging Equation (28) back to Equation (27), we have:

$$\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}\left[\|\bar{\mathbf{g}}^{(t)} - \nabla\bar{\mathcal{L}}_{\text{sur}}^{\mathcal{D},\sigma}(\mathbf{w}^{(t)})\|_2^2\right] \lesssim \frac{\alpha^2 dB}{\rho N}\|\mathbf{w}^{(t)} - \mathbf{w}^*\|_2^2.$$

Finally, choosing ζ to be a sufficiently small multiple of $\sqrt{\frac{\alpha^2 dB}{\delta\rho N}}\|\mathbf{w}^{(t)} - \mathbf{w}^*\|_2$, Equation (21) follows.

□

Corollary D.11. *Suppose N samples $\{(\mathbf{x}(j), y(j))\}_{j=1}^N$ are drawn from $\mathcal{D}$ independently and suppose Assumptions 2.2 to 2.4 hold. Let $\mathbf{g}^{(t)}$ be the empirical gradient of $\mathcal{L}_{\text{sur}}^{\mathcal{D},\sigma}(\mathbf{w}^{(t)})$. Moreover, let $H_4(r)$ be defined as in Definition D.5. Then for a fixed positive real number r_ϵ satisfying $H_4(r_\epsilon) \lesssim \epsilon$ and $r_\epsilon \geq 1$, with probability at least $1-\delta$ it holds*

$$\|\mathbf{g}^{(t)} - \nabla\mathcal{L}_{\text{sur}}^{\mathcal{D},\sigma}(\mathbf{w}^{(t)})\|_2 \lesssim \sqrt{\frac{d\alpha^2 B}{\delta\rho N}}\left(\|\mathbf{w}^{(t)} - \mathbf{w}^*\|_2 + \sqrt{r_\epsilon^2\text{OPT} + M^2\epsilon}\right). \tag{29}$$

Corollary D.12. *Let $\mathcal{D}$ be a distribution in $\mathbb{R}^d \times \mathbb{R}$ and suppose Assumptions 2.2 to 2.4 hold. Moreover, let $H_4(r)$ be defined as in Definition D.5. Fix a positive real number r_ϵ satisfying $H_4(r_\epsilon) \lesssim \epsilon$ and $r_\epsilon \geq 1$. It holds that*

$$\mathbf{E}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}\left[\|\nabla\mathcal{L}_{\text{sur}}^{\mathcal{D},\sigma}(\mathbf{w}^{(t)})\|_2^2\right] \lesssim \frac{d\alpha^2 B}{\rho}\left(\|\mathbf{w}^{(t)} - \mathbf{w}^*\|_2^2 + r_\epsilon^2\text{OPT} + M^2\epsilon\right). \tag{30}$$

We further show that the norm of empirical gradients $\mathbf{g}^*$ and $\bar{\mathbf{g}}^{(t)}$ can be bounded with respect to OPT, ϵ and $\|\mathbf{w}^{(t)} - \mathbf{w}^*\|_2$.

Corollary D.13. *Suppose the conditions in Lemma D.9 are satisfied. Fix $r_\epsilon \geq 1$ such that $H_4(r_\epsilon)$ is a sufficiently small multiple of ϵ. Then with probability at least $1-\delta$, we have:*

$$\|\mathbf{g}^*\|_2 \lesssim \sqrt{(B/\rho)\text{OPT}} + \sqrt{\frac{d(r_\epsilon^2\text{OPT} + \alpha^2 M^2\epsilon)}{\delta N}}, \tag{31}$$

and

$$\|\bar{\mathbf{g}}^{(t)}\|_2 \lesssim \frac{\alpha B}{\rho}\left(1+\sqrt{\frac{d\rho}{\delta BN}}\right)\|\mathbf{w}^{(t)}-\mathbf{w}^*\|_2. \tag{32}$$

Proof. We first estimate the norm of $\nabla\mathcal{L}_{\text{sur}}^{\mathcal{D},\sigma}(\mathbf{w}^*)$ and $\nabla\bar{\mathcal{L}}_{\text{sur}}^{\mathcal{D},\sigma}(\mathbf{w}^{(t)})$. For the former, applying the Cauchy-Schwarz inequality, we get:

$$\begin{aligned}
\|\nabla\mathcal{L}_{\text{sur}}^{\mathcal{D},\sigma}(\mathbf{w}^*)\|_2 &= \|\,\mathbf{E}[(\sigma(\mathbf{w}^*\cdot\mathbf{x})-y)\mathbf{x}]\|_2 \\
&= \max_{\|\mathbf{u}\|_2=1}\mathbf{E}[(\sigma(\mathbf{w}^*\cdot\mathbf{x})-y)\mathbf{u}\cdot\mathbf{x}] \\
&\le \max_{\|\mathbf{u}\|_2=1}\sqrt{\mathbf{E}[(\sigma(\mathbf{w}^*\cdot\mathbf{x})-y)^2]\,\mathbf{E}[(\mathbf{u}\cdot\mathbf{x})^2]} \\
&\le \sqrt{\frac{5B}{\rho}\mathrm{OPT}}\,,
\end{aligned}$$

where we used that $H_2(0)\le 5B/\rho$ from Fact C.4. In addition, by Lemma D.9, with probability at least $1-\delta$, we have:

$$\|\mathbf{g}^*-\nabla\mathcal{L}_{\text{sur}}^{\mathcal{D},\sigma}(\mathbf{w}^*)\|_2 \lesssim \sqrt{\frac{d}{\delta N}(r_\epsilon^2\mathrm{OPT}+\alpha^2M^2\epsilon)},$$

given that r_ϵ is chosen large enough so that $H_4(r_\epsilon)\lesssim\epsilon$. Then combining with the bound of $\|\nabla\mathcal{L}_{\text{sur}}^{\mathcal{D},\sigma}(\mathbf{w}^*)\|_2$ above, it holds:

$$\|\mathbf{g}^*\|_2 \lesssim \sqrt{\frac{B}{\rho}\mathrm{OPT}}+\sqrt{\frac{d(r_\epsilon^2\mathrm{OPT}+\alpha^2M^2\epsilon)}{\delta N}}.$$

For the second claim, following the exact same approach and utilizing the fact that σ is α-Lipschitz continuous again, we have:

$$\begin{aligned}
\|\nabla\bar{\mathcal{L}}_{\text{sur}}^{\mathcal{D},\sigma}(\mathbf{w}^{(t)})\|_2 &= \|\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}\left[(\sigma(\mathbf{w}^{(t)}\cdot\mathbf{x})-\sigma(\mathbf{w}^*\cdot\mathbf{x}))\mathbf{x}\right]\|_2 \\
&= \max_{\|\mathbf{u}\|_2=1}\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}\left[|\sigma(\mathbf{w}^{(t)}\cdot\mathbf{x})-\sigma(\mathbf{w}^*\cdot\mathbf{x})|\mathbf{u}\cdot\mathbf{x}\right] \\
&\le \alpha\max_{\|\mathbf{u}\|_2=1}\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}\left[|(\mathbf{w}^{(t)}-\mathbf{w}^*)\cdot\mathbf{x}|\mathbf{u}\cdot\mathbf{x}\right].
\end{aligned}$$

Applying Cauchy-Schwarz inequality, we have

$$\|\nabla\bar{\mathcal{L}}_{\text{sur}}^{\mathcal{D},\sigma}(\mathbf{w}^{(t)})\|_2 \le \alpha\max_{\|\mathbf{u}\|_2=1}\sqrt{\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}\left[((\mathbf{w}^{(t)}-\mathbf{w}^*)\cdot\mathbf{x})^2\right]\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}\left[(\mathbf{u}\cdot\mathbf{x})^2\right]} \le \frac{5\alpha B}{\rho}\|\mathbf{w}^{(t)}-\mathbf{w}^*\|_2.$$

Then combining with Equation (21), we get the desired claim:

$$\|\bar{\mathbf{g}}^{(t)}\|_2 \lesssim \frac{\alpha B}{\rho}\left(1+\sqrt{\frac{d\rho}{\delta BN}}\right)\|\mathbf{w}^{(t)}-\mathbf{w}^*\|_2.$$

□

Finally, we can turn to the proof of Theorem D.7.

Proof of Theorem D.7. Recall that for a vector $\hat{\mathbf{w}}$, we have

$$\begin{aligned}
\mathcal{L}_2^{\mathcal{D},\sigma}(\hat{\mathbf{w}}) = \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}\left[(\sigma(\hat{\mathbf{w}}\cdot\mathbf{x})-y)^2\right] &\le 2\mathcal{L}_2^{\mathcal{D},\sigma}(\mathbf{w}^*)+2\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}\left[(\sigma(\hat{\mathbf{w}}\cdot\mathbf{x})-\sigma(\mathbf{w}^*\cdot\mathbf{x}))^2\right] \\
&\le 2\mathcal{L}_2^{\mathcal{D},\sigma}(\mathbf{w}^*)+10B\alpha^2/\rho\|\hat{\mathbf{w}}-\mathbf{w}^*\|_2^2,
\end{aligned}$$

where in the last inequality we used the fact that σ is α-Lipschitz and $\mathbf{E}_{\mathbf{x}\sim\mathcal{D}_{\mathbf{x}}}[\mathbf{x}\mathbf{x}^\top] \preceq (5B/\rho)I$ according to Fact C.4. Thus when the algorithm generates some $\hat{\mathbf{w}}$ such that $\|\hat{\mathbf{w}} - \mathbf{w}^*\|_2^2 \leq \epsilon'$, it holds

$$\mathcal{L}_2^{\mathcal{D},\sigma}(\mathbf{w}^{(T)}) \leq 2\mathrm{OPT} + (10B\alpha^2/\rho)\epsilon' \tag{33}$$

yielding a C-approximate solution to the Problem 1.1. Therefore, our ultimate goal is to minimize $\|\mathbf{w} - \mathbf{w}^*\|_2$ efficiently. To this aim, we study the difference of $\|\mathbf{w}^{(t+1)} - \mathbf{w}^*\|_2^2$ and $\|\mathbf{w}^{(t)} - \mathbf{w}^*\|_2^2$. We remind the reader that for convenience of notation, we denote the empirical gradients as the following

$$\mathbf{g}^{(t)} = \frac{1}{N}\sum_{j=1}^{N}(\sigma(\mathbf{w}^{(t)} \cdot \mathbf{x}(j)) - y(j))\mathbf{x}(j),$$
$$\mathbf{g}^{*} = \frac{1}{N}\sum_{j=1}^{N}(\sigma(\mathbf{w}^{*} \cdot \mathbf{x}(j)) - y(j))\mathbf{x}(j).$$

Moreover, we denote the "noise-free" empirical gradient by $\bar{\mathbf{g}}^{(t)}$, i.e.,

$$\bar{\mathbf{g}}^{(t)} = \mathbf{g}^{(t)} - \mathbf{g}^* = \frac{1}{N}\sum_{j=1}^{N}(\sigma(\mathbf{w}^{(t)} \cdot \mathbf{x}(j)) - \sigma(\mathbf{w}^{*} \cdot \mathbf{x}(j)))\mathbf{x}(j).$$

Plugging in the iteration scheme $\mathbf{w}^{(t+1)} = \mathbf{w}^{(t)} - \eta\mathbf{g}^{(t)}$ while expanding the squared norm, we get

$$\begin{aligned}\|\mathbf{w}^{(t+1)} - \mathbf{w}^*\|_2^2 &= \|\mathbf{w}^{(t)} - \mathbf{w}^*\|_2^2 - 2\eta\mathbf{g}^{(t)} \cdot (\mathbf{w}^{(t)} - \mathbf{w}^*) + \eta^2\|\mathbf{g}^{(t)}\|_2^2 \\ &\leq \underbrace{\|\mathbf{w}^{(t)} - \mathbf{w}^*\|_2^2 - 2\eta\nabla\mathcal{L}_{\mathrm{sur}}^{\mathcal{D},\sigma}(\mathbf{w}^{(t)}) \cdot (\mathbf{w}^{(t)} - \mathbf{w}^*)}_{Q_1} \\ &\quad \underbrace{-2\eta(\mathbf{g}^{(t)} - \nabla\mathcal{L}_{\mathrm{sur}}^{\mathcal{D},\sigma}(\mathbf{w}^{(t)})) \cdot (\mathbf{w}^{(t)} - \mathbf{w}^*) + \eta^2\|\mathbf{g}^{(t)}\|_2^2}_{Q_2}\ .\end{aligned}$$

Observe that we decomposed the right-hand side into two parts, the true contribution of the gradient (Q_1) and the estimation error (Q_2).

Note that in order to utilize the sharpness property of surrogate loss at the point $\mathbf{w}^{(t)}$, the conditions

$$\begin{aligned}\mathbf{w}^{(t)} &\in \mathcal{B}(2||\mathbf{w}^*||_2) \text{ and} \\ \mathbf{w}^{(t)} &\in \{\mathbf{w} : \|\mathbf{w}^{(t)} - \mathbf{w}^*\|_2^2 \geq 20B/(\bar{\mu}^2\rho)\mathrm{OPT}\}\end{aligned} \tag{34}$$

need to be satisfied. For the first condition, recall that we initialized $\mathbf{w}^{(0)} = \mathbf{0}$, hence Equation (34) is valid for $t = 0$. By induction rule, it suffices to show that assuming $\mathbf{w}^{(t)} \in \mathcal{B}(2||\mathbf{w}^*||_2)$ holds, we have $||\mathbf{w}^{(t+1)} - \mathbf{w}^*||_2 \leq (1-C)||\mathbf{w}^{(t)} - \mathbf{w}^*||_2$ for some constant $0 < C < 1$. Thus, we assume temporarily Equation (34) is true at iteration t, and we will show in the remainder of the proof that $||\mathbf{w}^{(t+1)} - \mathbf{w}^*||_2 \leq (1-C)||\mathbf{w}^{(t)} - \mathbf{w}^*||_2$ until we arrived at some final iteration T. Then by induction, the first part of Equation (34) is satisfied at each step $t \leq T$. For the second condition, note that if it is violated at some iteration T, then $\|\mathbf{w}^{(T)} - \mathbf{w}^*\|_2 \leq O(\mathrm{OPT})$ implying that this would be the solution we are looking for and the algorithm could be terminated at T. Therefore, whenever $\|\mathbf{w}^{(t)} - \mathbf{w}^*\|_2$ is far away from OPT, the prerequisites of Proposition D.3 are satisfied and the sharpness property of $\mathcal{L}_{\mathrm{sur}}^{\mathcal{D},\sigma}$ is allowed to use.

Now for the first term (Q_1), using the fact that $\mathcal{L}_{\mathrm{sur}}^{\mathcal{D},\sigma}(\mathbf{w}^{(t)})$ is $\mu(\gamma, \lambda, \beta, \rho, B)$-sharp according to Corollary D.4, we immediately get a sufficient decrease at each iteration: $||\mathbf{w}^{(t+1)} - \mathbf{w}^*||_2^2 \leq (1-C)||\mathbf{w}^{(t)} - \mathbf{w}^*||_2^2$. Namely, denote $\mu(\gamma, \lambda, \beta, \rho, B)$ as μ for simplicity, applying Corollary D.4 we have

$$(Q_1) = \|\mathbf{w}^{(t)} - \mathbf{w}^*\|_2^2 - 2\eta\nabla\mathcal{L}_{\mathrm{sur}}^{\mathcal{D},\sigma}(\mathbf{w}^{(t)}) \cdot (\mathbf{w}^{(t)} - \mathbf{w}^*) \leq (1 - 2\eta\mu)\|\mathbf{w}^{(t)} - \mathbf{w}^*\|_2^2\ ,$$

where $\mu = 1/2\bar{\mu}$, and $\bar{\mu} = C\lambda^2\gamma\beta\rho/B$ for some sufficiently small constant C.

Now it suffices to show that (Q_2) can be bounded above by $C'||\mathbf{w}^{(t)} - \mathbf{w}^*||_2^2$, where C' is a parameter depending on η and μ that can be made comparatively small. Formally, we show the following claim.

Claim D.14. Suppose $\eta \leq 1$. Fix $r_\epsilon \geq 1$ such that $H_4(r_\epsilon)$ is a sufficiently small multiple of ϵ. Choosing N to be a sufficiently large constant multiple of $\frac{d}{\delta}(r_\epsilon^2 + \alpha^2 M^2)$, then we have with probability at least $1-\delta$

$$(Q_2) \leq \left(\frac{3}{2}\eta\mu + \frac{8\eta^2\alpha^2 B^2}{\rho^2}\right)\|\mathbf{w}^{(t)} - \mathbf{w}^*\|_2^2 + \frac{4\eta}{\mu}\left(\frac{2B}{\rho}\mathrm{OPT} + \epsilon\right).$$

Proof of Claim D.14. Observe that by applying the Arithmetic-Geometric Mean inequality and Cauchy-Schwarz inequality, we get $\mathbf{x}\cdot\mathbf{y} \leq (a/2)\|\mathbf{x}\|_2^2 + (1/2a)\|\mathbf{y}\|_2^2$ for any vector $\mathbf{x}$ and $\mathbf{y}$, thus applying this inequality to the inner product $(\mathbf{g}^{(t)} - \nabla\mathcal{L}_{\mathrm{sur}}^{\mathcal{D},\sigma}(\mathbf{w}^{(t)}))\cdot(\mathbf{w}^{(t)} - \mathbf{w}^*)$ with coefficient $a = \mu$, we get

$$\begin{aligned}
(Q_2) &= -2\eta(\mathbf{g}^{(t)} - \nabla\mathcal{L}_{\mathrm{sur}}^{\mathcal{D},\sigma}(\mathbf{w}^{(t)}))\cdot(\mathbf{w}^{(t)} - \mathbf{w}^*) + 2\eta^2\|\mathbf{g}^{(t)}\|_2^2 \\
&\leq -2\eta(\mathbf{g}^{(t)} - \nabla\mathcal{L}_{\mathrm{sur}}^{\mathcal{D},\sigma}(\mathbf{w}^{(t)}))\cdot(\mathbf{w}^{(t)} - \mathbf{w}^*) + 2\eta^2\|\bar{\mathbf{g}}^{(t)}\|_2^2 + 2\eta^2\|\mathbf{g}^*\|_2^2 \\
&\leq \frac{\eta}{\mu}\|\mathbf{g}^{(t)} - \nabla\mathcal{L}_{\mathrm{sur}}^{\mathcal{D},\sigma}(\mathbf{w}^{(t)})\|_2^2 + \eta\mu\|\mathbf{w}^{(t)} - \mathbf{w}^*\|_2^2 + 2\eta^2\|\bar{\mathbf{g}}^{(t)}\|_2^2 + 2\eta^2\|\mathbf{g}^*\|_2^2\,,
\end{aligned}$$

where μ is the sharpness parameter and we used the definition that $\bar{\mathbf{g}}^{(t)} = \mathbf{g}^{(t)} - \mathbf{g}^*$ in the first inequality. Note that

$$\begin{aligned}
\|\mathbf{g}^{(t)} - \nabla\mathcal{L}_{\mathrm{sur}}^{\mathcal{D},\sigma}(\mathbf{w}^{(t)})\|_2^2 &= \|\mathbf{g}^{(t)} - \mathbf{g}^* - (\nabla\mathcal{L}_{\mathrm{sur}}^{\mathcal{D},\sigma}(\mathbf{w}^{(t)}) - \nabla\mathcal{L}_{\mathrm{sur}}^{\mathcal{D},\sigma}(\mathbf{w}^*)) + \mathbf{g}^* - \nabla\mathcal{L}_{\mathrm{sur}}^{\mathcal{D},\sigma}(\mathbf{w}^*)\|_2^2 \\
&\leq 2\|\bar{\mathbf{g}}^{(t)} - \nabla\bar{\mathcal{L}}_{\mathrm{sur}}^{\mathcal{D},\sigma}(\mathbf{w}^{(t)})\|_2^2 + 2\|\mathbf{g}^* - \nabla\mathcal{L}_{\mathrm{sur}}^{\mathcal{D},\sigma}(\mathbf{w}^*)\|_2^2,
\end{aligned}$$

since we have $\bar{\mathcal{L}}_{\mathrm{sur}}^{\mathcal{D},\sigma}(\mathbf{w}^{(t)}) = \mathcal{L}_{\mathrm{sur}}^{\mathcal{D},\sigma}(\mathbf{w}^{(t)}) - \mathcal{L}_{\mathrm{sur}}^{\mathcal{D},\sigma}(\mathbf{w}^*)$. Thus, it holds

$$\begin{aligned}
(Q_2) \leq &\frac{2\eta}{\mu}\|\bar{\mathbf{g}}^{(t)} - \nabla\bar{\mathcal{L}}_{\mathrm{sur}}^{\mathcal{D},\sigma}(\mathbf{w}^{(t)})\|_2^2 + \frac{2\eta}{\mu}\|\mathbf{g}^* - \nabla\mathcal{L}_{\mathrm{sur}}^{\mathcal{D},\sigma}(\mathbf{w}^*)\|_2^2 + \eta\mu\|\mathbf{w}^{(t)} - \mathbf{w}^*\|_2^2 \\
&+ 2\eta^2\|\bar{\mathbf{g}}^{(t)}\|_2^2 + 2\eta^2\|\mathbf{g}^*\|_2^2
\end{aligned} \tag{35}$$

Furthermore, recall that as shown in Lemma D.9 and Corollary D.13, $\|\bar{\mathbf{g}}^{(t)} - \nabla\bar{\mathcal{L}}_{\mathrm{sur}}^{\mathcal{D},\sigma}(\mathbf{w}^{(t)})\|_2^2$, $\|\nabla\bar{\mathcal{L}}_{\mathrm{sur}}^{\mathcal{D},\sigma}(\mathbf{w}^{(t)})\|_2^2$, $\|\mathbf{g}^*\|_2^2$ and $\|\bar{\mathbf{g}}^{(t)}\|_2^2$ can be made small by increasing the batch size N. In particular, when r_ϵ satisfies $H_4(r_\epsilon) \lesssim \epsilon$, we have proved that with probability at least $1-\delta$, it holds

$$\|\bar{\mathbf{g}}^{(t)} - \nabla\bar{\mathcal{L}}_{\mathrm{sur}}^{\mathcal{D},\sigma}(\mathbf{w}^{(t)})\|_2^2 \lesssim \frac{\alpha^2 dB}{\delta\rho N}\|\mathbf{w}^{(t)} - \mathbf{w}^*\|_2^2, \|\bar{\mathbf{g}}^{(t)}\|_2^2 \lesssim \frac{\alpha^2 B^2}{\rho^2}\left(1 + \sqrt{\frac{d\rho}{\delta BN}}\right)^2\|\mathbf{w}^{(t)} - \mathbf{w}^*\|_2^2,$$

$$\|\mathbf{g}^* - \nabla\mathcal{L}_{\mathrm{sur}}^{\mathcal{D},\sigma}(\mathbf{w}^*)\|_2^2 \lesssim \frac{d}{\delta N}(r_\epsilon^2\mathrm{OPT} + \alpha^2 M^2\epsilon),$$

and

$$\|\mathbf{g}^*\|_2^2 \lesssim \left(\sqrt{(B/\rho)\mathrm{OPT}} + \sqrt{\frac{d(r_\epsilon^2\mathrm{OPT} + \alpha^2 M^2\epsilon)}{\delta N}}\right)^2 \lesssim \frac{B}{\rho}\left(1 + \frac{r_\epsilon^2 d}{\delta N}\right)\mathrm{OPT} + \frac{\alpha^2 M^2 d}{\delta N}\epsilon\,.$$

Therefore, choosing

$$N \geq C\max\left\{\frac{dr_\epsilon^2}{\delta}, \frac{\alpha^2 M^2 d}{\delta}, \frac{B\alpha^2 d}{\rho\mu^2\delta}\right\}, \tag{36}$$

where C is a sufficiently large absolute constant, then with probability at least $1-\delta$, it holds

$$\|\bar{\mathbf{g}}^{(t)} - \nabla\bar{\mathcal{L}}_{\mathrm{sur}}^{\mathcal{D},\sigma}(\mathbf{w}^{(t)})\|_2^2 \leq \frac{\mu^2}{4}\|\mathbf{w}^{(t)} - \mathbf{w}^*\|_2^2,\ \|\bar{\mathbf{g}}^{(t)}\|_2^2 \leq \frac{4\alpha^2 B^2}{\rho^2}\|\mathbf{w}^{(t)} - \mathbf{w}^*\|_2^2,$$

and

$$\|\mathbf{g}^* - \nabla\mathcal{L}_{\mathrm{sur}}^{\mathcal{D},\sigma}(\mathbf{w}^*)\|_2^2 \leq \mathrm{OPT} + \epsilon\,, \quad \|\mathbf{g}^*\|_2^2 \leq \frac{2B}{\rho}\mathrm{OPT} + \epsilon.$$

Plugging these bounds back to Equation (35), we get

$$\begin{aligned}
(Q_2) &\leq \left(\frac{3}{2}\eta\mu + 8\frac{\eta^2\alpha^2 B^2}{\rho^2}\right)\|\mathbf{w}^{(t)} - \mathbf{w}^*\|_2^2 + 2\eta\left(\eta + \frac{1}{\mu}\right)\left(\frac{2B}{\rho}\mathrm{OPT} + \epsilon\right) \\
&\leq \left(\frac{3}{2}\eta\mu + 8\frac{\eta^2\alpha^2 B^2}{\rho^2}\right)\|\mathbf{w}^{(t)} - \mathbf{w}^*\|_2^2 + \frac{4\eta}{\mu}\left(\frac{2B}{\rho}\mathrm{OPT} + \epsilon\right),
\end{aligned}$$

where in the last inequality we used the assumption that $\eta \leq 1$ and $\mu \leq 1$. The proof is now complete. □

Now combining the upper bounds on (Q_1) and (Q_2) and choosing $\eta = \frac{\mu\rho^2}{32\alpha^2 B^2}$, we have:

$$\begin{aligned}\|\mathbf{w}^{(t+1)} - \mathbf{w}^*\|_2^2 &\leq \left(1 - \frac{1}{2}\eta\mu + \frac{8\eta^2\alpha^2 B^2}{\rho^2}\right)\|\mathbf{w}^{(t)} - \mathbf{w}^*\|_2^2 + \frac{4\eta}{\mu}\left(\frac{2B}{\rho}\mathrm{OPT} + \epsilon\right) \\ &\leq \left(1 - \frac{\mu^2\rho^2}{128\alpha^2 B^2}\right)\|\mathbf{w}^{(t)} - \mathbf{w}^*\|_2^2 + \frac{\rho}{4\alpha^2 B}\left(\mathrm{OPT} + \epsilon\right). \end{aligned} \tag{37}$$

When $\|\mathbf{w}^{(t)} - \mathbf{w}^*\|_2^2 \geq (64B/(\rho\mu^2))(\mathrm{OPT} + \epsilon)$, in other words when $\mathbf{w}^{(t)}$ is still away from the minimizer $\mathbf{w}^*$, it further holds with probability $1 - \delta$:

$$\|\mathbf{w}^{(t+1)} - \mathbf{w}^*\|_2^2 \leq \left(1 - \frac{\mu^2\rho^2}{256\alpha^2 B^2}\right)\|\mathbf{w}^{(t)} - \mathbf{w}^*\|_2^2, \tag{38}$$

which proves the sufficient decrease of $||\mathbf{w}^{(t)} - \mathbf{w}^*||_2^2$ that we proposed at the beginning.

Let T be the first iteration such that $\mathbf{w}^{(T)}$ satisfies $\|\mathbf{w}^{(T)} - \mathbf{w}^*\|_2^2 \leq (64B/(\rho\mu^2))(\mathrm{OPT} + \epsilon)$. Recall that we need Equation (34) for every $t \leq T$ to be satisfied to implement sharpness. The first condition is satisfied naturally for $\|\mathbf{w}^{(t+1)} - \mathbf{w}^*\|_2^2 \leq \|\mathbf{w}^*\|_2^2$ as a consequence of Equation (38) (recall that $\mathbf{w}^{(0)} = 0$). For the second condition, when $t + 1 \leq T$, since $\mu = 1/2\bar{\mu}$, we have

$$\|\mathbf{w}^{(t+1)} - \mathbf{w}^*\|_2^2 \geq \frac{64B}{\rho\mu^2}(\mathrm{OPT} + \epsilon) \geq \frac{20B}{\rho\bar{\mu}^2}\mathrm{OPT},$$

hence the second condition is also satisfied.

When $t \leq T$, the contraction of $\|\mathbf{w}^{(t)} - \mathbf{w}^*\|_2^2$ indicates a linear convergence rate of stochastic gradient descent. Since $\mathbf{w}^{(0)} = 0$, $\|\mathbf{w}^*\|_2 \leq W$, it holds $\|\mathbf{w}^{(t)} - \mathbf{w}^*\|_2^2 \leq (1 - \mu^2\rho^2/(256\alpha^2 B^2))^t\|\mathbf{w}^{(0)} - \mathbf{w}^*\|_2^2 \leq \exp(-t\mu^2\rho^2/(256\alpha^2 B^2))W^2$. Thus, to generate a point $\mathbf{w}^{(T)}$ such that $\|\mathbf{w}^{(T)} - \mathbf{w}^*\|_2^2 \leq (64B/(\rho\mu^2))(\mathrm{OPT} + \epsilon)$, it suffices to run Algorithm 2 for

$$T = \widetilde{\Theta}\left(\frac{B^2\alpha^2}{\rho^2\mu^2}\log\left(\frac{W}{\epsilon}\right)\right) \tag{39}$$

iterations, where the logarithmic dependence on parameters α, B, ρ and μ are hidden in the $\widetilde{\Theta}(\cdot)$ notation. Further, recall that at each step t the contraction $\|\mathbf{w}^{(t+1)} - \mathbf{w}^*\|_2^2 \leq (1 - \frac{\mu^2\rho^2}{256\alpha^2 B^2})\|\mathbf{w}^{(t)} - \mathbf{w}^*\|_2^2$ holds with probability $1 - \delta$, thus the union bound inequality implies $\|\mathbf{w}^{(T)} - \mathbf{w}^*\|_2^2 \leq (64B/(\rho\mu^2))(\mathrm{OPT} + \epsilon)$ holds with probability $1 - T\delta$. Let $\delta = 1/(3T)$, we get with probability at least $2/3$, $\|\mathbf{w}^{(T)} - \mathbf{w}^*\|_2^2 \leq (64B/(\rho\mu^2))(\mathrm{OPT} + \epsilon)$, and thus from Equation (33),

$$\mathcal{L}_2^{\mathcal{D},\sigma}(\mathbf{w}^{(T)}) \leq 2\mathrm{OPT} + \frac{640\alpha^2 B^2}{\rho^2\mu^2}(\mathrm{OPT} + \epsilon) = O\left(\left(\frac{B\alpha}{\rho\mu}\right)^2\mathrm{OPT}\right) + \epsilon,$$

and the proof is now complete. □

In the final part of this section we apply Theorem D.7 to sub-Exponential and k-Heavy tail distributions. Before we dig into the details, some upper bounds on $H_2(r)$ and $H_4(r)$ are needed. We provide the following simple fact.

Fact D.15. Let $H_2(r)$ and $H_4(r)$ be as in Definition D.5. Then, we have the following bounds:

$$\begin{aligned} H_2(r) &\leq r^2\min\{1, h(r)\} + \int_r^\infty 2s\min\{1, h(s)\}\,\mathrm{d}s, \\ H_4(r) &\leq r^4\min\{1, h(r)\} + \int_r^\infty 4s^3\min\{1, h(s)\}\,\mathrm{d}s. \end{aligned}$$

Moreover, if $\mathcal{D}_{\mathbf{x}}$ is sub-exponential with $h(r) = \exp(-r/B)$ or k-heavy tailed with $h(r) = B/r^k$, $k > 4 + \rho$, $\rho > 0$, and $r \geq \max\{1, B^{-4-\rho}\}$ then

$$H_2(r) \lesssim r^2 h(r) \quad \text{and} \quad H_4(r) \lesssim r^4 h(r).$$

Proof. To prove the fact, we bound the expectation $\mathbf{E}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}\left[|\mathbf{u}\cdot\mathbf{x}|^i\mathbb{1}\{|\mathbf{u}\cdot\mathbf{x}|\geq r\}\right]$ for any vector $\mathbf{u}\in\mathcal{B}(1)$, where $i=2,4$. To calculate the expectation, observe that when $t<r^i$, it holds

$$\mathbf{Pr}\left[|\mathbf{u}\cdot\mathbf{x}|^i\mathbb{1}\{|\mathbf{u}\cdot\mathbf{x}|\geq r\}\geq t\right]=\mathbf{Pr}\left[|\mathbf{u}\cdot\mathbf{x}|\geq r\right].$$

Thus, we have

$$\begin{aligned}\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}\left[|\mathbf{u}\cdot\mathbf{x}|^i\mathbb{1}\{|\mathbf{u}\cdot\mathbf{x}|\geq r\}\right]&=\int_0^\infty\mathbf{Pr}\left[|\mathbf{u}\cdot\mathbf{x}|^i\mathbb{1}\{|\mathbf{u}\cdot\mathbf{x}|\geq r\}\geq t\right]\mathrm{d}t\\&=\mathbf{Pr}\left[|\mathbf{u}\cdot\mathbf{x}|\geq r\right]\int_0^{r^i}1\,\mathrm{d}t+\int_{r^i}^\infty\mathbf{Pr}\left[|\mathbf{u}\cdot\mathbf{x}|^i\mathbb{1}\{|\mathbf{u}\cdot\mathbf{x}|\geq r\}\geq t\right]\mathrm{d}t\\&=r^i\mathbf{Pr}\left[|\mathbf{u}\cdot\mathbf{x}|\geq r\right]+\int_r^\infty i\mathbf{Pr}\left[|\mathbf{u}\cdot\mathbf{x}|^i\mathbb{1}\{|\mathbf{u}\cdot\mathbf{x}|\geq r\}\geq s^i\right]s^{i-1}\,\mathrm{d}s.\end{aligned}$$

Since (by Assumption 2.4) $\mathbf{Pr}\left[|\mathbf{u}\cdot\mathbf{x}|\geq r\right]\leq\min\{1,h(r)\}$, and further note that when $s\geq r$ it holds

$$\mathbf{Pr}\left[|\mathbf{u}\cdot\mathbf{x}|^i\mathbb{1}\{|\mathbf{u}\cdot\mathbf{x}|\geq r\}\geq s^i\right]=\mathbf{Pr}\left[|\mathbf{u}\cdot\mathbf{x}|\mathbb{1}\{|\mathbf{u}\cdot\mathbf{x}|\geq r\}\geq s\right]=\mathbf{Pr}\left[|\mathbf{u}\cdot\mathbf{x}|\geq s\right]\leq\min\{1,h(s)\},$$

then we get

$$\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}\left[|\mathbf{u}\cdot\mathbf{x}|^i\mathbb{1}\{|\mathbf{u}\cdot\mathbf{x}|\geq r\}\right]\leq r^i\min\{1,h(r)\}+\int_r^\infty is^{i-1}\min\{1,h(s)\}\,\mathrm{d}s,$$

which holds for any $\mathbf{u}\in\mathcal{B}(1)$. Therefore, we proved the first part of the claim by taking the maximum over $\mathbf{u}\in\mathcal{B}(1)$ on both sides of the inequality.

Now, consider $r\geq\max\{1,B^{-4-\rho}\}$. Then for sub-Exponential distributions, as $h(s)=\exp(-\frac{s}{B})\leq 1$ when $s\geq r$, we have:

$$\int_r^\infty sh(s)\,\mathrm{d}s=\int_r^\infty s\exp(-s/B)\,\mathrm{d}s=B(r-B)\exp(-r/B)\leq Br^2h(r),$$

and

$$\int_r^\infty s^3h(s)\,\mathrm{d}s=\int_r^\infty s^3\exp(-s/B)\,\mathrm{d}s=B^4((r/B)^3+3(r/B)^2+6r/B+6)\exp(-r/B)\leq 16B^4r^4h(r),$$

where we assumed without loss of generality that $c\leq 1$. Hence $H_2(r)\leq(1+2B)r^2h(r)$ and $H_4(r)\leq(1+64B^4)r^4h(r)$, proving the desired claim.

Finally, for k-Heavy tail distributions with $k>4+\rho$, $\rho>0$, $h(r)=B/r^k$. Since $h(r)\leq 1$ when $r\geq\max\{1,B^{-4-\rho}\}$, we have:

$$H_2(r)\leq r^2h(r)+\int_r^\infty\frac{2B}{s^{k-1}}\,\mathrm{d}s\leq(1+2B)r^2h(r),$$

and in addition,

$$H_4(r)\leq r^4h(r)+\int_r^\infty\frac{4B}{s^{k-3}}\,\mathrm{d}s\leq\left(1+\frac{4B}{\rho}\right)r^4h(r).$$

The claim is now complete. □

Applying Theorem D.7 to sub-Exponential distributions yeilds an L_2^2 error of order $O(\mathrm{OPT})+\epsilon$ with $\tilde{\Theta}(\log(1/\epsilon))$ convergence rate, using $\tilde{\Omega}(\mathrm{polylog}(1/\epsilon))$ samples. Formally, we have the following corollaries.

Corollary D.16 (Sub-Exponential Distributions)**.** *Fix $\epsilon>0$ and $W>0$ and suppose Assumptions 2.2 and 2.3 hold. Moreover, assume that Assumption 2.4 holds for $h(r)=\exp(-r/B)$ for some $B\geq 1$. Let OPT denote the minimum value of the L_2^2, i.e., $\mathrm{OPT}=\min_{\mathbf{w}\in\mathcal{B}(W)}\mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}\left[(\sigma(\mathbf{w}\cdot\mathbf{x})-y)^2\right]$. Let $\mu:=\mu(\lambda,\gamma,\beta,B)$ be a sufficiently small multiple of $\lambda^2\gamma\beta$, and let $M=O(\alpha WB\log\left(\frac{\alpha W}{\epsilon}\right))$. Then after*

$$T=\widetilde{\Theta}\left(\frac{B^2\alpha^2}{\mu^2}\log\left(\frac{W}{\epsilon}\right)\right)$$

iterations with batch size

$$N = \widetilde{\Omega}\left(\frac{dB^4\alpha^6W^2}{\mu^2}\mathrm{polylog}(1/\epsilon)\right),$$

Algorithm 2 converges to a point $\mathbf{w}^{(T)}$ *such that* $\mathcal{L}_2^{\mathcal{D},\sigma}(\mathbf{w}^{(T)}) = O\big(\big(\frac{B\alpha}{\mu}\big)^2\mathrm{OPT}\big) + \epsilon$ *with probability at least* $2/3$.

Proof of Corollary D.16. Since by assumption it holds $h(r) = \exp(-r/B) \le B/r^{4+\rho}$ for $\rho = 1$, we can set $\rho = 1$ in Theorem D.7. Thus, a direct application of Theorem D.7 with parameter $\rho = 1$ gives the desired L_2^2 error and the required number of iterations. It remains to determine the batch size with respect to the sub-Exponential distributions. To this aim, note that $N = \Omega(dT(r_\epsilon^2 + \alpha^2M^2))$, thus we need to find the truncation bound M, which is defined in Lemma D.8, and calculate r_ϵ such that $H_4(r_\epsilon) \lesssim \epsilon$.

Denote $\kappa = \frac{\epsilon}{4\alpha^2W^2}$. Recall that $M = \alpha W H_2^{-1}(\kappa)$. To determine $H_2^{-1}(\kappa)$, note that $H_2(r)$ is a non-increasing function, therefore it suffices to find a r_κ such that $H_2(r_\kappa) \le \kappa$, then it holds $H_2^{-1}(\kappa) \le r_\kappa$. For sub-Exponential distributions where $h(r) = \exp(-r/B)$, choosing $r_\kappa = B\log(1/\kappa^2)$ satisfies

$$r_\kappa^2 h(r_\kappa) = 4B^2\log^2\left(\frac{1}{\kappa}\right)\kappa^2 \le 4B^2\kappa,$$

since $\log^2(1/\kappa)\kappa \le 1$ as $\kappa = \epsilon/(4\alpha^2W^2) \le 1$. Further note that in Fact D.15 we showed $H_2(r) \le (1+2B)r^2h(r)$, thus $H_2(r_\kappa) \lesssim \kappa$ hence $M = O(\alpha WB\log((\alpha W)/\epsilon))$.

For r_ϵ, by the same idea one can show that for $r_\epsilon = B\log(1/\epsilon^2)$, it holds

$$H_4(r_\epsilon) \le \left(1+64B^4\right)r_\epsilon^4\exp(-r_\epsilon/B) = \left(1+64B^4\right)16B^4\log^4(1/\epsilon)\epsilon^2 \le \left(1+64B^4\right)80B^4\epsilon,$$

where the first inequality is due to Fact D.15 and in the last inequality we used the fact that $\log^4(1/\epsilon)\epsilon \le 5$ when $\epsilon \le 1$.

Therefore, combining the bounds on M and r_ϵ, we get

$$N = \widetilde{\Omega}(dT(r_\epsilon^2 + \alpha^2M^2)) = \widetilde{\Omega}\left(\frac{dB^4\alpha^6W^2}{\rho^2\mu^2}\log^3\left(\frac{1}{\epsilon}\right)\right).$$

□

Next, we apply Theorem D.7 to Heavy-Tail distributions.

Corollary D.17 (Heavy-Tail Distributions)**.** *Fix* $\epsilon > 0$ *and* $W > 0$ *and suppose Assumptions 2.2 and 2.3 hold. Moreover, assume that Assumption 2.4 holds for* $h(r) = B/r^k$ *for some* $k > 4+\rho$ *where* $\rho > 0$ *and* $B \ge 1$. *Let* OPT *denote the minimum value of the* L_2^2*, i.e.,* $\mathrm{OPT} = \min_{\mathbf{w}\in\mathcal{B}(W)}\mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}\left[(\sigma(\mathbf{w}\cdot\mathbf{x}) - y)^2\right]$. *Let* $\mu := \mu(\lambda,\gamma,\beta,\rho,B)$ *be a sufficiently small multiple of* $\lambda^2\gamma\beta\rho/B$*, and let* $M = \Theta(\alpha W\big(\frac{\alpha WB}{\epsilon}\big)^{1/(k-2)})$. *Then after*

$$T = \widetilde{\Theta}\left(\frac{B^2\alpha^2}{\rho^2\mu^2}\log\left(\frac{W}{\epsilon}\right)\right)$$

iterations with batch size

$$N = \widetilde{\Omega}\left(\frac{dB^2\alpha^6W^2}{\rho^2\mu^2}\left(\frac{B}{\epsilon}\right)^{\frac{2}{k-4}}\right),$$

Algorithm 2 converges to a point $\mathbf{w}^{(T)}$ *such that* $\mathcal{L}_2^{\mathcal{D},\sigma}(\mathbf{w}^{(T)}) = O\big(\big(\frac{B\alpha}{\rho\mu}\big)^2\mathrm{OPT}\big) + \epsilon$ *with probability at least* $2/3$.

Proof of Corollary D.17. Applying Theorem D.7 directly we get the desired convergence rate and L_2^2 loss. Now for batch size, we need to determine the truncation bound $M = \alpha W H_2^{-1}(\epsilon/(4\alpha^2W^2))$ (see Lemma D.8) as well as r_ϵ such that $H_4(r_\epsilon) \lesssim \epsilon$.

First, denote $\kappa = \frac{\epsilon}{4\alpha^2W^2}$. Let $r_\kappa = (\frac{2B}{\kappa})^{1/(k-2)}$. By Fact D.15, $H_2(r_\kappa) \lesssim r_\kappa^2h(r_\kappa) = \frac{\kappa}{2}$. Since $H_2(r)$ is non-increasing, we know $H_2^{-1}(\kappa) \lesssim r_\kappa$. Thus, $M = O(\alpha W(\frac{\alpha WB}{\epsilon})^{1/(k-2)})$. Next, choose $r_\epsilon = (\frac{B}{\epsilon})^{1/(k-4)}$, then it holds $H_4(r_\epsilon) \lesssim r_\epsilon^4h(r_\epsilon) = \epsilon$ satisfying the condition.

Combining the bounds on r_ϵ and M, we get the batch size

$$N = \Omega(dT(r_\epsilon^2 + \alpha^2 M^2)) = \widetilde{\Omega}\left(\frac{dB^2\alpha^2}{\rho^2\mu^2}\log\left(\frac{W}{\epsilon}\right)\left(\left(\frac{B}{\epsilon}\right)^{\frac{2}{k-4}} + \alpha^4 W^2\left(\frac{B}{\epsilon}\right)^{\frac{2}{k-2}}\right)\right)$$
$$= \widetilde{\Omega}\left(\frac{dB^2\alpha^6 W^2}{\rho^2\mu^2}\left(\frac{B}{\epsilon}\right)^{\frac{2}{k-4}}\right).$$

□

Thus, Algorithm 2 yields an L_2^2 error of $O(\mathrm{OPT}) + \epsilon$ in $\widetilde{\Theta}(\log(1/\epsilon))$ iterations with batch size $\widetilde{\Omega}((1/\epsilon)^{2/(k-4)})$ when applied on k-Heavy Tailed distributions.

E. Distributions Satisfying Our Assumptions

In this section, we show that many natural distributions satisfy Assumptions 2.3 and 2.4

E.1. Well-Behaved Distributions from Diakonikolas et al. (2022b)

We first consider the class of distributions defined by Diakonikolas et al. (2022b) and termed "well-behaved". This distribution class contains many natural distributions like log-concave and s-concave distributions.

Definition E.1 (Well-Behaved Distributions)**.** Let $L, R > 0$. An isotropic (i.e., zero mean and identity covariance) distribution $\mathcal{D}_{\mathbf{x}}$ on $\mathbb{R}^d$ is called (L, R)-well-behaved if for any projection $(\mathcal{D}_{\mathbf{x}})_V$ of $\mathcal{D}_{\mathbf{x}}$ onto a subspace V of dimension at most two, the corresponding pdf ϕ_V on $\mathbb{R}^2$ satisfies the following:

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

The distribution class that is (L, B)-well-behaved satisfies Assumption 2.4. Therefore, we need to show that the distributions in this class satisfy Assumption 2.3.

Lemma E.2. *Let $\mathcal{D}_{\mathbf{x}}$ be a (L, B)-well-behaved distribution. Then, $\mathcal{D}_{\mathbf{x}}$ satisfies Assumption 2.3, for $\gamma = R/2$ and $\lambda = LR^4/16$.*

Proof. Let $\mathbf{u}, \mathbf{v} \in \mathbb{R}^d$ be any two orthonormal vectors and let V be the subspace spanned by $\mathbf{u}, \mathbf{v}$. We have that

$$\mathbf{E}_{\mathbf{x}\sim\mathcal{D}_{\mathbf{x}}}\left[(\mathbf{u}\cdot\mathbf{x})^2\mathbb{1}\{\mathbf{v}\cdot\mathbf{x} \geq R/2\}\right] \geq \mathbf{E}_{\mathbf{x}\sim\mathcal{D}_{\mathbf{x}}}\left[(\mathbf{u}\cdot\mathbf{x})^2\mathbb{1}\{R \geq \mathbf{u}\cdot\mathbf{x} \geq R/2, R \geq \mathbf{v}\cdot\mathbf{x} \geq R/2\}\right]$$
$$\geq (R^2/4)\mathbf{E}_{\mathbf{x}\sim\mathcal{D}_{\mathbf{x}}}\left[\mathbb{1}\{R \geq \mathbf{u}\cdot\mathbf{x} \geq R/2, R \geq \mathbf{v}\cdot\mathbf{x} \geq R/2\}\right]$$
$$= (R^2/4)\int_{R/2}^{R}\int_{R/2}^{R}\phi_V(\mathbf{x})\mathrm{d}\mathbf{x} \geq (R^2/4)L\int_{R/2}^{R}\int_{R/2}^{R}\mathrm{d}\mathbf{x} = LR^4/16\,.$$

Furthermore, similarly, we have that $\mathbf{E}_{\mathbf{x}\sim\mathcal{D}_{\mathbf{x}}}\left[(\mathbf{v}\cdot\mathbf{x})^2\mathbb{1}\{\mathbf{v}\cdot\mathbf{x} \geq R/2\}\right] \geq LR^4/16$. Therefore, $\mathcal{D}_{\mathbf{x}}$ satisfies Assumption 2.3 with $\gamma = R/2$ and $\lambda = LR^4/16$. □

E.2. Symmetric Product Distributions with Strong Concentration

E.2.1. k-Heavy Tailed Symmetric Distributions, $k \geq 7$

Here we show that symmetric product distributions with sufficiently large polynomial tails satisfy our assumptions.

Proposition E.3. *Let $\mathcal{D}_{\mathbf{x}}$ be a k-Heavy Tailed symmetric distribution with $k \geq 7$ and i.i.d. coordinates, i.e., it satisfies $\mathbf{Pr}\left[|\mathbf{u}\cdot\mathbf{x}| \geq r\right] \leq B/r^k$ for some absolute constant $B \geq 1$. Let $\alpha = \mathbf{E}_{\mathbf{x}\sim\mathcal{D}_{\mathbf{x}}}\left[\mathbf{x}_i^2\right]$ and $\beta = \mathbf{E}_{\mathbf{x}\sim\mathcal{D}_{\mathbf{x}}}\left[\mathbf{x}_i^4\right]$. Suppose $\beta - \alpha^2 \geq c\alpha^2$, where $c > 0$ is an absolute constant and let C to be suffciently small absolute multiple of $(k-6)c^2\alpha^4/B$. Then, $\mathcal{D}_{\mathbf{x}}$ satisfies Assumptions 2.3 and 2.4 with $\gamma = \frac{C}{2}(\frac{C^3}{16B})^{1/k}$ and $\lambda = \frac{C^5}{64}(\frac{C^3}{16B})^{2/k}$.*

Proof of Proposition E.3. First, observe that by definition, $\mathcal{D}_{\mathbf{x}}$ satisfies Assumption 2.4 with $h(r) = B/r^k$. We show that $\mathcal{D}_{\mathbf{x}}$ satisfies Assumption 2.3 with some absolute constants γ and λ. Let $\mathbf{u}, \mathbf{v} \in \mathbb{R}^d$ be any two orthonormal vectors. We have that

$$\mathbf{E}_{\mathbf{x}\sim\mathcal{D}_{\mathbf{x}}}\left[|\mathbf{u}\cdot\mathbf{x}|\mathbf{v}\cdot\mathbf{x}\right] = 0 \ ,$$

since the distribution is symmetric. Therefore, we have that $\mathbf{E}_{\mathbf{x}\sim\mathcal{D}_{\mathbf{x}}}\left[|\mathbf{u}\cdot\mathbf{x}||\mathbf{v}\cdot\mathbf{x}|\mathbb{1}\{\mathbf{v}\cdot\mathbf{x}\geq 0\}\right] = \mathbf{E}_{\mathbf{x}\sim\mathcal{D}_{\mathbf{x}}}\left[|\mathbf{u}\cdot\mathbf{x}||\mathbf{v}\cdot\mathbf{x}|\right]/2$. Let $V = |\mathbf{u}\cdot\mathbf{x}||\mathbf{v}\cdot\mathbf{x}|$. We show that $\mathbf{E}_{\mathbf{x}\sim\mathcal{D}_{\mathbf{x}}}\left[V\right] \gtrsim c^2\alpha^4(k-6)/B$.

Claim E.4. Let $V = |\mathbf{u}\cdot\mathbf{x}||\mathbf{v}\cdot\mathbf{x}|$. Assume that $\mathbf{x}$ has i.i.d. zero mean coordinates and that $\mathbf{x}$ is k-Heavy tailed with parameter $k \geq 7$, $B \geq 1$. Let $\alpha = \mathbf{E}_{\mathbf{x}\sim\mathcal{D}_{\mathbf{x}}}\left[\mathbf{x}_i^2\right]$ and $\beta = \mathbf{E}_{\mathbf{x}\sim\mathcal{D}_{\mathbf{x}}}\left[\mathbf{x}_i^4\right]$. If $\beta - \alpha^2 \geq c\alpha^2$, where $c > 0$ is an absolute constant then, $\mathbf{E}_{\mathbf{x}\sim\mathcal{D}_{\mathbf{x}}}\left[V\right] \gtrsim c^2\alpha^4(k-6)/B$.

Proof of Claim E.4. First, note that we can write V^2 as $\sqrt{V}V^{3/2}$, because $V \geq 0$. Therefore, by applying the Cauchy-Schwarz inequality, we have that

$$\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_{\mathbf{x}}}\left[V^2\right]^2 \leq \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_{\mathbf{x}}}\left[V\right]\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_{\mathbf{x}}}\left[V^3\right] \ .$$

Therefore, we have that $\mathbf{E}_{\mathbf{x}\sim\mathcal{D}_{\mathbf{x}}}\left[V\right] \geq (\mathbf{E}_{\mathbf{x}\sim\mathcal{D}_{\mathbf{x}}}\left[V^2\right])^2/\mathbf{E}_{\mathbf{x}\sim\mathcal{D}_{\mathbf{x}}}\left[V^3\right]$. We first bound $\mathbf{E}_{\mathbf{x}\sim\mathcal{D}_{\mathbf{x}}}\left[V^2\right]$ from below. Let $\alpha = \mathbf{E}_{\mathbf{x}\sim\mathcal{D}_{\mathbf{x}}}\left[\mathbf{x}_i^2\right]$ and $\beta = \mathbf{E}_{\mathbf{x}\sim\mathcal{D}_{\mathbf{x}}}\left[\mathbf{x}_i^4\right]$. Observe that

$$\begin{aligned}
\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_{\mathbf{x}}}\left[V^2\right] &= \sum_{i_1,i_2,i_3,i_4}\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_{\mathbf{x}}}\left[\mathbf{u}_{i_1}\mathbf{u}_{i_2}\mathbf{v}_{i_3}\mathbf{v}_{i_4}\mathbf{x}_{i_1}\mathbf{x}_{i_2}\mathbf{x}_{i_3}\mathbf{x}_{i_4}\right] \\
&= \sum_{i_1,i_2,i_1\neq i_2}\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_{\mathbf{x}}}\left[\mathbf{u}_{i_1}^2\mathbf{v}_{i_2}^2\mathbf{x}_{i_1}^2\mathbf{x}_{i_2}^2 + 2\mathbf{u}_{i_1}\mathbf{v}_{i_1}\mathbf{u}_{i_2}\mathbf{v}_{i_2}\mathbf{x}_{i_1}^2\mathbf{x}_{i_2}^2\right] + \sum_i \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_{\mathbf{x}}}\left[\mathbf{u}_i^2\mathbf{v}_i^2\mathbf{x}_i^4\right] \\
&= \alpha^2\sum_{i_1,i_2,i_1\neq i_2}\mathbf{u}_{i_1}^2\mathbf{v}_{i_2}^2 + 2\alpha^2\sum_{i_1,i_2,i_1\neq i_2}\mathbf{u}_{i_1}\mathbf{v}_{i_1}\mathbf{u}_{i_2}\mathbf{v}_{i_2} + \beta\sum_i\mathbf{u}_i^2\mathbf{v}_i^2 \\
&= \alpha^2\sum_{i_1,i_2,i_1\neq i_2}\mathbf{u}_{i_1}^2\mathbf{v}_{i_2}^2 - 2\alpha^2\sum_i\mathbf{u}_i^2\mathbf{v}_i^2 + \beta\sum_i\mathbf{u}_i^2\mathbf{v}_i^2 \\
&= \alpha^2\sum_i\mathbf{u}_i^2(1-\mathbf{v}_i^2) - 2\alpha^2\sum_i\mathbf{u}_i^2\mathbf{v}_i^2 + \beta\sum_i\mathbf{u}_i^2\mathbf{v}_i^2 \\
&= (\beta - 3\alpha^2)\sum_i\mathbf{u}_i^2\mathbf{v}_i^2 + \alpha^2 \ ,
\end{aligned}$$

where we used that $\sum_{i=1}^d \mathbf{v}_i\mathbf{u}_i = 0$ and $\|\mathbf{u}\|_2 = \|\mathbf{v}\|_2 = 1$, because $\mathbf{v}, \mathbf{u}$ are orthonormal. Next, we show that $\sum_i \mathbf{u}_i^2\mathbf{v}_i^2$ is less than $1/2$.

Claim E.5. Let $\mathbf{v}, \mathbf{u}$ be two orthonormal vectors. Then, $\sum_i \mathbf{u}_i^2\mathbf{v}_i^2 \leq 1/2$.

Proof of Claim E.5. Note that since $\sum_i \mathbf{u}_i^2 = \sum_i \mathbf{v}_i^2 = 1$ and $\sum_i \mathbf{u}_i\mathbf{v}_i = 0$, it holds

$$\begin{aligned}
1 &= \left(\sum_i \mathbf{u}_i^2\right)\left(\sum_i \mathbf{v}_i^2\right) = \sum_i \mathbf{u}_i^2\mathbf{v}_i^2 + \sum_{1\leq i<j\leq d}(\mathbf{u}_i^2\mathbf{v}_j^2 + \mathbf{u}_j^2\mathbf{v}_i^2) \ , \\
0 &= \left(\sum_i \mathbf{u}_i\mathbf{v}_i\right)^2 = \sum_i \mathbf{u}_i^2\mathbf{v}_i^2 + 2\sum_{1\leq i<j\leq d}\mathbf{u}_i\mathbf{v}_i\mathbf{u}_j\mathbf{v}_j \ .
\end{aligned}$$

Thus, summing the equalities above, we get

$$1 = 2\sum_i \mathbf{u}_i^2\mathbf{v}_i^2 + \sum_{1\leq i<j\leq d}(\mathbf{u}_i\mathbf{v}_j + \mathbf{u}_j\mathbf{v}_i)^2 \geq 2\sum_i \mathbf{u}_i^2\mathbf{v}_i^2,$$

therefore, we have $\sum_i \mathbf{u}_i^2\mathbf{v}_i^2 \leq 1/2$. □

Therefore, we have that if $c \geq 2$ then $\beta - 3\alpha^2 \geq 0$ hence $\mathbf{E}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}\left[V^2\right] \geq \alpha^2$. If $c \leq 2$, then it holds $\mathbf{E}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}\left[V^2\right] \geq (c/2)\alpha^2$. In summary we have $\mathbf{E}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}\left[V^2\right] \geq (c/2)\alpha^2$ for any $c > 0$. Furthermore, we can bound $\mathbf{E}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}\left[V^3\right]$ from above as the following. Using Cauchy-Schwarz, we have that $\mathbf{E}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}\left[V^3\right] \leq \max_{\mathbf{u}\in\mathcal{B}(1)} \mathbf{E}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}\left[(\mathbf{u}\cdot\mathbf{x})^6\right]$. Recall that $\mathbf{x}$ is a k-Heavy Tailed random variable with $k \geq 7$, hence $\mathbf{E}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}\left[(\mathbf{u}\cdot\mathbf{x})^6\right] \leq 1 + 6B/(k-6)$. Therefore, we have that

$$\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[V] = \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[|\mathbf{u}\cdot\mathbf{x}||\mathbf{v}\cdot\mathbf{x}|] \geq \frac{c^2\alpha^4}{4 + \frac{24B}{k-6}}. \tag{40}$$

This completes the proof of Claim E.4. □

Lemma E.6. *Let $Z = |\mathbf{u}\cdot\mathbf{x}||\mathbf{v}\cdot\mathbf{x}|\mathbb{1}\{\mathbf{v}\cdot\mathbf{x} \geq 0\}$. Assume that, there exists a constant $1 > C > 0$, so that $\mathbf{E}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[Z] \geq C$ and $\mathbf{E}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}\left[Z^2\right] \leq 1/C$. Then it holds*

$$\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}\left[(\mathbf{u}\cdot\mathbf{x})^2\mathbb{1}\left\{\mathbf{v}\cdot\mathbf{x} \geq \frac{C}{2}\left(\frac{C^3}{16B}\right)^{1/k}\right\}\right] \geq \frac{C^5}{64}\left(\frac{C^3}{16B}\right)^{2/k},$$

and

$$\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}\left[(\mathbf{v}\cdot\mathbf{x})^2\mathbb{1}\left\{\mathbf{v}\cdot\mathbf{x} \geq \frac{C}{2}\left(\frac{C^3}{16B}\right)^{1/k}\right\}\right] \geq \frac{C^5}{64}\left(\frac{C^3}{16B}\right)^{2/k}.$$

Proof of Lemma E.6. Using Paley-Zigmund inequality, we have that

$$\mathbf{Pr}\left[Z \geq \zeta \mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[Z]\right] \geq (1-\zeta)^2 \frac{\mathbf{E}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[Z]^2}{\mathbf{E}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}\left[Z^2\right]}.$$

Therefore, we have that $\mathbf{Pr}\left[Z \geq C/2\right] \geq C^3/4$, where $Z = |\mathbf{u}\cdot\mathbf{x}||\mathbf{v}\cdot\mathbf{x}|\mathbb{1}\{\mathbf{v}\cdot\mathbf{x} \geq 0\}$. For simplicity of notation, let's denote $|\mathbf{u}\cdot\mathbf{x}|$ and $|\mathbf{v}\cdot\mathbf{x}|\mathbb{1}\{\mathbf{v}\cdot\mathbf{x} \geq 0\}$ as a and b respectively. Then fixing some $t > 0$, it holds:

$$\begin{aligned}
\frac{C^3}{4} &\leq \mathbf{Pr}\left[ab \geq \frac{C}{2}\right] \\
&= \mathbf{Pr}\left[ab \geq \frac{C}{2}, a \geq t\sqrt{\frac{C}{2}}, b \leq t\sqrt{\frac{C}{2}}\right] + \mathbf{Pr}\left[ab \geq \frac{C}{2}, a \leq t\sqrt{\frac{C}{2}}, b \geq t\sqrt{\frac{C}{2}}\right] \\
&+ \mathbf{Pr}\left[ab \geq \frac{C}{2}, a \geq t\sqrt{\frac{C}{2}}, b \geq t\sqrt{\frac{C}{2}}\right] + \mathbf{Pr}\left[ab \geq \frac{C}{2}, \frac{1}{t}\sqrt{\frac{C}{2}} \leq a \leq t\sqrt{\frac{C}{2}}, \frac{1}{t}\sqrt{\frac{C}{2}} \leq b \leq t\sqrt{\frac{C}{2}}\right].
\end{aligned}$$

Note that $\mathcal{D}_\mathbf{x}$ is k-Heavy Tailed, thus, when $t \geq \sqrt{\frac{2}{C}}\left(\frac{16B}{C^3}\right)^{1/k}$, it holds

$$\mathbf{Pr}\left[a \geq t\sqrt{C/2}\right] = \mathbf{Pr}\left[|\mathbf{u}\cdot\mathbf{x}| \geq t\sqrt{C/2}\right] \leq \frac{B}{(t\sqrt{C/2})^k} \leq \frac{C^3}{16}.$$

Similarly, for $b = |\mathbf{v}\cdot\mathbf{x}|\mathbb{1}\{\mathbf{v}\cdot\mathbf{x} \geq 0\} \leq |\mathbf{v}\cdot\mathbf{x}|$ it holds $\mathbf{Pr}\left[b \geq t/\sqrt{C/2}\right] \leq C^3/16$. Therefore, $\mathbf{Pr}\left[ab \geq C/2\right]$ can be upper-bounded by

$$\frac{C^3}{4} \leq \mathbf{Pr}\left[ab \geq \frac{C}{2}\right] \leq \frac{3C^3}{16} + \mathbf{Pr}\left[ab \geq \frac{C}{2}, \frac{1}{t}\sqrt{\frac{C}{2}} \leq a \leq t\sqrt{\frac{C}{2}}, \frac{1}{t}\sqrt{\frac{C}{2}} \leq b \leq t\sqrt{\frac{C}{2}}\right].$$

Hence, we get

$$\mathbf{Pr}\left[a \geq \frac{1}{t}\sqrt{\frac{C}{2}}, b \geq \frac{1}{t}\sqrt{\frac{C}{2}}\right] \geq \mathbf{Pr}\left[ab \geq \frac{C}{2}, \frac{1}{t}\sqrt{\frac{C}{2}} \leq a \leq t\sqrt{\frac{C}{2}}, \frac{1}{t}\sqrt{\frac{C}{2}} \leq b \leq t\sqrt{\frac{C}{2}}\right] \geq \frac{C^3}{16},$$

where we choose $t = \sqrt{\frac{2}{C}}\left(\frac{16B}{C^3}\right)^{1/k}$. As a result,

$$\begin{aligned}
&\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}\left[(\mathbf{u}\cdot\mathbf{x})^2\mathbb{1}\left\{\mathbf{v}\cdot\mathbf{x}\geq\frac{C}{2}\left(\frac{C^3}{16B}\right)^{1/k}\right\}\right]\\
&\geq\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}\left[(\mathbf{u}\cdot\mathbf{x})^2\mathbb{1}\left\{\mathbf{v}\cdot\mathbf{x}\geq\frac{C}{2}\left(\frac{C^3}{16B}\right)^{1/k},|\mathbf{u}\cdot\mathbf{x}|\geq\frac{C}{2}\left(\frac{C^3}{16B}\right)^{1/k}\right\}\right]\\
&\geq\frac{C^5}{64}\left(\frac{C^3}{16B}\right)^{2/k}.
\end{aligned}$$

Similarly, we also have $\mathbf{E}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}[(\mathbf{v}\cdot\mathbf{x})^2\mathbb{1}\{\mathbf{v}\cdot\mathbf{x}\geq\frac{C}{2}(\frac{C^3}{16B})^{1/k}\}]\geq\frac{C^5}{64}(\frac{C^3}{16B})^{2/k}$. □

From Claim E.4 we know that $\mathbf{E}[Z]=\mathbf{E}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}\left[|\mathbf{u}\cdot\mathbf{x}||\mathbf{v}\cdot\mathbf{x}|\mathbb{1}\{\mathbf{v}\cdot\mathbf{x}\geq 0\}\right]=\frac{1}{2}\mathbf{E}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}\left[|\mathbf{u}\cdot\mathbf{x}||\mathbf{v}\cdot\mathbf{x}|\right]\gtrsim(k-6)c^2\alpha^4/B$. In addition, using Cauchy-Schwarz we have $\mathbf{E}[Z^2]\leq\max_{\mathbf{u}\in\mathcal{B}(1)}\mathbf{E}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}\left[(\mathbf{u}\cdot\mathbf{x})^4\right]\leq 5B$, where the last inequality comes from Fact C.4 (note that $\rho=1$ suffices in when $\mathcal{D}_\mathbf{x}$ is k-Heavy Tailed, $k\geq 7$). Thus, choosing C to be suffciently small absolute multiple of $(k-6)c^2\alpha^4/B$, it holds $\mathbf{E}[Z]\geq C$ and $\mathbf{E}[Z^2]\leq 1/C$. Let $\mathbf{v}=\mathbf{w}^*/\|\mathbf{w}^*\|_2$ and let $\mathbf{u}$ be any vector that is orthonormal to $\mathbf{v}$. Then by the results of Lemma E.6, we know that choosing $\gamma=\frac{C}{2}(\frac{C^3}{16B})^{1/k}$ and $\lambda=\frac{C^5}{64}(\frac{C^3}{16B})^{2/k}$, it holds $\mathbf{E}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}\left[\mathbf{x}\mathbf{x}^\top\mathbb{1}\{\mathbf{w}^*\cdot\mathbf{x}\geq\gamma\|\mathbf{w}^*\|_2\}\right]\succeq\lambda\mathbf{I}$. □

E.2.2. Discrete Gaussians

Here we show that our assumptions are satisfied for discrete multivariate Gaussians.

We will use the following standard definition of a discrete Gaussian.

Definition E.7 (Discrete Gaussian)**.** We define the discrete standard Gaussian distribution as follows: Fix $\theta\in\mathbb{R}_+$ with $\theta>0$. Then, the pmf of the discrete Gaussian distribution is given by

$$p(z)=\frac{1}{Z}\exp\left(-\frac{z^2}{2}\right)\mathbb{1}\{z\in\theta\mathbb{Z}\}\ ,$$

where Z is a normalization constant. Similarly, we define the high dimensional analogous as follows, we say that a random vector $\mathbf{x}\in\mathbb{R}^d$ follows the d-dimensinal discrete Gaussian distribution if $\mathbf{x}$ is a vector of d independent random variables, each of which follows the discrete Gaussian distribution.

Corollary E.8. *Let $\theta\in(0,1]$ and let $\mathcal{D}_\mathbf{x}$ be a d-dimensional discrete Gaussian distribution with parameter θ. Then, there exists an absolute constant $C>0$, so that $\mathcal{D}_\mathbf{x}$ satisfies Assumptions 2.3 and 2.4 with $\rho=1$, $B=e^9$, $\gamma=C(C^3/B)^{1/7}$ and $\lambda=C^5(C^3/B)^{2/7}$.*

Proof of Corollary E.8. We first show that the discrete Gaussian distribution is subgaussian with an appropriate parameter.

Lemma E.9. *Let $\theta\in(0,1]$ then the discrete Gaussian distribution is subgaussian with parameter $2\sqrt{2}$.*

Proof of Lemma E.9. By definition, a random variable X is D-subgaussian if

$$\mathbf{Pr}\left[|X|\geq t\right]\leq\exp(-t^2/D^2)\ .$$

Let $X\sim\mathcal{D}_\mathbf{x}$, where $\mathcal{D}_\mathbf{x}$ is a discrete Gaussian distribution with parameter θ. Fix $t\geq\theta$, Then, we have that

$$\begin{aligned}
\mathbf{Pr}\left[|X|\geq t\right]&\leq\frac{1}{Z}\sum_{z\in\theta\mathbb{Z},|z|\geq t}\exp\left(-\frac{z^2}{2}\right)\leq\frac{2}{Z}\int_{t/\theta-1}^{\infty}\exp\left(-\frac{z^2\theta^2}{2}\right)\mathrm{d}z\\
&=\frac{2}{\theta Z}\int_{t-\theta}^{\infty}\exp\left(-\frac{z^2}{2}\right)\mathrm{d}z\leq\frac{2}{\theta Z}\int_{t/2}^{\infty}\exp\left(-\frac{x^2}{2}\right)\mathrm{d}x\ ,
\end{aligned}$$

where for the first inequality, we used the integral mean value theorem and the monotonicity, i.e., $\int_k^{k+1}f(t)\mathrm{d}t=f(\xi)$ for $\xi\in(k,k+1)$ and that $f(k+1)\leq f(\xi)\leq f(k)$ because f is a decreasing function. Further note that $2\int_{t/2}^{\infty}\exp(-x^2/2)\,\mathrm{d}x=$

$\sqrt{2\pi}\mathrm{erfc}\big(\frac{t}{2\sqrt{2}}\big) \le \sqrt{2\pi}\exp(-t^2/8)$. It remains to show that θZ is lower-bouned. Note that $\exp(-x^2/2)$ is a decreasing function when $x \ge 0$, therefore for any $z \in \mathbb{Z}_+$ it holds $\exp(-(\theta z)^2/2) \ge \int_z^{z+1} \exp(-(\theta x)^2/2)\,\mathrm{d}x$. Thus, by definition,

$$\begin{aligned}
\theta Z &= \theta + 2\theta \sum_{z\in\mathbb{Z}_+} \exp((\theta z)^2/2) \\
&\ge \theta \int_0^1 \exp(-(\theta x)^2/2)\,\mathrm{d}x + 2\theta \sum_{z\in\mathbb{Z}_+} \int_z^{z+1} \exp(-(\theta x)^2/2)\,\mathrm{d}x \\
&\ge \int_0^\infty \exp(-t^2/2)\,\mathrm{d}t + \int_1^\infty \exp(-t^2/2)\,\mathrm{d}t = \sqrt{\frac{\pi}{2}}(2-\mathrm{erf}(1/\sqrt{2})) \ge \sqrt{\frac{\pi}{2}}.
\end{aligned}$$

Thus, combining these results, we get $\mathbf{Pr}\left[|X| \ge t\right] \le 4\exp(-t^2/8)$, thus discrete Gaussian is sub-Gaussian with parameter $D = 2\sqrt{2}$.

□

It remains to show that the discrete Gaussian distribution satisfies the requirements of Proposition E.3. To be specific, we show that it holds

Claim E.10. Let X be a discrete Gaussian random variable. Denote $\mathbf{E}[X^2]$ as α and $\mathbf{E}[X^4]$ as β. Then it holds $\alpha \le 1$ and $\beta \ge 1.25$.

Proof of Claim E.10. By Poisson summation formula, we know that it holds $\sum_{z\in\mathbb{Z}} f(z) = \sum_{z\in\mathbb{Z}} \hat{f}(z)$ where $\hat{f}(t)$ is the fourier transform of f, i.e., $\hat{f}(z) = \int_{-\infty}^{+\infty} f(x)e^{-2\pi i x t}\mathrm{d}x$. It is easy to calculate that for $f(z) = \theta^2 z^2 \exp\big(-\frac{\theta^2 z^2}{2}\big)$ we have

$$\hat{f}(t) = \frac{\sqrt{2\pi}}{\theta^3}(\theta^2 - 4\pi^2 t^2)\exp\bigg(-\frac{2\pi^2 t^2}{\theta^2}\bigg),$$

and for $g(z) = \exp\big(-\frac{\theta^2 z^2}{2}\big)$, we have

$$\hat{g}(t) = \frac{\sqrt{2\pi}}{\theta}\exp\bigg(-\frac{2\pi^2 t^2}{\theta^2}\bigg).$$

Thus, by definition,

$$\alpha = \mathbf{E}[X^2] = \frac{\sum_{z\in\mathbb{Z}} f(z)}{\sum_{z\in\mathbb{Z}} g(z)} = \frac{\sum_{z\in\mathbb{Z}} \hat{f}(z)}{\sum_{z\in\mathbb{Z}} \hat{g}(z)} = 1 - \frac{\sum_{z\in\mathbb{Z}} \frac{4\pi^2 z^2}{\theta^2}\exp\big(-\frac{2\pi^2 z^2}{\theta^2}\big)}{\sum_{z\in\mathbb{Z}} \exp\big(-\frac{2\pi^2 z^2}{\theta^2}\big)} \le 1.$$

For $\mathbf{E}[X^4]$, note that $t^4\exp(-t^2/2)$ is an increasing function when $t \in (0,2)$ and is decreasing when $t \in (2,\infty)$. Thus, denote $\Delta z = 1$, then by the property of integral, it holds

$$\begin{aligned}
\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_{\mathbf{x}}}\left[X^4\right] &= \frac{1}{\theta Z}\sum_{z\in\mathbb{Z}} (\theta z)^4 \exp(-(\theta z)^2/2)\theta(\Delta z) \\
&\ge \frac{2}{\theta Z}\bigg(\int_0^{2-\theta} t^4\exp(-t^2/2)\,\mathrm{d}t + \int_2^\infty t^4\exp(-t^2/2)\,\mathrm{d}t\bigg) \\
&\ge \frac{2}{\theta Z}\bigg(\int_0^1 t^4\exp(-t^2/2)\,\mathrm{d}t + 2.06\bigg) \ge \frac{4.4}{\theta Z},
\end{aligned}$$

where the second inequality is due to the fact that $\int_2^\infty t^4\exp(-t^2/2)\,\mathrm{d}t \ge 2.06$ and $\theta \in (0,1]$. Further, in the last inequality we used the fact that $\int_0^1 t^4\exp(-t^2/2)\,\mathrm{d}t \ge 0.14$.

Now, note that for θZ, it holds

$$\theta Z = \theta + 2\sum_{z\in\mathbb{Z}_+} \exp(-(\theta z)^2/2)\theta(\Delta z) \le \theta + 2\int_0^\infty \exp\bigg(-\frac{t^2}{2}\bigg)\,\mathrm{d}t \le 1 + \sqrt{2\pi}.$$

Therefore, combining with upper-bound on θZ, it holds $\mathbf{E}[X^4] \ge 4.4/(1+\sqrt{2\pi}) \ge 1.25$. □

Now since $\alpha \leq 1$, $\beta \geq 1.25\alpha^2$ and discrete Gaussian is $2\sqrt{2}$-sub-Gaussian as proved in Lemma E.9, we have $\mathbf{Pr}\left[|\mathbf{u}\cdot\mathbf{x}| \leq r\right] \leq \exp(-r^2/8) \leq e^9/r^7$. Thus, the conditions in Proposition E.3 are satisfied with parameters $B = e^9$, $c = 0.25$, $k = 7$. Thus, choosing C to be a small multiple of $c^2\alpha^4/B$, then since according to Claim E.4, $\mathbf{E}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}\left[Z\right] \gtrsim c^2\alpha^4/B \geq C$ and $\mathbf{E}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}\left[Z^2\right] \leq 5B \leq 1/C$, thus, by Proposition E.3, we know that Assumption 2.3 is satisfied with parameters $\gamma = \frac{C}{2}(\frac{C^3}{16B})^{1/7}$ and $\lambda = \frac{C^5}{64}(\frac{C^3}{16B})^{2/7}$.

□

E.2.3. Uniform Discrete Distribution on $\{-1, 0, 1\}^d$

Finally, we show that the uniform distribution on a hyper-grid satisfies our assumptions.

Corollary E.11. *Let $\mathcal{D}_\mathbf{x}$ be a d-dimensional uniform distribution over the $\{-1, 0, 1\}^d$. Then, there exists an absolute constant $C > 0$, so that $\mathcal{D}_\mathbf{x}$ satisfies Assumptions 2.3 and 2.4 with $B = 1$, $\rho = 1$, $\gamma = C(C^3/B)^{1/7}$ and $\lambda = C^5(C^3/B)^{2/7}$.*

Proof of Corollary E.11. Note that the distribution is 1-sub-Gaussian. Now since $\beta = \mathbf{E}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}\left[\mathbf{x}_i^4\right] = 2/3$ and $\alpha = \mathbf{E}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}\left[\mathbf{x}_i^2\right] = 2/3$, therefore, $\beta = 1.5\alpha^2$. Thus, the conditions in Proposition E.3 are satisfied with parameters $B = 1$, $\alpha = 2/3$, $c = 0.5$, $k = 7$. Now choosing C to be a small multiple of $c^2\alpha^4/B$, then since according to Claim E.4, $\mathbf{E}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}\left[Z\right] \gtrsim c^2\alpha^4/B \geq C$ and $\mathbf{E}_{\mathbf{x}\sim\mathcal{D}_\mathbf{x}}\left[Z^2\right] \leq 5B \leq 1/C$, thus, by Proposition E.3, we know that Assumption 2.3 is satisfied with parameters $\gamma = \frac{C}{2}(\frac{C^3}{16B})^{1/7}$ and $\lambda = \frac{C^5}{64}(\frac{C^3}{16B})^{2/7}$.

□

F. Extension to Certain Non-Monotone Activations

In this section, we extend our algorithmic results to certain cases where the activation function is not monotone. Specifically, we will consider activations like GeLU (Hendrycks & Gimpel, 2016): $\sigma_{GeLU}(t) = t\Phi(t)$, where $\Phi(t)$ is the cdf. of the standard normal random variable $\mathcal{N}(0, 1)$ and Swish (Ramachandran et al., 2017) defined by $\sigma_{Swish}(t) = \frac{t}{1+\exp(-t)}$.

Definition F.1 (Non-Monotonic (α, β)-Unbounded Activations)**.** Let $\sigma : \mathbb{R} \mapsto \mathbb{R}$ be an activation function and let $\alpha, \beta > 0$. We say that σ is non-monotonic (α, β)-unbounded if it satisfies the following assumptions:

1. $\sigma(t_2) \geq 0 \geq \sigma(t_1)$ for any $t_2 \geq 0 \geq t_1$;
2. σ is α-Lipschitz; and
3. $\sigma'(t) \geq \beta$ for all $t \in (0, \infty)$.

As mentioned above, Definition F.1 contains GeLU and Swish. Indeed, one can show that $\sigma_{GeLU}(t)$ is actually non-monotonic $(1.1, 1/2)$-unbounded, and $\sigma_{Swish}(t)$ is non-monotonic $(1.2, 0.4)$-unbounded. We include the following Figure 1 of these activations to provide the readers with a better geometric intuition.

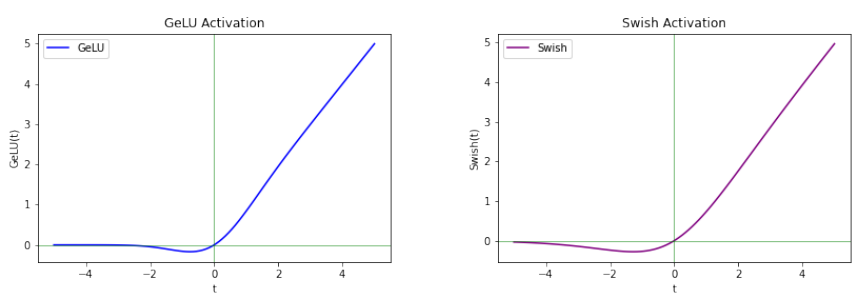

Figure 1: Non-Monotonic (α, β)-Unbounded Activation Examples: GeLU and Swish

Now, we show that truncating a non-monotonic (α, β)-unbounded activation σ to $\hat{\sigma}(t) = [\sigma(t)]_+$ and cutting off the negative part of y induces only a small L_2^2 error at point $\mathbf{w}^*$, i.e., $\mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}\left[(\hat{\sigma}(\mathbf{w}^*\cdot\mathbf{x}) - y\mathbb{1}\{y \geq 0\})^2\right] \leq \mathrm{OPT}$, implying that we can consider running an algorithm similar to Algorithm 1 on $\hat{\sigma}(t)$ and truncated y.

Lemma F.2. *Let* $\mathbf{w}^* = \operatorname{argmin}_{\mathbf{w}\in\mathcal{B}(W)} \mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}\left[(\sigma(\mathbf{w}\cdot\mathbf{x})-y)^2\right] = \operatorname{argmin}_{\mathbf{w}\in\mathcal{B}(W)} \mathcal{L}_2^{\mathcal{D},\sigma}(\mathbf{w})$ *and denote* $\mathcal{L}_2^{\mathcal{D},\sigma}(\mathbf{w}^*)$ *as* OPT. *Define* $\hat{y} = [y]_+$ *and* $\hat{\sigma}(t) = [\sigma(t)]_+$. *Then:*

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

Proof. The proof follows similar ideas in Lemma D.8. Since $[t]_+$ is a non-expansive projection from $\mathbb{R}$ to $\mathbb{R}^+$, we have $|[t_1]_+ - [t_2]_+| \le |t_1 - t_2|$ for any $t_1, t_2 \in \mathbb{R}$. Thus, we get

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

□

The truncated activation function is a monotonic (α,β)-unbounded since $\hat{\sigma}(t)$ is increasing when $t \ge 0$ and $\hat{\sigma}(t) = 0$ when $t \le 0$. Thus, when Assumptions 2.3 and 2.4 hold with respect to distribution $\mathcal{D}_{\mathbf{x}}$ and $\hat{\sigma}$, we can use a slightly modified algorithm Algorithm 3 that works as efficiently as Algorithm 1 since Lemma 2.5 and Theorem 3.3 can be applied to activation $\hat{\sigma}(t)$ with minor modifications. Formally, we have the following corollaries:

Corollary F.3. *Let* $\sigma(t)$ *be a non-monotonic* (α,β)*-unbounded activation function satisfying Definition F.1. Suppose that Assumption 2.3 and Assumption 2.4 holds. Further denote* $\hat{\sigma}(t) = \sigma(t)\mathbb{1}\{t \ge 0\}$. *Then the noise-free surrogate loss* $\bar{\mathcal{L}}_{\mathrm{sur}}^{\mathcal{D},\hat{\sigma}}$ *with respect to activation function* $\hat{\sigma}(t)$ *is* $\Omega(\lambda^2\gamma\beta\rho/B)$*-sharp in the ball* $\mathcal{B}(2\|\mathbf{w}^*\|_2)$*, i.e.,*

$$\nabla\bar{\mathcal{L}}_{\mathrm{sur}}^{\mathcal{D},\hat{\sigma}}(\mathbf{w})\cdot(\mathbf{w}-\mathbf{w}^*) \gtrsim \lambda^2\gamma\beta\rho/B\|\mathbf{w}-\mathbf{w}^*\|_2^2,\ \ \forall\mathbf{w}\in\mathcal{B}(2\|\mathbf{w}^*\|_2).$$

Proof. Since $\hat{\sigma}$ is monotonic (α,β)-unbounded, we have proven in Lemma 2.5 for monotonic (α,β)-unbounded activations and distribution $\mathcal{D}_{\mathbf{x}}$ satisfying Assumption 2.2 to Assumption 2.4, $\bar{\mathcal{L}}_{\mathrm{sur}}^{\mathcal{D},\hat{\sigma}}$ is $\bar{\mu}$ sharp with the parameter $\bar{\mu} = \Omega(\lambda^2\gamma\beta\rho/B)$. □

Corollary F.4. *Let* $\sigma(t)$ *be a non-monotonic* (α,β)*-unbounded activation function, satisfying Definition F.1. Fix* $\epsilon > 0$ *and* $W > 0$ *and suppose Assumptions 2.3 and 2.4 hold. Let* OPT *denote the minimum value of the* L_2^2 *loss i.e.,*

$$\mathrm{OPT} = \min_{\mathbf{w}\in\mathcal{B}(W)} \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}\left[(\sigma(\mathbf{w}\cdot\mathbf{x})-y)^2\right].$$

Let $\mu := \mu(\lambda,\gamma,\beta,\rho,B)$ *be a sufficiently small multiple of* $\lambda^2\gamma\beta\rho/B$*, and let* $M = \alpha W H_2^{-1}\left(\frac{\epsilon}{4\alpha^2W^2}\right)$. *Further, choose parameter* r_ϵ *large enough so that* $H_4(r_\epsilon)$ *is a sufficiently small multiple of* ϵ. *Then after*

$$T = \widetilde{\Theta}\left(\frac{B^2\alpha^2}{\rho^2\mu^2}\log\left(\frac{W}{\epsilon}\right)\right)$$

iterations with batch size $N = \widetilde{\Omega}(dT(r_\epsilon^2+\alpha^2M^2))$*, Algorithm 3 converges to a point* $\mathbf{w}^{(T)}$ *such that* $\mathcal{L}_2^{\mathcal{D},\sigma}(\mathbf{w}^{(T)}) = O\left(\frac{B^2\alpha^2}{\rho^2\mu^2}\mathrm{OPT}\right)+\epsilon$ *with probability at least* $2/3$.

Proof. First observe that since the α-Lipschitz property remains for non-monotonic (α,β)-unbounded functions, the following is still valid:

$$\begin{aligned}
\mathcal{L}_2^{\mathcal{D},\sigma}(\mathbf{w}) &= \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}\left[(\sigma(\mathbf{w}\cdot\mathbf{x})-y)^2\right] \\
&\le 2\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_{\mathbf{x}}}\left[(\sigma(\mathbf{w}\cdot\mathbf{x})-\sigma(\mathbf{w}^*\cdot\mathbf{x}))^2\right] + 2\mathop{\mathbf{E}}_{(\mathbf{x},y)\sim\mathcal{D}}\left[(\sigma(\mathbf{w}^*\cdot\mathbf{x})-y)^2\right] \\
&\le 2\alpha^2\mathop{\mathbf{E}}_{\mathbf{x}\sim\mathcal{D}_{\mathbf{x}}}\left[((\mathbf{w}-\mathbf{w}^*)\cdot\mathbf{x})^2\right] + 2\mathrm{OPT} \\
&\le (10B\alpha^2/\rho)\|\mathbf{w}-\mathbf{w}^*\|_2^2 + 2\mathrm{OPT}. \qquad (41)
\end{aligned}$$

Now denote $\hat{\epsilon} = \mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}\left[(\hat{\sigma}(\mathbf{w}\cdot\mathbf{x})-\hat{y})^2\right]$. If one can show that after $T = \widetilde{\Theta}\left(B^2\alpha^2/(\rho^2\mu^2)\log(W/\epsilon)\right)$ iterations with a large enough batch size $N = \widetilde{\Omega}(dT(r_\epsilon^2+\alpha^2M^2))$, Algorithm 3 generates a point $\mathbf{w}^{(T)}$ such that it holds

$$\|\mathbf{w}^{(T)}-\mathbf{w}^*\|_2^2 \le \frac{64B}{\rho\mu^2}\hat{\epsilon}+\epsilon, \qquad (42)$$

Algorithm 3 Stochastic Gradient Descent on Surrogate Loss For Non-Monotonic (α, β)-Unbounded Activations

Input: Iterations: T, sample access from $\mathcal{D}$, batch size N, step size η, bound M.
Initialize $\mathbf{w}^{(0)} \leftarrow \mathbf{0}$.
for $t = 1$ **to** T **do**
Draw N samples $\{(\mathbf{x}(j), y(j))\}_{j=1}^{N} \sim \mathcal{D}$.
for each $j \in [N]$, $y(j) \leftarrow \min([y(j)]_+, M)$
Let $\hat{\sigma}(t) = [\sigma(t)]_+$, calculate

$$\mathbf{g}^{(t)} \leftarrow \frac{1}{N} \sum_{j=1}^{N} (\hat{\sigma}(\mathbf{w}^{(t)} \cdot \mathbf{x}(j)) - y(j))\mathbf{x}(j),$$

$\mathbf{w}^{(t+1)} \leftarrow \mathbf{w}^{(t)} - \eta \mathbf{g}^{(t)}$.
end for
Output: The weight vector $\mathbf{w}^{(T)}$.

then, since we showed in Lemma F.2 that $\hat{\epsilon} \leq \mathrm{OPT}$, combining with Equation (41) we immediately get

$$\mathcal{L}_2^{\mathcal{D},\sigma}(\mathbf{w}) \leq \left(2 + \frac{640B^2\alpha^2}{\rho^2\mu^2}\right) \mathrm{OPT} + (10B\alpha^2/\rho)\epsilon,$$

thus completing the corollary.

In order to prove the claim above and Equation (42), one only needs to observe that $\hat{\sigma}(t)$ is monotonic (α, β)-unobounded, and Assumption 2.2 to Assumption 2.4 holds for $\hat{\sigma}(t)$ and the distribution $\mathcal{D}_{\mathbf{x}}$. Therefore, the exact same techniques for proving Theorem 3.3 (see Appendix D.2) can be applied. Results similar to Lemma D.8 and Lemma D.9 still hold for activation function $\hat{\sigma}(t)$ and data points $(\mathbf{x}, \hat{y})$, with the only difference being that we have $\mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}\left[(\hat{\sigma}(\mathbf{w}^* \cdot \mathbf{x}) - \hat{y})^2\right] = \hat{\epsilon}$ instead of OPT. Moreover, it has proven in Corollary F.3 that $\tilde{\mathcal{L}}_{\mathrm{sur}}^{\mathcal{D},\hat{\sigma}}$ is $\bar{\mu}$ sharp, therefore by Proposition 3.2 we know $\mathcal{L}_{\mathrm{sur}}^{\mathcal{D},\sigma}$ is $\bar{\mu}/2$ sharp. Thus, Equation (42) follows from the same steps and the same choice of parameters as in the proof of Theorem 3.3 (see Appendix D.2).

□