Learning a Single Neuron Robustly to Distributional Shifts and Adversarial Label Noise

Shuyao Li[*]
University of Wisconsin-Madison
shuyao.li@wisc.edu

Sushrut Karmalkar[*]
University of Wisconsin-Madison
skarmalkar@wisc.edu

Ilias Diakonikolas
University of Wisconsin-Madison
ilias@cs.wisc.edu

Jelena Diakonikolas
University of Wisconsin-Madison
jelena@cs.wisc.edu

Abstract

We study the problem of learning a single neuron with respect to the L_2^2-loss in the presence of adversarial distribution shifts, where the labels can be arbitrary, and the goal is to find a "best-fit" function. More precisely, given training samples from a reference distribution p_0, the goal is to approximate the vector $\boldsymbol{w}^*$ which minimizes the squared loss with respect to the worst-case distribution that is close in χ^2-divergence to p_0. We design a computationally efficient algorithm that recovers a vector $\hat{\boldsymbol{w}}$ satisfying $\mathbb{E}_{p^*}(\sigma(\hat{\boldsymbol{w}} \cdot \boldsymbol{x}) - y)^2 \leq C\, \mathbb{E}_{p^*}(\sigma(\boldsymbol{w}^* \cdot \boldsymbol{x}) - y)^2 + \epsilon$, where $C > 1$ is a dimension-independent constant and $(\boldsymbol{w}^*, p^*)$ is the witness attaining the min-max risk $\min_{\boldsymbol{w}\,:\,\|\boldsymbol{w}\| \leq W} \max_{p} \mathbb{E}_{(\boldsymbol{x},y)\sim p}(\sigma(\boldsymbol{w}\cdot\boldsymbol{x}) - y)^2 - \nu\chi^2(p, p_0)$. Our algorithm follows a primal-dual framework and is designed by directly bounding the risk with respect to the original, nonconvex L_2^2 loss. From an optimization standpoint, our work opens new avenues for the design of primal-dual algorithms under structured nonconvexity.

1 Introduction

The problem of learning a single neuron from randomly drawn labeled examples is a fundamental problem extensively studied in the machine learning literature. Given labeled examples $\{(\boldsymbol{x}_i, y_i) : (\boldsymbol{x}_i, y_i) \in \mathbb{R}^d \times \mathbb{R}\}_{i=1}^N$ drawn from a reference distribution p_0, the goal in this context is to recover a parameter vector $\boldsymbol{w}_0^*$ that minimizes the squared loss $\Lambda_{\sigma,p_0}(\boldsymbol{w})$ over a ball of radius $W > 0$:

$$\boldsymbol{w}_0^* := \operatorname*{arg\,min}_{\boldsymbol{w}\in\mathbb{R}^d : \|\boldsymbol{w}\|_2 \leq W} \Lambda_{\sigma,p_0}(\boldsymbol{w}); \quad \Lambda_{\sigma,p_0}(\boldsymbol{w}) := \mathbb{E}_{(\boldsymbol{x},y)\sim p_0}(\sigma(\boldsymbol{w}\cdot\boldsymbol{x}) - y)^2, \tag{1}$$

where $\sigma : \mathbb{R} \to \mathbb{R}$ is a known (typically non-linear) non-decreasing activation function (e.g., the ReLU activation $\sigma(t) = \max(0, t)$) and we denote by $\mathrm{OPT}_0 = \min_{\boldsymbol{w}:\|\boldsymbol{w}\|_2 \leq W} \Lambda_{\sigma,p_0}(\boldsymbol{w})$ the minimum squared loss. In the realizable setting — where $y = \sigma(\boldsymbol{w}_0^* \cdot \boldsymbol{x})$ and thus $\mathrm{OPT}_0 = 0$ — this problem is well-understood and by now part of the folklore (see, e.g., [KS09; Kak+11; Sol17; YS20]). The results for the realizable setting also naturally extend to zero-mean bounded-variance label noise.

The more realistic agnostic model [Hau92; KSS92] (a.k.a. adversarial label noise) aims to identify the best-fitting neuron for a reference distribution of the examples, without any assumptions on label structure. However, it is known that in this setting finding a parameter vector with square loss $\mathrm{OPT}_0 + \epsilon$ requires $d^{\mathrm{poly}(1/\epsilon)}$ time, even if the $\boldsymbol{x}$-marginal distribution is Gaussian [GKK19; DKZ20; GGK20; Dia+21; DKR23]. Even if we relax our goal to achieve error $O(\mathrm{OPT}_0) + \epsilon$, efficient

[*] Equal contribution.

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

algorithms only exist under strong distributional assumptions. In fact, without such assumptions, this problem is NP-hard [Sim02; MR18]. Recent work has also shown that (under cryptographic assumptions) no polynomial-time constant-factor improper learner exists even for distributions supported on the unit ball [Dia+22b]. Given these intractability results, recent work has focused on developing efficient constant-factor approximate learners under minimal distributional assumptions (see, e.g.,[Dia+20; FCG20; Dia+22a; ATV23; Gol+23; Wan+23a; Zar+24]).

This recent progress notwithstanding, prior work primarily focused on the setting where only the labels might be corrupted, without considering possible distributional shifts or heterogeneity of the data. Such distributional corruptions are frequently observed in practice and have motivated a long line of research in areas such as domain adaptation and (related to it) distributionally robust optimization (DRO); see e.g., [BEN09; ND16; RM22; Bla+24] and references therein. Thus, the main question motivating our work is:

How do adversarial changes in the underlying distribution *impact the learnability of a neuron?*

We study this question within the DRO framework, where the goal is to minimize the model's loss on a worst-case distribution from a set of distributions close to the reference distribution.[2] This set of distributions, known as the ambiguity set, models possible distributional shifts of the data. In addition to being interesting on its own merits, the DRO framework arises in diverse contexts, including algorithmic fairness [Has+18b] and class imbalance [Xu+20]. Moreover, it has recently found a range of applications in reinforcement learning [Kal+22; Liu+22; Lot+23; Wan+23b; Yan+23; Yu+23], robotics [Sha+20], language modeling [Liu+21], sparse neural network training [Sap+23], and defense against model extraction [Wan+23c].

Despite a range of impressive results in the DRO literature (see, e.g., recent surveys [Kuh+19; CP20; RM22; Bla+24] and references therein), algorithmic results with rigorous approximation guarantees for the loss have almost exclusively been obtained under fairly strong assumptions about the loss function involving both convexity and either smoothness or Lipschitzness, with linear regression being the prototypical example; see, e.g., [CP18; BMN21; DN21]. Unfortunately, this vanilla setting does not capture a range of machine learning applications, where a typical loss function is nonconvex. In particular, even the simplest ReLU learning problem in the realizable setting (with noise-free labels) is nonconvex. Further, existing DRO approaches for nonconvex loss functions such as [SND18; Qi+21] only guarantee convergence to a stationary point, which is insufficient for learning a ReLU neuron even without distributional ambiguity [YS20]. Motivated by this gap in our understanding, in this work we initiate a rigorous algorithmic investigation of learning a neuron (arguably the simplest non-convex problem) in the DRO setting. We hope that this work will stimulate future research in this direction, potentially addressing more complex models in a principled manner.

Due to space constraints, we defer further discussion of related work to Appendix A.

1.1 Problem Setup

To formally define our setting, we recall the definition of χ^2-divergence between distributions p and p', given by $\chi^2(p, p') := \int \big(\frac{\mathrm{d}p}{\mathrm{d}p'} - 1\big)^2 \mathrm{d}p'$. We focus on the class of monotone unbounded activations introduced in [Dia+22a], for which we additionally assume convexity. Example activations in this class include the ReLU, leaky ReLU, exponential linear unit (ELU), and normalized[3] SoftPlus.

Definition 1.1 (Unbounded [Dia+22a] + Convex Activation)**.** Let $\sigma : \mathbb{R} \to \mathbb{R}$ be a non-decreasing convex function, and let $\alpha, \beta > 0$. We say σ is (α, β)-unbounded if it satisfies the following: (i) σ is β-Lipschitz; (ii) $\sigma(t_1) - \sigma(t_2) \geq \alpha(t_1 - t_2)$ for all $t_1 \geq t_2 \geq 0$, and (iii) $\sigma(0) = 0$.

To formally state the problem, we further define the loss, risk, and optimal value (denoted by OPT).

Definition 1.2 (Loss, Risk, and OPT)**.** Given a regularization parameter ν and a reference distribution p_0, let $\mathscr{P} = \mathscr{P}(p_0)$ denote the set of all distributions that are absolutely continuous with respect to

[2] We contrast here robustness to perturbed data *distribution* studied within the DRO framework to robustness to perturbed *data examples* referred to as the adversarial robustness in modern deep learning literature (e.g., [GSS14]). Our paper is concerned with the former (and not the latter) model of robustness.

[3] Normalization, which ensures $\sigma(0) = 0$, is without loss of generality, as it corresponds to a simple change of variable: $\hat{\sigma}(t) \leftarrow \sigma(t) - \sigma(0)$ and $\hat{y} \leftarrow y + \sigma(0)$, which does not affect the loss value or its approximation.

$\mathscr{p}_0$ and $\mathscr{B}(W) := \{\boldsymbol{w} : \|\boldsymbol{w}\|_2 \leq W\}$. We define the following:

$$\begin{aligned}
L_\sigma(\boldsymbol{w}, \mathscr{p}; \mathscr{p}_0) &:= \mathbb{E}_{(\boldsymbol{x},y)\sim\mathscr{p}}(\sigma(\boldsymbol{w}\cdot\boldsymbol{x}) - y)^2 - \nu\chi^2(\mathscr{p}, \mathscr{p}_0) = \Lambda_{\sigma,\mathscr{p}}(\boldsymbol{w}) - \nu\chi^2(\mathscr{p}, \mathscr{p}_0),\\
R(\boldsymbol{w}; \mathscr{p}_0) &:= \max_{\mathscr{p}\in\mathscr{P}(\mathscr{p}_0)} L_\sigma(\boldsymbol{w}, \mathscr{p}; \mathscr{p}_0),\ \mathscr{q}_{\boldsymbol{w}} := \operatorname*{arg\,max}_{\mathscr{p}\in\mathscr{P}(\mathscr{p}_0)} L_\sigma(\boldsymbol{w}, \mathscr{p}; \mathscr{p}_0),\\
\boldsymbol{w}^* &:= \operatorname*{arg\,min}_{\boldsymbol{w}\in\mathscr{B}(W)} R(\boldsymbol{w}; \mathscr{p}_0), \quad \mathscr{p}^* := \mathscr{q}_{\boldsymbol{w}^*},\\
\mathrm{OPT} &:= \mathbb{E}_{(\boldsymbol{x},y)\sim\mathscr{p}^*}(\sigma(\boldsymbol{w}^*\cdot\boldsymbol{x}) - y)^2 = \Lambda_{\sigma,\mathscr{p}^*}(\boldsymbol{w}^*).
\end{aligned}$$

We say that $L_\sigma(\boldsymbol{w}, \mathscr{p}; \mathscr{p}_0)$ is the regularized square loss function of a vector $\boldsymbol{w}$ and a distribution $\mathscr{p} \in \mathscr{P}$; and $R(\boldsymbol{w}; \mathscr{p}_0)$ is the DRO risk of $\boldsymbol{w}$ with respect to $\mathscr{p}_0$. We call $\mathscr{p}^*$ the target distribution.

The minimization of the DRO risk as defined above corresponds to the regularized/penalized DRO formulation studied in prior work; see, e.g., [SND18; Wan+23c; MDH24]. An alternate formulation would have been to instead optimize over a restricted domain. The two are equivalent because of Lagrangian duality. We show in Claim E.1 a concrete relation between our regularization parameter ν and the chi-squared distance between the population distribution $\mathscr{p}_0$ and the target distribution $\mathscr{p}^*$. We further require that ν is sufficiently large to ensure that the resulting $\chi^2(\mathscr{p}^*, \mathscr{p}_0)$ is smaller than an absolute constant, which is in line with the DRO being used for not too large ambiguity sets [RM22].

Empirical Version If the reference distribution is the uniform distribution on N labeled examples $(\boldsymbol{x}_i, y_i) \in \mathbb{R}^d \times \mathbb{R}$ drawn from $\mathscr{p}_0$, we call it $\widehat{\mathscr{p}}_0 = \widehat{\mathscr{p}}_0(N)$, and similarly define $\widehat{\mathscr{p}} \in \mathscr{P}(\widehat{\mathscr{p}}_0)$. Note that $R(\boldsymbol{w}^*; \widehat{\mathscr{p}}_0) = \max_{\widehat{\mathscr{p}}\in\mathscr{P}(\widehat{\mathscr{p}}_0)} \mathbb{E}_{(\boldsymbol{x},y)\sim\widehat{\mathscr{p}}}(\sigma(\boldsymbol{w}^*\cdot\boldsymbol{x}) - y)^2 - \nu\chi^2(\widehat{\mathscr{p}}, \widehat{\mathscr{p}}_0)$; if we let $\widehat{\mathscr{p}}^*$ denote the distribution that achieves the maximum, $\widehat{\mathscr{p}}^*$ has the same support as $\widehat{\mathscr{p}}_0$ and can be interpreted as the reweighting of the samples that maximizes the regularized loss.

Formally, our goal is to solve the following learning problem.

Problem 1.3 (Robustly Learning a Single Neuron Under Distributional Shifts)**.** Given error parameters $\epsilon, \delta \in (0, 1)$, regularization parameter $\nu > 0$, set radius $W > 0$, and sample access to labeled examples $(\boldsymbol{x}, y)$ drawn i.i.d. from an unknown reference distribution $\mathscr{p}_0$, output a parameter vector $\hat{\boldsymbol{w}} \in \mathscr{B}(W)$ that is competitive with the DRO risk minimizer $\boldsymbol{w}^* = \arg\min_{\boldsymbol{w}\in\mathscr{B}(W)} R(\boldsymbol{w}; \mathscr{p}_0)$ in the sense that with probability at least $1 - \delta$, $\|\hat{\boldsymbol{w}} - \boldsymbol{w}^*\|_2^2 \leq C\mathrm{OPT} + \epsilon$ for an absolute constant C.

While the stated goal is expressed in terms of $\|\hat{\boldsymbol{w}} - \boldsymbol{w}^*\|_2$, under mild distributional assumptions that we make on the reference and target distributions, this guarantee implies being competitive with the best-fit function on $\mathscr{p}^*$ in terms of both the square loss and the risk, namely $\Lambda_{\sigma,\mathscr{p}^*}(\hat{\boldsymbol{w}}) = O(\mathrm{OPT}) + \epsilon$ and $R(\boldsymbol{w}, \mathscr{p}_0) - \min_{\boldsymbol{w}\in\mathscr{B}(W)} R(\hat{\boldsymbol{w}}, \mathscr{p}_0) \leq O(\mathrm{OPT}) + \epsilon$. Further, our algorithm is primal-dual and it outputs a distribution $\widehat{\mathscr{p}}$ that is close to $\widehat{\mathscr{p}}^*$ in the chi-squared divergence.

Since the solution to Problem 1.3 has an error of $O(\mathrm{OPT}) + \epsilon$, when we use the term "convergence" in our paper, we refer to the following weaker notion: the iterates of our algorithm *converge* to the (set of) solutions such that asymptotically all iterates lie within the set of $O(\mathrm{OPT}) + \epsilon$ solutions, which are the target solutions, as stated in Problem 1.3.

1.2 Main Result

Our main contribution is the first polynomial sample and time algorithm for learning a neuron in a distributionally robust setting for a broad class of activations (Definition 1.1) and under mild distributional assumption on the target distribution (Assumptions 2.1 and 2.2 in Section 2.1).

Theorem 1.4 (Main Theorem — Informal)**.** *Suppose that the learner has access to $N = \tilde{\Omega}(d/\epsilon^2)$ samples drawn from the reference distribution $\mathscr{p}_0$. If all samples are bounded and the distribution $\mathscr{p}^*$ satisfies the "margin-like" condition and concentration (Assumptions 2.1 and 2.2 in Section 2.1), then after $\widetilde{O}(d\log(1/\epsilon))$ iterations, each running in sample near-linear time, with high probability Algorithm 1 recovers $\hat{\boldsymbol{w}}$ such that $\|\hat{\boldsymbol{w}} - \boldsymbol{w}^*\|_2^2 \leq C\,\mathrm{OPT} + \epsilon$, for an absolute constant C.*

We emphasize that Theorem 1.4 simultaneously addresses two types of robustness: firstly, robustness concerning labels (y); and secondly, robustness due to shifts in the distribution ($\mathscr{p}_0$ being perturbed). This result is new even when specialized to any nontrivial activation like ReLU, realizable case (where $\mathrm{OPT} = 0$), and the simplest Gaussian $\boldsymbol{x}$-marginal distribution. Without distributional robustness, existing approaches, as previously discussed, yield an error of $O(\mathrm{OPT}) + \epsilon$ under certain

$\boldsymbol{x}$-marginal conditions. We demonstrate that this error rate can be also achieved with respect to $\mathpzc{p}^*$ in a distributionally robust context, as long as $\mathpzc{p}^*$ meets the same conditions specified in [Wan+23a] — among the mildest in the literature addressing non-distributionally robust agnostic setting.

1.3 Technical Overview

Our technical approach relies on three main components, described below:

Local Error Bounds Our work is inspired by optimization-theory local error bounds ("sharpness") obtained for learning a single neuron with monotone unbounded activations under structured distributions without considering distributional shift or ambiguity [MBM18; Wan+23a]. These bounds are crucial as they quantify growth of a loss function outside the set of target solutions, essentially acting as a "signal" to guide algorithms toward target solutions in our learning problems. Concretely, under distributional assumptions on $\mathpzc{p}^*$ from [Wan+23a], the following sharpness property can be established: there is an absolute constant $c_1 > 0$ such that $\forall \boldsymbol{w} \in \mathscr{B}(2\|\boldsymbol{w}^*\|_2)$,

$$\|\boldsymbol{w} - \boldsymbol{w}^*\|_2^2 = \Omega(\mathrm{OPT}) \;\Rightarrow\; \Lambda_{\sigma,\mathpzc{p}^*}(\boldsymbol{w}) - \Lambda_{\sigma,\mathpzc{p}^*}(\boldsymbol{w}^*) \geq c_1 \|\boldsymbol{w} - \boldsymbol{w}^*\|_2^2. \tag{2}$$

The local error bounds in [MBM18; Wan+23a] assume identical reference and target distributions. Introducing distributional ambiguity — as in our work — invalidates this assumption, and as a result necessary distributional assumptions for sharpness may not apply to all distributions in the ambiguity set. In this work, distributional assumptions are exclusively applied to the target distribution to exploit the sharpness property proved in [Wan+23a]. We also assume that the sample covariates from the reference distribution are polynomially bounded; this assumption, which is without loss of generality, impacts only the sample and computational complexities and is satisfied by standard distributions.

Primal-Dual Algorithm Our algorithm is a principled, primal-dual algorithm leveraging the sharpness property on the target distribution, the structure of the square loss, and properties of chi-squared divergence. We control a "gap-like" function of the iterates, $\mathrm{Gap}(\widehat{\boldsymbol{w}}, \widehat{\mathpzc{p}}; \widehat{\mathpzc{p}}_0) := L_\sigma(\widehat{\boldsymbol{w}}, \widehat{\mathpzc{p}}^*; \widehat{\mathpzc{p}}_0) - L_\sigma(\boldsymbol{w}^*, \widehat{\mathpzc{p}}; \widehat{\mathpzc{p}}_0)$. The idea of approximating a gap and showing it reduces at a rate $1/A_k$, where A_k is a monotonically increasing function of k, comes from [DO19] and has been extended to primal-dual methods, including DRO settings, in [SWD21; Dia+22c; Son+22; MDH24].

Unlike past work [SWD21; Dia+22c; Son+22; MDH24], our primal problem is nonconvex, even for ReLU activations without distributional ambiguity. Unfortunately, the previously mentioned results relying on convexity do not apply in our setting. Additionally, sharpness — which appears crucial to approximating the target loss — is a *local* property, applying only to $\boldsymbol{w}$ such that $\|\boldsymbol{w}\|_2 \leq 2\|\boldsymbol{w}^*\|_2$, where $\|\boldsymbol{w}^*\|_2$ is unknown. This condition is trivially met at initialization, but proving it holds for all iterates requires convergence. We address this issue via an inductive argument, effectively coupling convergence analysis with localization of the iterates.

Additionally, standard primal-dual methods [CP11; Cha+18; SWD21; ACW22; Son+22] rely on bilinear coupling between primal and dual variables in $L_\sigma(\boldsymbol{w}, \widehat{\mathpzc{p}}; \widehat{\mathpzc{p}}_0)$. In our case, $L_\sigma(\boldsymbol{w}, \widehat{\mathpzc{p}}; \widehat{\mathpzc{p}}_0)$ is *nonlinear* and *nonconvex* in the first argument. Recent work [MDH24] handled nonlinearity by linearizing the function using convexity of the loss, which makes the function bounded below by its linear approximation at any point. However, this approach cannot be applied to our problem as the loss is nonconvex. Instead, we control the chi-squared divergence between the target distribution and the algorithm dual iterates to bound $L_\sigma(\boldsymbol{w}, \widehat{\mathpzc{p}}^*; \widehat{\mathpzc{p}}_0)$ from below, using a key structural result that we establish in Lemma 3.4. The challenges involved in proving this structural result require us to rely on chi-squared regularization and convex activation σ. Generalizing our result to all monotone unbounded activations and other strongly convex divergences like KL would need a similar structural lemma under these broader assumptions.

An interesting aspect of our analysis is that we do not rely on a convex surrogate for our problem. Instead, we constructively bound a quantity related to the DRO risk of the original square loss, justifying our algorithmic choices directly from the analysis. Although we do not consider convex surrogates, the vector field $\boldsymbol{v}(\boldsymbol{w}; \boldsymbol{x}, y)$, scaled by 2β, corresponds to the gradient of the convex surrogate loss $\int_0^{\boldsymbol{w}\cdot\boldsymbol{x}} (\sigma(t) - y)\,\mathrm{d}t$, which has been used in prior literature on learning a single neuron under similar settings without distributional ambiguity [Kak+11; Dia+20; Wan+23a]. In our analysis, the vector field $\boldsymbol{v}(\boldsymbol{w}; \boldsymbol{x}, y)$ is naturally motivated by the argument in the proof of Lemma 3.4.

"Concentration" of the Target Distribution To prove that our primal-dual algorithm converges, we need to prove both an upper bound and a lower bound for $\mathrm{Gap}(\widehat{\boldsymbol{w}}, \widehat{\mathpzc{p}}; \widehat{\mathpzc{p}}_0)$. The lower bound relies on sharpness; however, we need it to hold for the *empirical target distribution* ($\widehat{\mathscr{p}}^*$). This requires us to translate distributional assumptions and/or their implications from $\mathscr{p}^*$ to $\widehat{\mathscr{p}}^*$. Unfortunately, $\widehat{\mathscr{p}}^*$ is not the uniform distribution over samples drawn from $\mathscr{p}^*$. Rather, it is the maximizing distribution in the empirical DRO risk, defined w.r.t. $\widehat{\mathscr{p}}_0$. This means that prior uniform convergence results do not apply. Additionally, minimax risk rates from prior statistical results, such as those in [DN21], relate $R(\boldsymbol{w}; \widehat{\mathscr{p}}_0)$ and $R(\boldsymbol{w}; \mathscr{p}_0)$. However, they do not help in our algorithmic analysis since they do not guarantee that the sharpness holds for $\widehat{\mathscr{p}}^*$.

To address these challenges, we prove (in Corollary C.2) that as long as ν is sufficiently large, there is a simple closed-form expression for $\widehat{\mathscr{p}}^*$ as a function of $\widehat{\mathscr{p}}_0$ and an analogous relationship holds between $\mathscr{p}^*$ and $\mathscr{p}_0$. This allows us to leverage the fact that expectations of bounded functions with respect to $\widehat{\mathscr{p}}_0$ closely approximate those with respect to $\mathscr{p}_0$ to show that expectations with respect to $\widehat{\mathscr{p}}^*$ and $\mathscr{p}^*$ are similarly close. This result then implies that the sharpness also holds for $\widehat{\mathscr{p}}^*$ (Lemma C.6). Full details are provided in Appendix C.

2 Preliminaries

In this section, we introduce the necessary notation and state basic facts used in our analysis.

Notation Given a positive integer N, $[N]$ denotes the set $\{1, 2, \ldots, N\}$. Given a set $\mathscr{E}$, $\mathscr{E}^c$ denotes the complement of $\mathscr{E}$ when the universe is clear from the context. We use $\mathbb{I}_{\mathscr{E}}$ to denote the characteristic function of a set $\mathscr{E}$: $\mathbb{I}_{\mathscr{E}}(x) = 1$ if $x \in \mathscr{E}$ and $\mathbb{I}_{\mathscr{E}}(x) = 0$ otherwise. For vectors $\boldsymbol{x}$ and $\widehat{\boldsymbol{x}}$ from the d-dimensional Euclidean space $\mathbb{R}^d$, we use $\langle \boldsymbol{x}, \widehat{\boldsymbol{x}} \rangle$ and $\boldsymbol{x} \cdot \widehat{\boldsymbol{x}}$ to denote the standard inner product, while $\| \cdot \|_2 = \sqrt{\langle \cdot, \cdot \rangle}$ denotes the ℓ_2 norm. We use $(\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}, \ldots, \boldsymbol{x}^{(d)})$ to denote the entries of $\boldsymbol{x} \in \mathbb{R}^d$. We write $\boldsymbol{x} \leq \widehat{\boldsymbol{x}}$ to indicate $\boldsymbol{x}^{(j)} \leq \widehat{\boldsymbol{x}}^{(j)}$ for all coordinates j. For $r > 0$, $\mathscr{B}(r) := \{\boldsymbol{x} : \|\boldsymbol{x}\|_2 \leq r\}$ denotes the centered ball of radius r. We use Δ_N to denote the probability simplex: $\Delta_n := \{\boldsymbol{x} \in \mathbb{R}^N : \sum_{j=1}^N \boldsymbol{x}^{(j)} = 1, \forall j \in [N] : \boldsymbol{x}^{(j)} \geq 0\}$. We denote by $\boldsymbol{I}_d$ the identity matrix of size $d \times d$. We write $A \succeq B$ to indicate that $\boldsymbol{x}^\top (A - B) \boldsymbol{x} \geq 0$ for all $\boldsymbol{x} \in \mathbb{R}^d$. For two functions f and g, we say $f = \tilde{O}(g)$ if $f = O(g \log^k(g))$ for some constant k, and similarly define $\tilde{\Omega}$. We use notation $\tilde{O}_c(\cdot)$ and $\tilde{\Omega}_c(\cdot)$ to hide polynomial factors in (typically absolute constant) parameters c. For two distributions $\mathscr{p}$ and $\mathscr{p}'$, we use $\mathscr{p} \ll \mathscr{p}'$ to denote that $\mathscr{p}$ is absolutely continuous with respect to $\mathscr{p}'$, i.e., for all measurable sets A, $\mathscr{p}'(A) = 0$ implies $\mathscr{p}(A) = 0$. Typically, $\hat{p}$ and $\hat{q}$ are empirical distributions, and $\hat{p} \ll \hat{q}$ is equivalent to the condition that the support of $\hat{p}$ is a subset of the support of $\hat{q}$. For $\mathscr{p} \ll \mathscr{p}'$, we use $\frac{\mathrm{d}\mathscr{p}}{\mathrm{d}\mathscr{p}'}$ to denote their Radon–Nikodym derivative, which is the quotient of probability mass functions for discrete distributions. We use $\chi^2(\mathscr{p}, \mathscr{p}')$ to denote the chi-squared divergence of $\mathscr{p}$ w.r.t. $\mathscr{p}'$, i.e., $\chi^2(\mathscr{p}, \mathscr{p}') = \int (\frac{\mathrm{d}\mathscr{p}}{\mathrm{d}\mathscr{p}'} - 1)^2 \mathrm{d}\mathscr{p}'$.

2.1 Distributional Assumptions

Similar to [Wan+23a], we make two assumptions about the target distribution of the covariates ($\mathscr{p}^*_{\boldsymbol{x}}$). First, we assume that the optimal solution $\boldsymbol{w}^*$ satisfies the following "margin-like" condition:

Assumption 2.1 (Margin)**.** *There exist absolute constants* $\lambda, \gamma \in (0, 1]$ *such that* $\mathbb{E}_{\boldsymbol{x} \sim \mathscr{p}^*_{\boldsymbol{x}}}[\boldsymbol{x}\boldsymbol{x}^T \mathbb{I}_{\boldsymbol{w}^* \cdot \boldsymbol{x} \geq \gamma \|\boldsymbol{w}^*\|_2}] \succeq \lambda \boldsymbol{I}$*, where* $\mathscr{p}^*_{\boldsymbol{x}}$ *is the* $\boldsymbol{x}$*-marginal distribution of* $\mathscr{p}^*$*.*

We also assume that $\mathscr{p}^*_{\boldsymbol{x}}$ is subexponential with parameter B, which is an absolute constant.

Assumption 2.2 (Subexponential Concentration)**.** *There exists a parameter* $B > 0$ *such that for any* $\boldsymbol{u} \in \mathscr{B}(1)$ *and any* $r \geq 1$*, it holds that* $\Pr_{\boldsymbol{x} \sim \mathscr{p}^*_{\boldsymbol{x}}}[|\boldsymbol{u} \cdot \boldsymbol{x}| \geq r] \leq \exp(-Br)$*.*

Appendix E of [Wan+23a] shows that Assumptions 2.1 and 2.2 are satisfied by several important families of distributions including Gaussians, discrete Gaussians, all isotropic log-concave distributions, the uniform distribution over $\{-1, 0, 1\}^d$, etc.

For simplicity, we assume the labeled samples $(\boldsymbol{x}^{(i)}, y^{(i)})$ drawn from the reference distribution are bounded. This assumption, which does not affect the approximation constant for Problem 1.3, only impacts iteration and sample complexities. We state the bound on the covariates below, while a bound on the labels follows from prior work (Fact 2.6 stated in the next subsection).

Assumption 2.3 (Boundedness)**.** *There exists a parameter* S *such that for any fixed* $\boldsymbol{u} \in \mathscr{B}(1)$ *it holds that* $\boldsymbol{u} \cdot \boldsymbol{x} \leq S$ *for all sample covariates* $\boldsymbol{x}$ *in the support of* $\widehat{\mathscr{p}}_0$*.*

We also assume without loss of generality that $\|\boldsymbol{w}^*\|_2^2 \geq C\,\mathrm{OPT} + \epsilon$ for some absolute constant C, since otherwise **0** would be a valid $O(\mathrm{OPT}) + \epsilon$ solution. Algorithmically, we can first compute the empirical risk (per Corollary C.3) of the output from our algorithm and of $\hat{\boldsymbol{w}} = \mathbf{0}$ and then output the solution with the lower risk to get an $O(\mathrm{OPT}) + \epsilon$ solution; see Claim E.2 for a detailed discussion.

2.2 Auxiliary Facts

To achieve the claimed guarantees, we leverage structural properties of the loss function on the target distribution, implied by our distributional assumptions (Assumptions 2.1 and 2.2). Specifically, we make use of Lemma 2.2 and Fact C.4 from [Wan+23a], summarized in the fact below.

Fact 2.4 (Sharpness ([Wan+23a]))**.** *Suppose p^* and $\boldsymbol{w}^*$ satisfy Assumptions 2.1 and 2.2. Let $c_0 = \frac{\gamma\lambda\alpha}{6B\log(20B/\lambda^2)}$. For all $\boldsymbol{w} \in \mathscr{B}(2\|\boldsymbol{w}^*\|)$ and $\boldsymbol{u} \in \mathscr{B}(1)$,*

$$\mathbb{E}_{\boldsymbol{x}\sim p^*_{\boldsymbol{x}}}[(\sigma(\boldsymbol{w}\cdot\boldsymbol{x}) - \sigma(\boldsymbol{w}^*\cdot\boldsymbol{x}))(\boldsymbol{w}\cdot\boldsymbol{x} - \boldsymbol{w}^*\cdot\boldsymbol{x})] \geq c_0\|\boldsymbol{w} - \boldsymbol{w}^*\|_2^2,$$

$$\mathbb{E}_{\boldsymbol{x}\sim p^*_{\boldsymbol{x}}}[(\boldsymbol{x}\cdot\boldsymbol{u})^\tau] \leq 5B \quad \textit{for } \tau = 2, 4.$$

Fact 2.4 applies to the population version of the problem. Such a result also holds for the target distribution of the empirical problem, which we state below. Note that this result cannot be obtained by appealing to uniform convergence results for learning a neuron (without distributional robustness).

Lemma 2.5 (Empirical Sharpness; Informal. See Lemma C.6)**.** *Under Assumptions 2.1 to 2.3, for a sufficiently large sample size N as a function of $B, W, S, \nu, \alpha, \gamma, \lambda, d$ and with high probability, for all $\boldsymbol{w} \in \mathscr{B}(2\|\boldsymbol{w}^*\|)$ with $\|\boldsymbol{w} - \boldsymbol{w}^*\| \geq \sqrt{\epsilon}$ and $\boldsymbol{u} \in \mathscr{B}(1)$,*

$$\mathbb{E}_{\boldsymbol{x}\sim \widehat{p}^*_{\boldsymbol{x}}}[(\sigma(\boldsymbol{w}\cdot\boldsymbol{x}) - \sigma(\boldsymbol{w}^*\cdot\boldsymbol{x}))(\boldsymbol{w}\cdot\boldsymbol{x} - \boldsymbol{w}^*\cdot\boldsymbol{x})] \geq (c_0/2)\|\boldsymbol{w} - \boldsymbol{w}^*\|_2^2 \tag{3}$$

$$\mathbb{E}_{\boldsymbol{x}\sim \widehat{p}^*_{\boldsymbol{x}}}[(\boldsymbol{x}\cdot\boldsymbol{u})^\tau] \leq 6B \quad \textit{for } \tau = 2, 4. \tag{4}$$

As a consequence, for $c_1 = c_0^2/(24B)$ and any $\boldsymbol{w} \in \mathscr{B}(W)$ (where c_0 is defined in Fact 2.4), we have

$$c_1\|\boldsymbol{w} - \boldsymbol{w}^*\|_2^2 \leq \mathbb{E}_{\boldsymbol{x}\sim \widehat{p}^*_{\boldsymbol{x}}}[(\sigma(\boldsymbol{w}\cdot\boldsymbol{x}) - \sigma(\boldsymbol{w}^*\cdot\boldsymbol{x}))^2] \leq 6B\beta^2\|\boldsymbol{w} - \boldsymbol{w}^*\|_2^2, \tag{5}$$

where the left inequality uses Cauchy-Schwarz and the right inequality uses β-Lipschitzness of $\sigma(\cdot)$.

[Wan+23a] also showed that the labels y can be assumed to be bounded without loss of generality.

Fact 2.6. *Suppose p^* and $\boldsymbol{w}^*$ satisfy Assumption 2.1 and Assumption 2.2. Let $y' = \operatorname{sign}(y)\max\{|y|, M\}$ where for some sufficiently large absolute constant C_M we define*

$$M = C_M W B \beta \log(\beta B W/\epsilon) \tag{6}$$

Then $\mathbb{E}_{p^*}(\sigma(\boldsymbol{w}^*\cdot\boldsymbol{x}) - y')^2 \leq \mathbb{E}_{p^*}(\sigma(\boldsymbol{w}^*\cdot\boldsymbol{x}) - y)^2 + \epsilon = \mathrm{OPT} + \epsilon$.

We also make use of the following facts from convex analysis. First, let $\phi : \mathbb{R}^N \to \mathbb{R}$ be a differentiable function and the Bregman divergence of ϕ for any $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^N$ be defined by

$$D_\phi(\boldsymbol{y}, \boldsymbol{x}) = \phi(\boldsymbol{y}) - \phi(\boldsymbol{x}) - \langle \nabla\phi(\boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x}\rangle.$$

Fact 2.7. *Let $\psi(\boldsymbol{x}) = \phi(\boldsymbol{x}) + \langle \boldsymbol{a}, \boldsymbol{x}\rangle + b$ for some $\boldsymbol{a} \in \mathbb{R}^N$ and $b \in \mathbb{R}$. Then $D_\psi(\boldsymbol{y}, \boldsymbol{x}) = D_\phi(\boldsymbol{y}, \boldsymbol{x})$ for all $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^N$, i.e., the Bregman divergence is blind to the addition of affine terms to function ϕ.*

Second, we state the first-order necessary conditions that a local maximizer must satisfy.

Fact 2.8 (First-Order Optimality Condition)**.** *Let Ω be a closed, convex, and nonempty set and let $f : \Omega \to \mathbb{R}$ be continuously differentiable. If $\boldsymbol{x}^*$ is a local maximizer of f on Ω, then it holds that*

$$\nabla f(\boldsymbol{x}^*) \cdot (\boldsymbol{y} - \boldsymbol{x}^*) \leq 0 \quad \textit{for all } \boldsymbol{y} \in \Omega. \tag{7}$$

If f is also concave, then Equation (7) *implies that $\boldsymbol{x}^*$ is a global maximizer of f.*

3 Algorithm and Convergence Analysis

In this section, we introduce our algorithm and state our main results, summarized in Theorem 3.1. We highlight the main components of our technical approach, while most of the technical details are deferred to the appendix, due to space constraints.

To facilitate the presentation of results, we introduce the following auxiliary notation: $\ell(\boldsymbol{w};\boldsymbol{x},y) := (\sigma(\boldsymbol{w}\cdot\boldsymbol{x})-y)^2$, $\boldsymbol{v}(\boldsymbol{w};\boldsymbol{x},y) := 2\beta(\sigma(\boldsymbol{w}\cdot\boldsymbol{x})-y)\boldsymbol{x}$ and $\widehat{\mathrm{OPT}} = \mathbb{E}_{(\boldsymbol{x},y)\sim\widehat{p}^*}\ell(\boldsymbol{w}^*;\boldsymbol{x},y)$. We also note that Assumption 2.3 implies that for all samples $\{\boldsymbol{x}_i, y_i\}$, the function $\boldsymbol{w}\mapsto\boldsymbol{v}(\boldsymbol{w};\boldsymbol{x}_i,y_i)$ is bounded above by G and κ-Lipschitz for all $i\in[N]$ and $\boldsymbol{w}\in\mathscr{B}(W)$, where $G = 2\beta S\sqrt{d}(\sqrt{2}\beta WS+M)$ and $\kappa = 2\beta^2S^2d$ (see Lemma B.4 in Appendix B). Starting from this section, we write $L(\boldsymbol{w},\widehat{p})$ to denote $L_\sigma(\boldsymbol{w},\widehat{p};\widehat{p}_0)$, hiding the dependence on $\widehat{p}_0$ and σ. We also write $\mathrm{Gap}(\boldsymbol{w},\widehat{p})$ for $\mathrm{Gap}(\boldsymbol{w},\widehat{p};\widehat{p}_0)$

Our main algorithm (Algorithm 1) is an iterative primal-dual method with extrapolation on the primal side via $\boldsymbol{g}_i$. The vector $\mathbb{E}_{\widehat{p}_i}[\boldsymbol{v}(\boldsymbol{w}_i;\boldsymbol{x},y)]$ equals the (scaled) gradient of a surrogate loss used in prior works [Kak+11; Dia+20; Wan+23a]. In contrast to prior work, we directly bound the original square loss, with $\mathbb{E}_{\widehat{p}_i}[\boldsymbol{v}(\boldsymbol{w}_i;\boldsymbol{x},y)]$ naturally arising from our analysis. Both updates $\boldsymbol{w}_i$ and $\widehat{p}_i$ are efficiently computable: $\boldsymbol{w}_i$ involves a simple projection onto a Euclidean ball, and $\widehat{p}_i$ involves a projection onto a probability simplex, computable in near-linear time [Duc+08].

Algorithm 1: Main algorithm

```
Input: $\nu > 0, \kappa, G, c_1, \nu_0 = 768\beta^4 B\epsilon/c_1$, sample set $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^N$
Initialization: $A_{-1} = a_{-1} = A_0 = a_0 = 0, \boldsymbol{w}_{-1} = \boldsymbol{w}_0 = \boldsymbol{0}, \widehat{p}_{-1} = \widehat{p}_0$;
for $i = 1, \ldots, k$ do
    $a_i = \left(1 + \frac{\min\{\nu, c_1/8\}}{2\max\{\kappa, G\}}\right)^{i-1} \min\{\nu_0, 1/4\}/(2\max\{\kappa, G\}), A_i = a_i + A_{i-1}$;
    $\boldsymbol{v}(\boldsymbol{w};\boldsymbol{x},y) = 2\beta(\sigma(\boldsymbol{w}\cdot\boldsymbol{x}) - \mathrm{sign}(y)\max\{|y|, M\})\boldsymbol{x}$, where $M$ is defined in Equation (6) ;
    $\boldsymbol{g}_{i-1} = \mathbb{E}_{\widehat{p}_{i-1}}[\boldsymbol{v}(\boldsymbol{w}_{i-1};\boldsymbol{x},y)] + \frac{a_{i-1}}{a_i}(\mathbb{E}_{\widehat{p}_{i-1}}[\boldsymbol{v}(\boldsymbol{w}_{i-1};\boldsymbol{x},y)] - \mathbb{E}_{\widehat{p}_{i-2}}[\boldsymbol{v}(\boldsymbol{w}_{i-2};\boldsymbol{x},y)])$;
    $\boldsymbol{w}_i = \arg\min_{\boldsymbol{w}\in\mathscr{B}(W)} \left\{a_i\langle\boldsymbol{g}_{i-1}, \boldsymbol{w}\rangle + \frac{1+0.5c_1A_{i-1}}{2}\|\boldsymbol{w} - \boldsymbol{w}_{i-1}\|_2^2\right\}$;
    $\widehat{p}_i = \arg\max_{\widehat{p}\in\mathscr{P}} \left\{a_iL(\boldsymbol{w}_i, \widehat{p}) - (\nu_0 + \nu A_{i-1})D_{\chi^2(\cdot,\widehat{p}_0)}(\widehat{p}, \widehat{p}_{i-1})\right\}$;
```

Theorem 3.1 (Main Theorem)**.** *Under Assumptions 2.1 to 2.3, suppose the sample size is such that* $N = \widetilde{\Omega}_{B,S,\beta,\alpha,\gamma,\lambda}\left(\frac{W^4}{\epsilon^2}\left(1+\frac{W^4}{\nu^2}\right)(d + W^4\log(1/\delta))\right)$ *and* $\nu \geq 8\beta^2\sqrt{6B}\sqrt{\mathrm{OPT}_{(2)}+\epsilon}/c_1$*, where* $\mathrm{OPT}_{(2)} = \mathbb{E}_{p^*}[\ell(\boldsymbol{w}^*;\boldsymbol{x},y)^2]$ *and* c_1 *is defined in Lemma 2.5. With probability at least* $1-\delta$*, for all iterates* $\boldsymbol{w}_k, \widehat{p}_k$*, it holds that*

$$\frac{c_1}{4}\|\boldsymbol{w}^* - \boldsymbol{w}_k\|_2^2 + \nu D_\phi(\widehat{p}^*, \widehat{p}_k) \leq \frac{D_0}{A_k} + \frac{60\beta^2B\,\mathrm{OPT}}{c_1} + \epsilon,$$

where $D_0 = \frac{1}{2}\|\boldsymbol{w}^* - \boldsymbol{w}_0\|_2^2 + \nu_0\chi^2(\widehat{p}^*, \widehat{p}_0)$ *and* $\chi^2(\widehat{p}^*, \widehat{p}_0) \leq c_1/(1536\beta^4B)$ *(and therefore* D_0 *does not depend on the sample size N).*

In particular, after at most $k = \widetilde{O}(\frac{\max\{\kappa,G\}}{\min\{\nu,c_1\}}\log(\frac{D_0}{\epsilon}))$ *iterations, it holds that*

$$\|\boldsymbol{w}_k - \boldsymbol{w}^*\|_2 \leq C_3\sqrt{\mathrm{OPT}} + \sqrt{\epsilon}, \tag{8}$$

$$\mathbb{E}_{(\boldsymbol{x},y)\sim p^*}[\ell(\boldsymbol{w}_k;\boldsymbol{x},y)] \leq (2 + 20B\beta^2C_3^2)\ \mathrm{OPT} + 10\beta^2B\epsilon, \tag{9}$$

$$R(\boldsymbol{w}_k;p_0) - \min_{\boldsymbol{w}\in\mathscr{B}(W)} R(\boldsymbol{w};p_0) = R(\boldsymbol{w}_k;p_0) - R(\boldsymbol{w}^*;p_0) \leq C_4(\mathrm{OPT}+\epsilon), \tag{10}$$

where $C_3 = 16\beta\sqrt{B}/c_1$ *and* $C_4 = 1 + 2(10B\beta^2 + c_1)C_3 + c_1\sqrt{5B}\beta^2C_3^2$*.*

We focus on the convergence of iterates $\boldsymbol{w}_i$ as claimed in Equation (8); the loss bound (Equation (9)) follows directly from the iterate convergence, while the risk bound (Equation (10)) requires a more involved analysis. Complete details for Equations (9) and (10) are provided in Appendix F.

Our strategy for the convergence analysis is as follows. Consider $\{a_i\}$, a sequence of positive step sizes, and define A_i as their cumulative sum $\sum_{j=1}^i a_j$. Our algorithm produces a sequence of primal-dual pairs $\boldsymbol{w}_i, \widehat{p}_i$, tracking a quantity related to the primal-dual gap, defined by:

$$\mathrm{Gap}(\boldsymbol{w}_i, \widehat{p}_i) := L(\boldsymbol{w}_i, \widehat{p}^*) - L(\boldsymbol{w}^*, \widehat{p}_i) = (L(\boldsymbol{w}_i, \widehat{p}^*) - L(\boldsymbol{w}^*, \widehat{p}^*)) + (L(\boldsymbol{w}^*, \widehat{p}^*) - L(\boldsymbol{w}^*, \widehat{p}_i)).$$

We view $(L(\boldsymbol{w}_i, \widehat{p}^*) - L(\boldsymbol{w}^*, \widehat{p}^*))$ as the "primal gap" and $(L(\boldsymbol{w}^*, \widehat{p}^*) - L(\boldsymbol{w}^*, \widehat{p}_i))$ as the "dual gap." Since the squared loss for ReLU and similar activations is nonconvex, $L(\boldsymbol{w}, \widehat{p}^*)$ is nonconvex in its first argument. Note that this gap function is not trivially non-negative (see Remark B.7), requiring an explicit lower bound proof.

Our strategy consists of deriving "sandwiching" inequalities for the (weighted) cumulative gap $\sum_{i=1}^{k} a_i \mathrm{Gap}(\boldsymbol{w}_i, \widehat{p}_i)$ and deducing convergence guarantees for the algorithm iterates from them. A combination of these two inequalities leads to the statement of Theorem 3.1, from which we can deduce that unless we already have an $O(\mathrm{OPT}) + \epsilon$ solution, the iterates must be converging to the target solutions at rate $1/A_k$, which we argue can be made geometrically fast.

Organization The rest of this section is organized as follows — under the standard assumptions we state in this paper, in Lemma 3.2, we prove a lower bound on $\mathrm{Gap}(\boldsymbol{w}, \hat{p})$ for any choice of $\boldsymbol{w}$ and $\hat{p}$. This can be used to get a corresponding lower bound on the weighted sum $\sum_{i=1}^{k} a_i \mathrm{Gap}(\boldsymbol{w}_i, \hat{\boldsymbol{p}}_i)$. In Lemma 3.3 we then state an upper bound on $\sum_{i=1}^{k} a_i \mathrm{Gap}(\boldsymbol{w}_i, \hat{\boldsymbol{p}}_i)$; the proof of this technical argument is deferred to Appendix D. These two bounds together give us the first inequality in Theorem 3.1. Claim B.6 then bounds below the convergence rate for our choice of a_i in Algorithm 1; and indicates that it is geometric. Finally, we put everything together to prove Theorem 3.1.

To simplify the notation, we use $\phi(\widehat{p}) := \chi^2(\widehat{p}, \widehat{p}_0)$ throughout this section. Note that $D_\phi(\widehat{p}, \widehat{q}) = D_\phi(\widehat{q}, \widehat{p}) = \sum_{i=1}^{N} \frac{(\widehat{q}^{(i)} - \widehat{p}^{(i)})^2}{\widehat{p}_0^{(i)}}$ for any $\widehat{p}$ and $\widehat{q}$ in the domain.

3.1 Lower Bound on the Gap Function

We begin the convergence analysis by demonstrating a lower bound on $\mathrm{Gap}(\boldsymbol{w}_i, \widehat{p}_i)$.

Lemma 3.2 (Gap Lower Bound)**.** *Under the setting in which Lemma 2.5 holds, for all* $\boldsymbol{w} \in \mathcal{B}(2\|\boldsymbol{w}^*\|_2)$, $\mathrm{Gap}(\boldsymbol{w}, \widehat{p}) \geq -\frac{12\beta^2 B}{c_1}\widehat{\mathrm{OPT}} + \frac{c_1}{2}\|\boldsymbol{w} - \boldsymbol{w}^*\|_2^2 + \nu D_\phi(\widehat{p}^*, \widehat{p})$.

Proof. Writing $(\sigma(\boldsymbol{w} \cdot \boldsymbol{x}) - y)^2 = \big((\sigma(\boldsymbol{w} \cdot \boldsymbol{x}) - \sigma(\boldsymbol{w}^* \cdot \boldsymbol{x})) + (\sigma(\boldsymbol{w}^* \cdot \boldsymbol{x}) - y)\big)^2$ and expanding the square, we have

$$\begin{aligned} L(\boldsymbol{w}, \widehat{p}^*) - L(\boldsymbol{w}^*, \widehat{p}^*) &= \mathbb{E}_{(\boldsymbol{x}, y) \sim \widehat{p}^*}[(\sigma(\boldsymbol{w} \cdot \boldsymbol{x}) - y)^2 - (\sigma(\boldsymbol{w}^* \cdot \boldsymbol{x}) - y)^2] \\ = &- 2\mathbb{E}_{\widehat{p}^*}[(\sigma(\boldsymbol{w}^* \cdot \boldsymbol{x}) - y)(\sigma(\boldsymbol{w} \cdot \boldsymbol{x}) - \sigma(\boldsymbol{w}^* \cdot \boldsymbol{x}))] + \mathbb{E}_{\widehat{p}^*}[((\sigma(\boldsymbol{w} \cdot \boldsymbol{x}) - \sigma(\boldsymbol{w}^* \cdot \boldsymbol{x}))^2]. \end{aligned}$$

By the Cauchy-Schwarz inequality, we further have that

$$\begin{aligned} &\mathbb{E}_{\widehat{p}^*}[(\sigma(\boldsymbol{w}^* \cdot \boldsymbol{x}) - y)(\sigma(\boldsymbol{w} \cdot \boldsymbol{x}) - \sigma(\boldsymbol{w}^* \cdot \boldsymbol{x}))] \\ &\leq \sqrt{\mathbb{E}_{\widehat{p}^*}[(\sigma(\boldsymbol{w}^* \cdot \boldsymbol{x}) - y)^2]\mathbb{E}_{\widehat{p}^*}[(\sigma(\boldsymbol{w} \cdot \boldsymbol{x}) - \sigma(\boldsymbol{w}^* \cdot \boldsymbol{x}))^2]} \\ &\leq \beta\sqrt{6B}\sqrt{\widehat{\mathrm{OPT}}}\|\boldsymbol{w} - \boldsymbol{w}^*\|_2, \end{aligned} \tag{11}$$

where in the second inequality we used the definition of $\widehat{\mathrm{OPT}}$ and $\mathbb{E}_{\boldsymbol{x} \sim \widehat{p}_{\boldsymbol{x}}^*}[(\sigma(\boldsymbol{w} \cdot \boldsymbol{x}) - \sigma(\boldsymbol{w}^* \cdot \boldsymbol{x}))^2] \leq 6B\beta^2\|\boldsymbol{w} - \boldsymbol{w}^*\|_2^2$ from the right inequality in Equation (5).

On the other hand, by the left inequality in Equation (5), we also have

$$\mathbb{E}_{\widehat{p}^*}[(\sigma(\boldsymbol{w} \cdot \boldsymbol{x}) - \sigma(\boldsymbol{w}^* \cdot \boldsymbol{x}))^2] \geq c_1\|\boldsymbol{w} - \boldsymbol{w}^*\|_2^2. \tag{12}$$

Thus, combining Equation (11) and Equation (12), we get

$$\begin{aligned} L(\boldsymbol{w}, \widehat{p}^*) - L(\boldsymbol{w}^*, \widehat{p}^*) &\geq -2\beta\sqrt{6B}\|\boldsymbol{w} - \boldsymbol{w}^*\|_2\sqrt{\widehat{\mathrm{OPT}}} + c_1\|\boldsymbol{w} - \boldsymbol{w}^*\|_2^2 \\ &\geq -\frac{12\beta^2 B}{c_1}\widehat{\mathrm{OPT}} + \frac{c_1}{2}\|\boldsymbol{w} - \boldsymbol{w}^*\|_2^2, \end{aligned} \tag{13}$$

where the last inequality is by $2\beta\sqrt{6B}\|\boldsymbol{w} - \boldsymbol{w}^*\|_2\sqrt{\widehat{\mathrm{OPT}}} \leq \frac{4\beta^2 6B}{2c_1}\widehat{\mathrm{OPT}} + \frac{c_1}{2}\|\boldsymbol{w} - \boldsymbol{w}^*\|_2^2$, which comes from an application of Young's inequality (Fact B.1).

Finally, we use the optimality of $\widehat{p}^*$, which achieves the maximum over all $\widehat{p} \in \mathcal{P}$ for $L(\boldsymbol{w}^*, \widehat{p})$. By the definition of a Bregman divergence, Fact 2.7, and first-order necessary condition in Fact 2.8:

$$-L(\boldsymbol{w}^*, \widehat{p}) - (-L(\boldsymbol{w}^*, \widehat{p}^*)) = -\langle \nabla_{\widehat{p}} L(\boldsymbol{w}^*, \widehat{p}^*), \widehat{p} - \widehat{p}^* \rangle + D_{-L(\boldsymbol{w}^*, \cdot)}(\widehat{p}, \widehat{p}^*) \geq \nu D_\phi(\widehat{p}^*, \widehat{p}). \tag{14}$$

Summing up Equation (13) and Equation (14) completes the proof. □

3.2 Upper Bound on the Gap Function

Having obtained a lower bound on the gap function, we now show an upper bound, leveraging our algorithmic choices. The proof is rather technical and involves individually bounding $L(\boldsymbol{w}_i, \widehat{p}^*)$ above and bounding $L(\boldsymbol{w}^*, \widehat{p}_i)$ below to obtain an upper bound on the gap function, which equals $L(\boldsymbol{w}_i, \widehat{p}^*) - L(\boldsymbol{w}^*, \widehat{p}_i)$. We state this result in the next lemma, while the proof is in Appendix D.

Lemma 3.3 (Gap Upper Bound)**.** *Let $\boldsymbol{w}_i, \widehat{p}_i, a_i, A_i$ evolve according to Algorithm 1, where we take, by convention, $a_{-1} = A_{-1} = a_0 = A_0 = 0$ and $\boldsymbol{w}_{-1} = \boldsymbol{w}_0$, $\widehat{p}_{-1} = \widehat{p}_0$. Assuming Lemma 2.5 applies, then, for all $k \geq 1$, $\sum_{i=1}^k a_i \mathrm{Gap}(\boldsymbol{w}_i, \widehat{p}_i)$ is bounded above by*

$$\frac{1}{2}\|\boldsymbol{w}^* - \boldsymbol{w}_0\|_2^2 + \nu_0 D_\phi(\widehat{p}^*, \widehat{p}_0) - \frac{1 + 0.5c_1 A_k}{2}\|\boldsymbol{w}^* - \boldsymbol{w}_k\|_2^2 - (\nu_0 + \nu A_k) D_\phi(\widehat{p}^*, \widehat{p}_k)$$
$$+ \sum_{i=1}^k a_i \frac{c_1}{4}\|\boldsymbol{w}^* - \boldsymbol{w}_i\|_2^2 + \frac{8\beta^2\sqrt{6B}\sqrt{\widehat{\mathrm{OPT}}_{(2)}}}{c_1}\sum_{i=1}^k a_i \chi^2(\widehat{p}_i, \widehat{p}^*) + \frac{48\beta^2 B\widehat{\mathrm{OPT}}A_k}{c_1}.$$

A critical technical component in the proof of Lemma 3.3 is how we handle issues related to nonconvexity. A key technical result that we prove and use is the following.

Lemma 3.4. *Let $S_i := \mathbb{E}_{\widehat{p}_i}[(\sigma(\boldsymbol{w}^* \cdot \boldsymbol{x}) - \sigma(\boldsymbol{w}_i \cdot \boldsymbol{x}))^2] + \mathbb{E}_{\widehat{p}_i}[2(\sigma(\boldsymbol{w}_i \cdot \boldsymbol{x}) - y)(\sigma(\boldsymbol{w}^* \cdot \boldsymbol{x}) - \sigma(\boldsymbol{w}_i \cdot \boldsymbol{x}))]$, $\boldsymbol{w}_i$ evolve according to Line 6 in Algorithm 1 and suppose we are in the setting where Lemma 2.5 holds. Then, $S_i \geq \mathbb{E}_{\widehat{p}_i}[\langle \boldsymbol{v}(\boldsymbol{w}; \boldsymbol{x}, y), \boldsymbol{w}^* - \boldsymbol{w}_i \rangle] - E_i$ where*

$$E_i = \frac{c_1}{4}\|\boldsymbol{w}^* - \boldsymbol{w}_i\|_2^2 + \Big(8\beta^2\sqrt{6B}\sqrt{\widehat{\mathrm{OPT}}_{(2)}}/c_1\Big)\chi^2(\widehat{p}_i, \widehat{p}^*) + (48\beta^2 B/c_1)\widehat{\mathrm{OPT}}. \tag{15}$$

This bound is precisely what forces us to choose chi-squared as the measure of divergence between distributions and introduce a dependence on $\widehat{\mathrm{OPT}}_{(2)}$. One pathway to generalize our results to other divergences would be to find a corresponding generalization to Lemma 3.4.

3.3 Proof of Main Theorem

Combining Lemma 3.2 and Lemma 3.3, we are now ready to prove our main result.

Proof of Theorem 3.1. Combining the lower bound on the gap function from Lemma 3.2 with the upper bound from Lemma 3.3 and rearranging, whenever $\|\boldsymbol{w}_i\|_2 \leq 2\|\boldsymbol{w}^*\|_2$ for all $i \leq k$ so that Lemma 2.5 applies, we get that

$$-\frac{12\beta^2 B}{c_1}\widehat{\mathrm{OPT}}A_k + \sum_{i=1}^k a_i \frac{c_1}{2}\|\boldsymbol{w}_i - \boldsymbol{w}^*\|_2^2 + \sum_{i=1}^k \nu a_i D_\phi(\widehat{p}^*, \widehat{p}_i) \leq \sum_{i=1}^k a_i \mathrm{Gap}(\boldsymbol{w}_i, \widehat{p}_i)$$
$$\leq \frac{1}{2}\|\boldsymbol{w}^* - \boldsymbol{w}_0\|_2^2 + \nu_0 D_\phi(\widehat{p}^*, \widehat{p}_0) - \frac{1 + 0.5c_1 A_k}{2}\|\boldsymbol{w}^* - \boldsymbol{w}_k\|_2^2 - (\nu_0 + \nu A_k) D_\phi(\widehat{p}^*, \widehat{p}_k)$$
$$+ \sum_{i=1}^k a_i \frac{c_1}{4}\|\boldsymbol{w}^* - \boldsymbol{w}_i\|_2^2 + \frac{8\beta^2\sqrt{6B}\sqrt{\widehat{\mathrm{OPT}}_{(2)}}}{c_1}\sum_{i=1}^k a_i \chi^2(\widehat{p}_i, \widehat{p}^*) + \frac{48\beta^2 B\widehat{\mathrm{OPT}}A_k}{c_1}.$$

To reach the first claim of the theorem, we first argue that $\sum_{i=1}^k a_i \Big((4\beta^2\sqrt{6B}\sqrt{\widehat{\mathrm{OPT}}_{(2)}}/c_1)\chi^2(\widehat{p}_i, \widehat{p}^*) - \nu D_\phi(\widehat{p}^*, \widehat{p}_i)\Big) \leq 0$. This follows from (1) Corollary C.2, by which we have $\widehat{p}^{*(j)} \geq \widehat{p}_0^{(j)}/2$ for all $j \in [N]$, hence

$$\chi^2(\widehat{p}_i, \widehat{p}^*) = \sum_{j \in [N]} (\widehat{p}^{*(j)} - \widehat{p}_i^{(j)})^2/\widehat{p}^{*(j)} \leq 2\sum_{j \in [N]} (\widehat{p}^{*(j)} - \widehat{p}_i^{(j)})^2/\widehat{p}_0^{(j)} = 2D_\phi(\widehat{p}^*, \widehat{p}_i)$$

and (2) our choice of ν, which ensures, with high probability, that $\nu \geq 8\beta^2\sqrt{6B}\sqrt{\mathrm{OPT}_{(2)} + \epsilon}/c_1 \geq 8\beta^2\sqrt{6B}\sqrt{\widehat{\mathrm{OPT}}_{(2)}}/c_1$, where the last inequality is because for the specified sample size, we have that $\widehat{\mathrm{OPT}}_{(2)} + \epsilon \geq \mathrm{OPT}_{(2)}$ by Corollary C.9.

Second, we similarly have that with probability $1-\delta$, $\widehat{\mathrm{OPT}} \le \mathrm{OPT}+\epsilon$. Hence, since Bregman divergence of a convex function is non-negative, whenever $\|\boldsymbol{w}_i\|_2 \le 2\|\boldsymbol{w}^*\|_2$ for all $i \le k$, we have

$$\|\boldsymbol{w}^* - \boldsymbol{w}_k\|_2^2 \le \frac{\|\boldsymbol{w}^* - \boldsymbol{w}_0\|_2^2 + 2\nu_0 D_\phi(\widehat{\boldsymbol{p}}^*, \widehat{\boldsymbol{p}}_0)}{1 + 0.5 c_1 A_k} + \frac{240\beta^2 B}{c_1}(\mathrm{OPT}+\epsilon) \tag{16}$$

$$D_\phi(\widehat{\boldsymbol{p}}^*, \widehat{\boldsymbol{p}}_k) \le \frac{\|\boldsymbol{w}^* - \boldsymbol{w}_0\|_2^2/2 + \nu_0 D_\phi(\widehat{\boldsymbol{p}}^*, \widehat{\boldsymbol{p}}_0)}{\nu_0 + \nu A_k} + \frac{60\beta^2 B}{\nu}(\mathrm{OPT}+\epsilon) \tag{17}$$

The bound $\chi^2(\widehat{\boldsymbol{p}}^*, \widehat{\boldsymbol{p}}_0) \le c_1/(1536\beta^4 B)$ is proved in Claim E.1. Finally, in Appendix E, we inductively prove the following claim so that assumptions in Lemma 2.5 are satisfied.

Claim 3.5. *For all iterations* $k \ge 0$, $\|\boldsymbol{w}_k\|_2 \le 2\|\boldsymbol{w}^*\|_2$.

The bound on the growth of A_k follows by standard arguments and is provided as Claim B.6. Since A_k grows exponentially with $(1+\eta)^k$ where $\eta = \frac{\min\{\nu, c_1/8\}}{2\max\{\kappa, G\}}$ and since $D_0(1+\eta)^{-k} \le \epsilon$ can be enforced by setting $k = (1+1/\eta)\log(D_0/\epsilon) \ge \log(D_0/\epsilon)/\log(1+\eta)$, we have that after $\widetilde{O}(\frac{\max\{\kappa, G\}}{\min\{\nu, c_1\}}\log(D_0/\epsilon))$ iterations either $\|\boldsymbol{w}_k - \boldsymbol{w}^*\|_2 \le \sqrt{\epsilon}$ or $\|\boldsymbol{w}_i - \boldsymbol{w}^*\|_2 \le C_3\sqrt{\mathrm{OPT}}$. □

4 Conclusion

In this paper, we study the problem of learning a single neuron in the distributionally robust setting, with the square loss regularized by the chi-squared distance between the reference and target distributions. Our results serve as a preliminary exploration in this area, paving the way for several potential extensions. Future work includes generalizing our approach to single index models with unknown activations, expanding to neural networks comprising multiple neurons, and considering alternative ambiguity sets such as those based on the Wasserstein distance or Kullback-Leibler divergence.

Acknowledgement

Shuyao Li was supported in part by AFOSR Award FA9550-21-1-0084 and the U.S. Office of Naval Research under award number N00014-22-1-2348. Sushrut Karmalkar was supported by NSF under Grant #2127309 to the Computing Research Association for the CIFellows 2021 Project. Ilias Diakonikolas was supported by NSF Medium Award CCF-2107079 and an H.I. Romnes Faculty Fellowship. Jelena Diakonikolas was supported in part by the U.S. Office of Naval Research under contract number N00014-22-1-2348. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the U.S. Department of Defense.

References

[ACW22] A. Alacaoglu, V. Cevher, and S. J. Wright. “On the complexity of a practical primal-dual coordinate method”. In: *arXiv preprint arXiv:2201.07684* (2022).

[AHW95] P. Auer, M. Herbster, and M. K. K. Warmuth. “Exponentially many local minima for single neurons”. In: *Advances in Neural Information Processing Systems*. 1995.

[ATV23] P. Awasthi, A. Tang, and A. Vijayaraghavan. “Agnostic Learning of General ReLU Activation Using Gradient Descent”. In: *The Eleventh International Conference on Learning Representations, ICLR*. 2023.

[Bec17] A. Beck. *First-order methods in optimization*. SIAM, 2017.

[Ben+10] S. Ben-David, J. Blitzer, K. Crammer, A. Kulesza, F. Pereira, and J. W. Vaughan. “A theory of learning from different domains”. In: *Machine learning* 79 (2010).

[BEN09] A. Ben-Tal, L. El Ghaoui, and A. Nemirovski. *Robust optimization*. Vol. 28. Princeton university press, 2009.

[BBS07] S. Bickel, M. Brückner, and T. Scheffer. “Discriminative learning for differing training and test distributions”. In: *Proceedings of the 24th international conference on Machine learning*. 2007.

[BMN21] J. Blanchet, K. Murthy, and V. A. Nguyen. “Statistical analysis of Wasserstein distributionally robust estimators”. In: *Tutorials in Operations Research: Emerging Optimization Methods and Modeling Techniques with Applications*. INFORMS, 2021.

[Bla+24] J. Blanchet, J. Li, S. Lin, and X. Zhang. “Distributionally Robust Optimization and Robust Statistics”. In: 2024.

[Cha+18] A. Chambolle, M. J. Ehrhardt, P. Richtárik, and C.-B. Schonlieb. “Stochastic primal-dual hybrid gradient algorithm with arbitrary sampling and imaging applications”. In: *SIAM Journal on Optimization* 28.4 (2018).

[CP11] A. Chambolle and T. Pock. “A first-order primal-dual algorithm for convex problems with applications to imaging”. In: *Journal of mathematical imaging and vision* 40 (2011).

[CP18] R. Chen and I. C. Paschalidis. “A robust learning approach for regression models based on distributionally robust optimization”. In: *Journal of Machine Learning Research* 19.13 (2018).

[CP20] R. Chen and I. C. Paschalidis. “Distributionally robust learning”. In: *Foundations and Trends® in Optimization* 4.1-2 (2020), pp. 1–243.

[Che+20] S. Chen, F. Koehler, A. Moitra, and M. Yau. “Classification under misspecification: Halfspaces, generalized linear models, and evolvability”. In: *Advances in Neural Information Processing Systems* 33 (2020).

[Dia+20] I. Diakonikolas, S. Goel, S. Karmalkar, A. R. Klivans, and M. Soltanolkotabi. “Approximation Schemes for ReLU regression”. In: *Conference on Learning Theory, COLT 2020*. 2020.

[DKZ20] I. Diakonikolas, D. Kane, and N. Zarifis. “Near-Optimal SQ Lower Bounds for Agnostically Learning Halfspaces and ReLUs under Gaussian Marginals”. In: *Annual Conference on Neural Information Processing Systems*. 2020.

[Dia+21] I. Diakonikolas, D. M. Kane, T. Pittas, and N. Zarifis. “The Optimality of Polynomial Regression for Agnostic Learning under Gaussian Marginals in the SQ Model”. In: *Proceedings of Machine Learning Research vol* 134 (2021).

[Dia+22a] I. Diakonikolas, V. Kontonis, C. Tzamos, and N. Zarifis. “Learning a Single Neuron with Adversarial Label Noise via Gradient Descent”. In: *Conference on Learning Theory (COLT)*. 2022, pp. 4313–4361.

[DPT21] I. Diakonikolas, J. H. Park, and C. Tzamos. “ReLU Regression with Massart Noise”. In: *Advances in Neural Information Processing Systems* 34 (2021).

[Dia+22b] I. Diakonikolas, D. Kane, P. Manurangsi, and L. Ren. “Hardness of Learning a Single Neuron with Adversarial Label Noise”. In: *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*. 2022.

[DKR23] I. Diakonikolas, D. Kane, and L. Ren. “Near-optimal cryptographic hardness of agnostically learning halfspaces and relu regression under gaussian marginals”. In: *International Conference on Machine Learning*. PMLR. 2023.

[Dia+22c] J. Diakonikolas, C. Li, S. Padmanabhan, and C. Song. “A fast scale-invariant algorithm for non-negative least squares with non-negative data”. In: *Advances in Neural Information Processing Systems* (2022).

[DO19] J. Diakonikolas and L. Orecchia. “The Approximate Duality Gap Technique: A Unified Theory of First-Order Methods”. In: *SIAM Journal on Optimization* 29.1 (2019), pp. 660–689.

[DN19] J. Duchi and H. Namkoong. “Variance-based regularization with convex objectives”. In: *Journal of Machine Learning Research* (2019).

[Duc+08] J. Duchi, S. Shalev-Shwartz, Y. Singer, and T. Chandra. “Efficient projections onto the ℓ_1-ball for learning in high dimensions”. In: *Proceedings of the 25th international conference on Machine learning*. 2008.

[DGN21] J. C. Duchi, P. W. Glynn, and H. Namkoong. “Statistics of robust optimization: A generalized empirical likelihood approach”. In: *Mathematics of Operations Research* 46.3 (2021).

[DN21] J. C. Duchi and H. Namkoong. “Learning models with uniform performance via distributionally robust optimization”. In: *The Annals of Statistics* 49.3 (2021).

[Dwo+12] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. Zemel. “Fairness through awareness”. In: *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*. 2012.

[FCG20] S. Frei, Y. Cao, and Q. Gu. “Agnostic learning of a single neuron with gradient descent”. In: *Advances in Neural Information Processing Systems* 33 (2020).

[GGK20] S. Goel, A. Gollakota, and A. R. Klivans. “Statistical-Query Lower Bounds via Functional Gradients”. In: *Annual Conference on Neural Information Processing Systems*. 2020.

[GKK19] S. Goel, S. Karmalkar, and A. R. Klivans. “Time/Accuracy Tradeoffs for Learning a ReLU with respect to Gaussian Marginals”. In: *Advances in Neural Information Processing Systems 32*. 2019.

[Gol+23] A. Gollakota, P. Gopalan, A. R. Klivans, and K. Stavropoulos. “Agnostically Learning Single-Index Models using Omnipredictors”. In: *Thirty-seventh Conference on Neural Information Processing Systems*. 2023.

[GSS14] I. J. Goodfellow, J. Shlens, and C. Szegedy. “Explaining and harnessing adversarial examples”. In: *arXiv preprint arXiv:1412.6572* (2014).

[HM13] M. Hardt and A. Moitra. “Understanding Alternating Minimization for Matrix Completion”. In: 2013.

[Has+18a] T. Hashimoto, M. Srivastava, H. Namkoong, and P. Liang. “Fairness Without Demographics in Repeated Loss Minimization”. In: *ICML*. 2018.

[Has+18b] T. Hashimoto, M. Srivastava, H. Namkoong, and P. Liang. “Fairness without demographics in repeated loss minimization”. In: *International Conference on Machine Learning*. PMLR. 2018.

[Hau92] D. Haussler. “Decision theoretic generalizations of the PAC model for neural net and other learning applications”. In: *Information and Computation* 100 (1992), pp. 78–150.

[Hu+18] W. Hu, G. Niu, I. Sato, and M. Sugiyama. “Does distributionally robust supervised learning give robust classifiers?” In: *ICML*. 2018.

[Hua+06] J. Huang, A. Gretton, K. Borgwardt, B. Schölkopf, and A. Smola. “Correcting sample selection bias by unlabeled data”. In: *Advances in neural information processing systems* 19 (2006).

[Kak+11] S. M. Kakade, V. Kanade, O. Shamir, and A. Kalai. “Efficient learning of generalized linear and single index models with isotonic regression”. In: *Advances in Neural Information Processing Systems* 24 (2011).

[KS09] A. T. Kalai and R. Sastry. “The Isotron Algorithm: High-Dimensional Isotonic Regression.” In: *COLT*. 2009.

[KSA19] S. M. M. Kalan, M. Soltanolkotabi, and S. Avestimehr. “Fitting relus via sgd and quantized sgd”. In: *2019 IEEE International Symposium on Information Theory (ISIT)*. IEEE. 2019.

[Kal+22] N. Kallus, X. Mao, K. Wang, and Z. Zhou. “Doubly Robust Distributionally Robust Off-Policy Evaluation and Learning”. In: *ICML*. 2022.

[KMM20] S. Karmakar, A. Mukherjee, and R. Muthukumar. “A Study of Neural Training with Iterative Non-Gradient Methods”. In: *arXiv e-prints* (2020), arXiv–2005.

[KSS92] M. J. Kearns, R. E. Schapire, and L. M. Sellie. “Toward efficient agnostic learning”. In: *Proceedings of the fifth annual workshop on Computational learning theory*. 1992.

[KKM17] A. R. Klivans, P. K. Kothari, and R. Meka. “Learning Halfspaces and Neural Networks with Random Initialization”. In: *Proceedings of the 30th Annual Conference on Learning Theory*. 2017.

[Kuh+19] D. Kuhn, P. M. Esfahani, V. A. Nguyen, and S. Shafieezadeh-Abadeh. “Wasserstein distributionally robust optimization: Theory and applications in machine learning”. In: *Operations research & management science in the age of analytics*. Informs, 2019, pp. 130–166.

[Lam13] H. Lam. “Robust Sensitivity Analysis for Stochastic Systems”. In: *Mathematics of Operations Research* (2013).

[Liu+21] E. Z. Liu, B. Haghgoo, A. S. Chen, A. Raghunathan, P. W. Koh, S. Sagawa, P. Liang, and C. Finn. “Just Train Twice: Improving Group Robustness without Training Group Information”. In: *ICML*. 2021.

[Liu+22] Z. Liu, Q. Bai, J. Blanchet, P. Dong, W. Xu, Z. Zhou, and Z. Zhou. “Distributionally Robust Q-Learning”. In: *ICML*. 2022.

[Lot+23] K. Lotidis, N. Bambos, J. Blanchet, and J. Li. “Wasserstein Distributionally Robust Linear-Quadratic Estimation under Martingale Constraints”. In: *AISTATS*. 2023.

[MMR09] Y. Mansour, M. Mohri, and A. Rostamizadeh. “Domain adaptation: Learning bounds and algorithms”. In: *arXiv preprint arXiv:0902.3430* (2009).

[MR18] P. Manurangsi and D. Reichman. “The computational complexity of training relu (s)”. In: *arXiv preprint arXiv:1810.04207* (2018).

[MDH24] R. Mehta, J. Diakonikolas, and Z. Harchaoui. “A Primal-Dual Algorithm for Faster Distributionally Robust Optimization”. In: *arXiv preprint arXiv:2403.10763* (2024).

[MBM18] S. Mei, Y. Bai, and A. Montanari. “THE LANDSCAPE OF EMPIRICAL RISK FOR NONCONVEX LOSSES”. In: *The Annals of Statistics* 46.6A (2018), pp. 2747–2774.

[ND16] H. Namkoong and J. C. Duchi. “Stochastic gradient methods for distributionally robust optimization with f-divergences”. In: *Advances in neural information processing systems* 29 (2016).

[NW72] J. A. Nelder and R. W. M. Wedderburn. “Generalized linear models”. In: *Journal of the Royal Statistical Society: Series A (General)* 135.3 (1972).

[Ore+19] Y. Oren, S. Sagawa, T. B. Hashimoto, and P. Liang. “Distributionally robust language modeling”. In: *arXiv preprint arXiv:1909.02060* (2019).

[PY09] S. J. Pan and Q. Yang. “A survey on transfer learning”. In: *IEEE Transactions on knowledge and data engineering* 22.10 (2009).

[Pat+15] V. M. Patel, R. Gopalan, R. Li, and R. Chellappa. “Visual domain adaptation: A survey of recent advances”. In: *IEEE signal processing magazine* 32.3 (2015).

[Qi+21] Q. Qi, Z. Guo, Y. Xu, R. Jin, and T. Yang. “An online method for a class of distributionally robust optimization with non-convex objectives”. In: *Advances in Neural Information Processing Systems* 34 (2021), pp. 10067–10080.

[RM22] H. Rahimian and S. Mehrotra. “Frameworks and results in distributionally robust optimization”. In: *Open Journal of Mathematical Optimization* (2022).

[Sap+23] H. Sapkota, D. Wang, Z. Tao, and Q. Yu. “Distributionally Robust Ensemble of Lottery Tickets Towards Calibrated Sparse Network Training”. In: *NeurIPS*. 2023.

[Sha17] A. Shapiro. “Distributionally robust stochastic programming”. In: *SIAM Journal on Optimization* 27.4 (2017).

[Sha+20] V. D. Sharma, M. Toubeh, L. Zhou, and P. Tokekar. “Risk-Aware Planning and Assignment for Ground Vehicles using Uncertain Perception from Aerial Vehicles”. In: *2020 IEEE/RSJ (IROS)*. 2020.

[Shi00] H. Shimodaira. “Improving predictive inference under covariate shift by weighting the log-likelihood function”. In: *Journal of statistical planning and inference* 90.2 (2000).

[Sim02] J. Sima. “Training a single sigmoidal neuron is hard”. In: *Neural computation* 14.11 (2002).

[SND18] A. Sinha, H. Namkoong, and J. Duchi. “Certifying Some Distributional Robustness with Principled Adversarial Training”. In: *International Conference on Learning Representations*. 2018.

[Sol17] M. Soltanolkotabi. “Learning ReLUs via gradient descent”. In: *Advances in neural information processing systems* 30 (2017).

[Son+22] C. Song, C. Y. Lin, S. Wright, and J. Diakonikolas. “Coordinate Linear Variance Reduction for Generalized Linear Programming”. In: *NeurIPS*. 2022.

[SWD21] C. Song, S. J. Wright, and J. Diakonikolas. “Variance Reduction via Primal-Dual Accelerated Dual Averaging for Nonsmooth Convex Finite-Sums”. In: *Proc. ICML’21*. 2021.

[SJ19] M. Staib and S. Jegelka. “Distributionally robust optimization and generalization in kernel methods”. In: *Advances in Neural Information Processing Systems* 32 (2019).

[Tan+18] C. Tan, F. Sun, T. Kong, W. Zhang, C. Yang, and C. Liu. “A survey on deep transfer learning”. In: *Artificial Neural Networks and Machine Learning–ICANN 2018*. Springer. 2018.

[Wan+23a] P. Wang, N. Zarifis, I. Diakonikolas, and J. Diakonikolas. “Robustly Learning a Single Neuron via Sharpness”. In: *Proceedings of the 40th International Conference on Machine Learning*. Vol. 202. Proceedings of Machine Learning Research. PMLR, July 2023, pp. 36541–36577.

[Wan+23b] S. Wang, N. Si, J. Blanchet, and Z. Zhou. “A Finite Sample Complexity Bound for Distributionally Robust Q-learning”. In: *AISTATS*. 2023.

[Wan+23c] Z. Wang, L. Shen, T. Liu, T. Duan, Y. Zhu, D. Zhan, D. S. Doermann, and M. Gao. “Defending against Data-Free Model Extraction by Distributionally Robust Defensive Training”. In: *NIPS*. 2023.

[Xu+20] Z. Xu, C. Dan, J. Khim, and P. Ravikumar. “Class-weighted classification: Trade-offs and robust approaches”. In: *International conference on machine learning*. PMLR. 2020.

[Yan+23] Z. Yang, Y. Guo, P. Xu, A. Liu, and A. Anandkumar. “Distributionally Robust Policy Gradient for Offline Contextual Bandits”. In: *AISTATS*. 2023.

[YS20] G. Yehudai and O. Shamir. “Learning a single neuron with gradient methods”. In: *Conference on Learning Theory*. 2020, pp. 3756–3786.

[Yu+23] Z. Yu, L. Dai, S. Xu, S. Gao, and C. P. Ho. “Fast Bellman Updates for Wasserstein Distributionally Robust MDPs”. In: *NeurIPS*. Curran Associates, Inc., 2023.

[Zar+24] N. Zarifis, P. Wang, I. Diakonikolas, and J. Diakonikolas. “Robustly Learning Single-Index Models via Alignment Sharpness”. In: *41th International Conference on Machine Learning, arXiv preprint: 2402.17756* (2024).

[Zha+21] R. Zhai, C. Dan, Z. Kolter, and P. Ravikumar. “DORO: Distributional and Outlier Robust Optimization”. In: *Proceedings of the 38th International Conference on Machine Learning*. Vol. 139. Proceedings of Machine Learning Research. PMLR, July 2021, pp. 12345–12355.

[Zhu+20] J.-J. Zhu, W. Jitkrittum, M. Diehl, and B. Schölkopf. “Kernel distributionally robust optimization”. In: *arXiv preprint arXiv:2006.06981* (2020).

Supplementary Material

Organization In Appendix A we briefly discuss related work. In Appendix B we set up some additional preliminaries for the rest of the appendix. In Appendix C we show that expectations of some important functions with respect to $\widehat{p}^*$ are close to their expectation with respect to p^*. In Appendix D we give a detailed proof of an upper bound on the gap of the iterates our algorithm generates (i.e. Lemma 3.3). Finally, in Appendix F we show that the estimate of $\boldsymbol{w}^*$ our algorithm returns is a constant factor approximation to the squared loss of $\boldsymbol{w}^*$ with respect to the target distribution.

A Related Work

Learning Noisy Neurons The problem of learning noisy neurons has been extensively explored in recent decades; notable works include [Kak+11; KS09; KKM17; NW72]. In the recent past, the focus has shifted towards specific activation functions such as ReLUs, under both easy noise models such as realizable/random additive noise [KSA19; Sol17; YS20] and more difficult ones like adversarial label noise [Dia+20; DKZ20; Dia+21; Dia+22a; Dia+22b; GGK20; GKK19; Wan+23a].

Even with clean labels, this problem has exponentially many local minima when using squared loss [AHW95]. Unfortunately, directly minimizing the squared loss using (S)GD on a bounded distribution does not converge to the global optimum with probability 1 [YS20]. Even so, gradient based methods can achieve suboptimal rates in the agnostic setting ($\sqrt{\mathrm{OPT}_0}$ vs OPT_0) for distributions with mild distributional assumptions [FCG20]. Making slightly stronger assumptions on the marginal does allow us to get efficient constant factor approximations. [Dia+20] developed an efficient learning method that is able to handle this in the presence of adversarial label noise and for isotropic logconcave distributions of the covariates. This was later extended to broader classes of activation functions and under weaker distributional assumptions by [Dia+22a; Wan+23a]. Without specific distributional assumptions, learning remains computationally difficult [Dia+22b; HM13]. The challenges extend to distribution-free scenarios with semi-random label noise, where methods like those in [DPT21] address bounded noise, and [KMM20] and [Che+20] explore stricter forms of Massart noise in learning a neuron with noise. In this paper, we consider the harder setting of distributionally robust optimization, where an adversary is allowed to impose not only errors in the labels, but also adversarial shifts in the underlying distribution of the covariates.

Distributionally Robust Optimization Distributional mismatches in data have been extensively studied in the context of learning from noisy data. This includes covariate shift, where the marginal distributions might be perturbed, [BBS07; Hua+06; Shi00], and changes in label proportions [Dwo+12; Xu+20]. This research also extends to domain adaptation and transfer learning [Ben+10; MMR09; PY09; Pat+15; Tan+18]. Distributionally robust optimization (DRO) has a rich history in optimization [BEN09; Sha17] and has gained traction in machine learning [DGN21; DN21; Kuh+19; ND16; SJ19; Zhu+20], showing mixed success across applications like language modeling [Ore+19], class imbalance correction [Xu+20], and group fairness [Has+18b].

Specifically this has also been studied in the context of linear regression and other function approximation [BMN21; CP18; DN21]. Typically, DRO is often very sensitive to additional sources of noise, such as outliers ([Has+18a; Hu+18; Zha+21]). However, prior work makes strong assumptions on the label noise as well as requiring convexity of the loss. We study the problem of learning a neuron where the labels have no guaranteed structure, effectively studying the setting for a combination of two notions of robustness — agnostic learning as well as covariate shift.

B Supplementary Preliminaries

B.1 Additional Notation

Given an $m \times n$ matrix $\boldsymbol{A}$, the operator norm of $\boldsymbol{A}$ is defined in the usual way as $\|\boldsymbol{A}\|_{\mathrm{op}} = \sup\{\|\boldsymbol{A}\boldsymbol{x}\|_2 : \boldsymbol{x} \in \mathbb{R}^n, \|\boldsymbol{x}\|_2 \leq 1\}$. For problems (P) and (P'), we use $(P) \equiv (P')$ to denote the equivalence of (P) and (P'). For a vector space $\mathbb{E}$, we use $\mathbb{E}^*$ to denote its dual space.

B.2 Standard Facts and Proofs

Fact B.1 (Young's inequality)**.** *If $a \geq 0$ and $b \geq 0$ are nonnegative real numbers and if $p > 1$ and $q > 1$ are real numbers such that*

$$\frac{1}{p} + \frac{1}{q} = 1,$$

then

$$ab \leq \frac{a^p}{p} + \frac{b^q}{q}.$$

Equality holds if and only if $a^p = b^q$.

Fact B.2 (Hoeffding's Inequality)**.** *Let $X_1, X_2, \ldots, X_n$ be independent random variables such that $a_i \leq X_i \leq b_i$ almost surely for all i. Let $\overline{X} = \frac{1}{n}\sum_{i=1}^n X_i$. Then, for any $t > 0$,*

$$\Pr\left[|\overline{X} - \mathbb{E}[\overline{X}]| \geq t\right] \leq 2\exp\left(-\frac{2nt^2}{\sum_{i=1}^n (b_i - a_i)^2}\right).$$

Fact 2.7. *Let $\psi(\boldsymbol{x}) = \phi(\boldsymbol{x}) + \langle \boldsymbol{a}, \boldsymbol{x}\rangle + b$ for some $\boldsymbol{a} \in \mathbb{R}^N$ and $b \in \mathbb{R}$. Then $D_\psi(\boldsymbol{y}, \boldsymbol{x}) = D_\phi(\boldsymbol{y}, \boldsymbol{x})$ for all $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^N$, i.e., the Bregman divergence is blind to the addition of affine terms to function ϕ.*

Proof of Fact 2.7. The Bregman divergence D_ϕ and D_ψ are defined by:

$$D_\phi(\boldsymbol{y}, \boldsymbol{x}) = \phi(\boldsymbol{y}) - \phi(\boldsymbol{x}) - \langle \nabla\phi(\boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x}\rangle,$$

$$D_\psi(\boldsymbol{y}, \boldsymbol{x}) = \psi(\boldsymbol{y}) - \psi(\boldsymbol{x}) - \langle \nabla\psi(\boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x}\rangle.$$

Since $\nabla\psi(\boldsymbol{x}) = \nabla\phi(\boldsymbol{x}) + \boldsymbol{a}$, substituting in the definition gives:

$$\begin{aligned} D_\psi(\boldsymbol{y}, \boldsymbol{x}) &= \phi(\boldsymbol{y}) + \langle \boldsymbol{a}, \boldsymbol{y}\rangle + b - (\phi(\boldsymbol{x}) + \langle \boldsymbol{a}, \boldsymbol{x}\rangle + b) - \langle \nabla\phi(\boldsymbol{x}) + \boldsymbol{a}, \boldsymbol{y} - \boldsymbol{x}\rangle \\ &= \phi(\boldsymbol{y}) - \phi(\boldsymbol{x}) - \langle \nabla\phi(\boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x}\rangle = D_\phi(\boldsymbol{y}, \boldsymbol{x}). \end{aligned}$$

Thus, the Bregman divergence is blind to the addition of linear terms to the function ϕ. □

B.3 Auxiliary Facts

We first state and prove Lemma B.3 to obtain upper bounds on the norm of each point, projections onto vectors of norm at most W, and the loss value at each point.

Lemma B.3 (Boundedness)**.** *Fix $\boldsymbol{w} \in \mathscr{B}(W)$. For all samples $(\boldsymbol{x}_i, y_i)$ with truncated labels $|y_i| < M$ as per Fact 2.6 and bounded covariates as per Assumption 2.3, it holds that*

$$\boldsymbol{w} \cdot \boldsymbol{x}_i \leq WS \tag{18}$$

$$\|\boldsymbol{x}_i\|_2 \leq S\sqrt{d} \tag{19}$$

$$(\sigma(\boldsymbol{w} \cdot \boldsymbol{x}_i) - y_i)^2 \leq 2\beta^2 W^2 (S^2 + C_M^2 B^2 \log^2(WB\beta/\epsilon)) \tag{20}$$

Proof. Equation (18) follows from Assumption 2.3, as

$$\boldsymbol{w} \cdot \boldsymbol{x}_i = \|\boldsymbol{w}\|_2 \frac{\boldsymbol{w}}{\|\boldsymbol{w}\|_2} \cdot \boldsymbol{x}_i \leq \|\boldsymbol{w}\|_2 S \leq WS.$$

To prove Equation (19), for each coordinate $j \in [d]$, we have $|\boldsymbol{x}^{(j)}| = \operatorname{sign}(\boldsymbol{x}^{(j)})\boldsymbol{e}_j \cdot \boldsymbol{x} \leq S$, by again using Assumption 2.3. Therefore, $\|\boldsymbol{x}\|_2 \leq S\sqrt{d}$.

For Equation (20), we recall Fact 2.6 that for some sufficiently large absolute constant C_M, it holds that $|y| \leq M := C_M WB\beta \log(\beta BW/\epsilon)$. Thus, Equation (20) follows from Young's inequality (Fact B.1) and Equation (19), since $|\sigma(t)| \leq \beta|t|$, which follows from β-Lipschitzness of σ and $\sigma(0) = 0$. □

Lemma B.4 (Bounds on $\boldsymbol{v}$)**.** *Let*

$$\boldsymbol{v}(\boldsymbol{w};\boldsymbol{x},y) = 2(\sigma(\boldsymbol{w}\cdot\boldsymbol{x}) - y)\beta\boldsymbol{x}$$

Then $\boldsymbol{v}$ is uniformly bounded by G in ℓ_2-norm and κ-Lipschitz for all samples $\boldsymbol{x}, y$ with truncated labels $|y| < M$ and bounded covariates as per Assumption 2.3, where $G = 2\beta S\sqrt{d}(\sqrt{2}\beta WS + M)$ and $\kappa = 2\beta^2 S^2 d$.

Proof. We first uniformly upper bound $\boldsymbol{v}$. An application of Lemma B.3 gives us,

$$\begin{aligned}\|\boldsymbol{v}(\boldsymbol{w};\boldsymbol{x},y)\|_2^2 &= 4\beta^2\,(\sigma(\boldsymbol{w}\cdot\boldsymbol{x}) - y)^2\,\|\boldsymbol{x}\|^2\\ &\le 4\,\beta^2\,(2\beta^2W^2S^2 + M^2)\,S^2d.\end{aligned}$$

Taking square roots, we get $\|\boldsymbol{v}(\boldsymbol{w};\boldsymbol{x},y)\|_2 \le 2\beta S\sqrt{d}(\sqrt{2}\beta WS + M) =: G$.

We now upper bound the Lipschitz constant κ. We will use the fact that σ is β-Lipschitz.

$$\begin{aligned}\|\nabla_{\boldsymbol{w}}\boldsymbol{v}(\boldsymbol{w};\boldsymbol{x},y)\|_2 &= 2\beta|\sigma'(\boldsymbol{w}\cdot\boldsymbol{x})|\|\boldsymbol{x}\boldsymbol{x}^T\|_2\\ &= 2\beta|\sigma'(\boldsymbol{w}\cdot\boldsymbol{x})|\|\boldsymbol{x}\|_2^2\\ &= 2\beta\cdot\beta\,S^2d = 2\beta^2S^2d =: \kappa.\end{aligned}$$

□

Corollary B.5. *Fix a reference distribution p_0. Suppose $\|\boldsymbol{v}(\boldsymbol{w};\boldsymbol{x},y)\|_2 \le G$ for all $\boldsymbol{w}$ almost surely. Then for all distributions $p, q \in \mathscr{P}(p_0)$ it holds that*

$$\|\mathbb{E}_p[\boldsymbol{v}(\boldsymbol{w};\boldsymbol{x},y)] - \mathbb{E}_q[\boldsymbol{v}(\boldsymbol{w};\boldsymbol{x},y)]\|_2^2 \le G^2 D_\phi(p,q).$$

Proof.

$$\begin{aligned}\|\mathbb{E}_p[\boldsymbol{v}(\boldsymbol{w};\boldsymbol{x},y)] - \mathbb{E}_q[\boldsymbol{v}(\boldsymbol{w};\boldsymbol{x},y)]\|_2^2 &= \Big\|\int \boldsymbol{v}(\boldsymbol{w};\boldsymbol{x},y)(\mathrm{d}p - \mathrm{d}q)\Big\|_2^2\\ &= \Big\|\int \boldsymbol{v}(\boldsymbol{w})\Big(\frac{\mathrm{d}p}{\mathrm{d}p_0} - \frac{\mathrm{d}q}{\mathrm{d}p_0}\Big)\mathrm{d}p_0\Big\|_2^2\\ &\overset{(i)}{\le} \int\Big\|\boldsymbol{v}(\boldsymbol{w})\Big(\frac{\mathrm{d}p}{\mathrm{d}p_0} - \frac{\mathrm{d}q}{\mathrm{d}p_0}\Big)\Big\|_2^2\mathrm{d}p_0\\ &\overset{(ii)}{\le} G^2\int\Big(\frac{\mathrm{d}p}{\mathrm{d}p_0} - \frac{\mathrm{d}q}{\mathrm{d}p_0}\Big)^2\mathrm{d}p_0\\ &= G^2 D_{\chi^2(\cdot,p_0)}(p,q),\end{aligned}$$

where (i) is an application of Jensen's inquality and (ii) follows from Lemma B.4. □

Claim B.6 (Convergence Rate)**.** *For all $i \ge 0$, let a_i be defined as in Line 3. Then it holds that $\frac{2G^2{a_i}^2}{1+0.5c_1A_i} \le \nu_0 + \nu A_{i-1}$ and $\frac{2\kappa^2{a_i}^2}{1+0.5c_1A_i} \le \frac{1+0.5c_1A_{i-1}}{4}$ for all i. Moreover, $A_k = \sum_{i=0}^k a_i = ((1 + \frac{\min\{\nu,c_1/8\}}{2\max\{\kappa,G\}})^k - 1)\min\{\nu_0,1/4\}/\min\{\nu,c_1/8\}$.*

Proof. In order for both $\frac{2G^2{a_i}^2}{1+0.5c_1A_i} \le \nu_0 + \nu A_{i-1}$ and $\frac{2\kappa^2{a_i}^2}{1+0.5c_1A_i} \le \frac{1+0.5c_1A_{i-1}}{4}$ to hold for all iterations i, it suffices that

$$\frac{4\max\{G,\kappa\}^2{a_i}^2}{1+0.5c_1A_i} \le \min\{\nu_0,1/4\} + \min\{\nu,c_1/8\}A_{i-1},$$

for which it suffices to enforce

$$4\max\{G,\kappa\}^2{a_i}^2 = (\min\{\nu_0,1/4\} + \min\{\nu,c_1/8\}A_{i-1})^2,$$

where we used $A_{i-1} \le A_i$.

Taking a square root on both sides using $a_i > 0$, we obtain

$$2\max\{G,\kappa\}a_i = \min\{\nu_0,1/4\} + \min\{\nu,c_1/8\}A_{i-1}.$$

Solving this recurrence relation using Mathematica, we compute that for all iterations i and k,

$$a_i = \Big(1 + \frac{\min\{\nu, c_1/8\}}{2\max\{\kappa, G\}}\Big)^{i-1} \min\{\nu_0, 1/4\}/(2\max\{\kappa, G\})$$

$$A_k = \sum_{i=0}^{k} a_i = ((1 + \frac{\min\{\nu, c_1/8\}}{2\max\{\kappa, G\}})^k - 1) \min\{\nu_0, 1/4\}/\min\{\nu, c_1/8\}.$$

□

Remark B.7. Note that in our case, the gap is not guarenteed to be non-negative as is usually the case for convex-concave min-max problems. Recall that $\mathrm{Gap}(\boldsymbol{w}, \widehat{p}) = (L(\boldsymbol{w}, \widehat{p}^*) - L(\boldsymbol{w}^*, \widehat{p}^*)) + (L(\boldsymbol{w}^*, \widehat{p}^*) - L(\boldsymbol{w}^*, \widehat{p}))$

Consider the following example:

Let p_0 be the uniform distribution over $\{(-2, 2), (2, 1.5)\}$, $\sigma \equiv$ ReLU and $\nu = 0$. Then, $\boldsymbol{w}^* = -1$ and p^* is the distribution which places all its mass on $(2, 1.5)$. Then, $\mathrm{Gap}(1, p^*) = L(1, p^*) - L(-1, p^*) = 0.25 - 2.25 < 0$.

This is why we also need an explicit lower bound on the Gap that we have shown in Lemma 3.2.

C Concentration

Recall that that $\widehat{q}_{\boldsymbol{w}}$ is not guaranteed to act as an empirical estimate of $q_{\boldsymbol{w}}$, because we cannot draw samples from the (unknown) distribution $q_{\boldsymbol{w}}$ but only from p^0. In this section, we show that for certain important functions f, it holds that $\mathbb{E}_{\widehat{q}_{\boldsymbol{w}}}[f] \approx \mathbb{E}_{q_{\boldsymbol{w}}}[f]$. We will abuse terminology and say that f "concentrates" with respect to $q_{\boldsymbol{w}}$.

Organization: In Appendix C.1 we derive closed-form expressions for $q_{\boldsymbol{w}}$ and $\widehat{q}_{\boldsymbol{w}}$ in terms of p_0 and $\widehat{p}_0$ respectively. Note that bounded functions concentrate with respect to p_0. In Appendix C.2 we use the closed-form expressions found in Appendix C.1 to translate these concentration properties to $\widehat{q}_{\boldsymbol{w}}$. Finally, in Appendix C.3 we show that $\widehat{p}^*$ satisfies sharpness, $\widehat{\mathrm{OPT}} \approx \mathrm{OPT}$ and $\widehat{\mathrm{OPT}}_{(2)} \approx \mathrm{OPT}_{(2)}$.

C.1 Closed-form expression

The following lemma gives us a closed-form expression for $q_{\boldsymbol{w}}$ and $\widehat{q}_{\boldsymbol{w}}$ in terms of p_0 and $\widehat{p}_0$, respectively.

We start with an additional definition to Definition 1.2:

$R(\boldsymbol{w}; \widehat{p}_0) := \max_{\widehat{p} \in \mathcal{P}} \mathbb{E}_{(\boldsymbol{x},y)\sim\widehat{p}}(\sigma(\boldsymbol{w} \cdot \boldsymbol{x}) - y)^2 - \nu\chi^2(\widehat{p}, \widehat{p}_0)$, with the maximum achieved by $\widehat{q}_{\boldsymbol{w}}$,

Lemma C.1 (Closed-form $q_{\boldsymbol{w}}$). *Let p_0 be a fixed distribution. Then, there exists $\xi \in \mathbb{R}$ such that,*

$$\frac{\mathrm{d}q_{\boldsymbol{w}}}{\mathrm{d}p_0}(\boldsymbol{x}, y) = \frac{\max\{\ell(\boldsymbol{w}; \boldsymbol{x}, y) - \xi + 2\nu, 0\}}{2\nu}.$$

When p_0 is the empirical distribution $\widehat{p}_0(N)$, this result implies that there exists $\hat{\xi} \in \mathbb{R}$ such that

$$\widehat{q}_{\boldsymbol{w}}^{(i)} = \widehat{p}_0^{(i)} \frac{\max\{\ell(\boldsymbol{w}; \boldsymbol{x}_i, y_i) - \hat{\xi} + 2\nu, 0\}}{2\nu} \quad \text{for all } i \in [N].$$

The constants ξ and $\hat{\xi}$ can be interpreted as normalization that ensures $\int \mathrm{d}q_{\boldsymbol{w}} = \int \mathrm{d}\widehat{q}_{\boldsymbol{w}} = 1$.

Proof. Recall that the dual feasible set is given by $\mathcal{P} = \mathcal{P}(p_0) = \{p \ll p_0 : \int \mathrm{d}p = 1, p \geq 0\} = \{p \ll p_0 : \int \frac{\mathrm{d}p}{\mathrm{d}p_0}\mathrm{d}p_0 = 1, \frac{\mathrm{d}p}{\mathrm{d}p_0} \geq 0\}$ and the function $p \mapsto L(\boldsymbol{w}, p)$ is strongly concave.

Consider the following optimization problem

$$\max_{p \in \mathcal{P}(p_0)} L(\boldsymbol{w}, p) \equiv \max_{p \in \mathcal{P}(p_0)} \mathbb{E}_{(\boldsymbol{x},y)\sim p}\ell(\boldsymbol{w}; \boldsymbol{x}, y) - \nu\chi^2(p_0, p).$$

By Fact 2.8, the first-order necessary and sufficient condition that corresponds to $q_{\boldsymbol{w}} := \arg\max_{p \in \mathscr{P}(p_0)} L(\boldsymbol{w}, p)$ is the following: for any $p \in \mathscr{P}(p_0)$,

$$0 \geq \int \nabla_p L(\boldsymbol{w}, q_{\boldsymbol{w}}) \mathrm{d}(p - q_{\boldsymbol{w}}) = \int \nabla_p L(\boldsymbol{w}, q_{\boldsymbol{w}}) \Big(\frac{\mathrm{d}p}{\mathrm{d}p_0} - \frac{\mathrm{d}q_{\boldsymbol{w}}}{\mathrm{d}p_0} \Big) \mathrm{d}p_0, \tag{21}$$

where we recall that both $\nabla_p L(\boldsymbol{w}, q_{\boldsymbol{w}})$ and Radon–Nikodym derivatives $\frac{\mathrm{d}p}{\mathrm{d}p_0}$, $\frac{\mathrm{d}q_{\boldsymbol{w}}}{\mathrm{d}p_0}$ are real-valued measurable functions on $\mathbb{R}^d \times \mathbb{R}$. We will also write $\ell = \ell(\boldsymbol{w}^*, \cdot, \cdot)$ for short.

We claim Equation (21) is satisfied if there exists $\xi \in \mathbb{R}$ and a bounded measurable function $\psi \geq 0$ such that

$$\nabla_p L(\boldsymbol{w}, q_{\boldsymbol{w}})(\boldsymbol{x}, y) = \begin{cases} \xi & \text{if } \frac{\mathrm{d}q_{\boldsymbol{w}}}{\mathrm{d}p_0} > 0 \\ \xi - \psi(\boldsymbol{x}, y) & \text{otherwise.} \end{cases} \tag{22}$$

Indeed, for any $p \in \mathscr{P}(p_0)$,

$$\begin{aligned}
&\int \nabla_p L(\boldsymbol{w}, q_{\boldsymbol{w}}) \Big(\frac{\mathrm{d}p}{\mathrm{d}p_0} - \frac{\mathrm{d}q_{\boldsymbol{w}}}{\mathrm{d}p_0} \Big) \mathrm{d}p_0 \\
&= \int_{\frac{\mathrm{d}q_{\boldsymbol{w}}}{\mathrm{d}p_0} > 0} \nabla_p L(\boldsymbol{w}, q_{\boldsymbol{w}}) \Big(\frac{\mathrm{d}p}{\mathrm{d}p_0} - \frac{\mathrm{d}q_{\boldsymbol{w}}}{\mathrm{d}p_0} \Big) \mathrm{d}p_0 + \int_{\frac{\mathrm{d}q_{\boldsymbol{w}}}{\mathrm{d}p_0} = 0} \nabla_p L(\boldsymbol{w}, q_{\boldsymbol{w}}) \frac{\mathrm{d}p}{\mathrm{d}p_0} \mathrm{d}p_0 \\
&= \int_{\frac{\mathrm{d}q_{\boldsymbol{w}}}{\mathrm{d}p_0} > 0} \xi \Big(\frac{\mathrm{d}p}{\mathrm{d}p_0} - \frac{\mathrm{d}q_{\boldsymbol{w}}}{\mathrm{d}p_0} \Big) \mathrm{d}p_0 + \int_{\frac{\mathrm{d}q_{\boldsymbol{w}}}{\mathrm{d}p_0} = 0} (\xi - \psi) \frac{\mathrm{d}p}{\mathrm{d}p_0} \mathrm{d}p_0 \\
&\leq \int_{\frac{\mathrm{d}q_{\boldsymbol{w}}}{\mathrm{d}p_0} > 0} \xi \Big(\frac{\mathrm{d}p}{\mathrm{d}p_0} - \frac{\mathrm{d}q_{\boldsymbol{w}}}{\mathrm{d}p_0} \Big) \mathrm{d}p_0 + \int_{\frac{\mathrm{d}q_{\boldsymbol{w}}}{\mathrm{d}p_0} = 0} \xi \frac{\mathrm{d}p}{\mathrm{d}p_0} \mathrm{d}p_0 \\
&= \int_{\frac{\mathrm{d}q_{\boldsymbol{w}}}{\mathrm{d}p_0} > 0} \xi \frac{\mathrm{d}p}{\mathrm{d}p_0} \mathrm{d}p_0 + \int_{\frac{\mathrm{d}q_{\boldsymbol{w}}}{\mathrm{d}p_0} = 0} \xi \frac{\mathrm{d}p}{\mathrm{d}p_0} \mathrm{d}p_0 + \int_{\frac{\mathrm{d}q_{\boldsymbol{w}}}{\mathrm{d}p_0} > 0} \xi \Big(- \frac{\mathrm{d}q_{\boldsymbol{w}}}{\mathrm{d}p_0} \Big) \mathrm{d}p_0 \\
&= \int \xi \frac{\mathrm{d}p}{\mathrm{d}p_0} \mathrm{d}p_0 + \int_{\frac{\mathrm{d}q_{\boldsymbol{w}}}{\mathrm{d}p_0} > 0} \xi \Big(- \frac{\mathrm{d}q_{\boldsymbol{w}}}{\mathrm{d}p_0} \Big) \mathrm{d}p_0 \overset{(i)}{=} \xi - \xi = 0,
\end{aligned}$$

where (i) is because $\int_{\frac{\mathrm{d}q_{\boldsymbol{w}}}{\mathrm{d}p_0} > 0} \Big(\frac{\mathrm{d}q_{\boldsymbol{w}}}{\mathrm{d}p_0} \Big) \mathrm{d}p_0 = \int \Big(\frac{\mathrm{d}q_{\boldsymbol{w}}}{\mathrm{d}p_0} \Big) \mathrm{d}p_0$.

Observe from the definition of $L(\boldsymbol{w}, q_{\boldsymbol{w}})$ that $\nabla_p L(\boldsymbol{w}, q_{\boldsymbol{w}})(\boldsymbol{x}, y) = \ell(\boldsymbol{w}; \boldsymbol{x}, y) - 2\nu(\frac{\mathrm{d}q_{\boldsymbol{w}}}{\mathrm{d}p_0}(\boldsymbol{x}, y) - 1)$. Plugging this into Equation (22) and rearranging, we have,

$$\frac{\mathrm{d}p^*}{\mathrm{d}p_0} = \begin{cases} \frac{2\nu + \ell - \xi}{2\nu} & \text{if } \frac{\mathrm{d}q_{\boldsymbol{w}}}{\mathrm{d}p_0} > 0 \\ \frac{2\nu + \ell - \xi + \psi}{2\nu} & \text{if } \frac{\mathrm{d}q_{\boldsymbol{w}}}{\mathrm{d}p_0} = 0 \end{cases}$$

For the case where $\frac{\mathrm{d}p^*}{\mathrm{d}p_0} > 0$, $\frac{\mathrm{d}p^*}{\mathrm{d}p_0} = \frac{2\nu + \ell - \xi}{2\nu}$, so the condition $\frac{\mathrm{d}q_{\boldsymbol{w}}}{\mathrm{d}p_0} > 0$ becomes $2\nu + \ell - \xi > 0$. On the other hand, if the above condition fails, it has to be the case that $\frac{\mathrm{d}q_{\boldsymbol{w}}}{\mathrm{d}p_0} = 0$. Combining, we have

$$\frac{\mathrm{d}q_{\boldsymbol{w}}}{\mathrm{d}p_0} = \begin{cases} \frac{2\nu + \ell - \xi}{2\nu} & \text{if } 2\nu + \ell - \xi > 0 \\ 0 & \text{otherwise} \end{cases} = \frac{\max\{2\nu + \ell - \xi, 0\}}{2\nu}.$$

□

Instead of using the expression in Lemma C.1, we will set ν to be big enough to ensure that there is no maximum in the expression for $q_{\boldsymbol{w}}$. This is captured in Corollary C.2.

Corollary C.2 (Simpler Closed-form $q_{\boldsymbol{w}}$)**.** *Fix $\boldsymbol{w} \in \mathbb{R}^d$. If $\nu \geq \frac{1}{2}\mathbb{E}_{p_0}\ell(\boldsymbol{w})$, then*

$$\frac{\mathrm{d}q_{\boldsymbol{w}}}{\mathrm{d}p_0}(\boldsymbol{x}, y) = 1 + \frac{\ell(\boldsymbol{w}; \boldsymbol{x}, y) - \mathbb{E}_{p_0}\ell(\boldsymbol{w})}{2\nu}.$$

Similarly, if $\nu \geq \frac{1}{2}\mathbb{E}_{\widehat{p}_0}\ell(\boldsymbol{w})$, then ${q_{\boldsymbol{w}}}^{(i)} > 0$ for all $i \in [N]$, and

$$\widehat{q}_{\boldsymbol{w}}^{(i)} = \widehat{p}_0^{(i)} + \frac{\ell(\boldsymbol{w}; \boldsymbol{x}_i, y_i) - \mathbb{E}_{\widehat{p}_0}\ell(\boldsymbol{w})}{2\nu} \widehat{p}_0^{(i)} \quad \text{for all } i \in [N].$$

Furthermore, if $\nu \geq \mathbb{E}_{\widehat{p}_0}\ell(\boldsymbol{w})$, then, in particular, for each coordinate $j \in [N]$, we have

$$\widehat{q}_{\boldsymbol{w}}^{(j)} \geq \widehat{p}_0^{(j)}/2$$

Similarly, if $\nu \geq \mathbb{E}_{p_0}\ell(\boldsymbol{w})$, then for any non-negative function g, we have

$$\int g \, \mathrm{d}p_{\boldsymbol{w}} \geq \frac{1}{2}\int g \, \mathrm{d}p_0$$

Recall from Definition 1.2 that when $\boldsymbol{w} = \boldsymbol{w}^*$, we define $p^* = q_{\boldsymbol{w}^*}$ and $\widehat{p}^* = \widehat{q}_{\boldsymbol{w}^*}$. If $\nu \geq 8\beta^2\sqrt{6B}\sqrt{\mathrm{OPT}_{(2)}+\epsilon}/c_1$ as assumed in Theorem 3.1, then both conditions $\nu \geq \mathbb{E}_{\widehat{p}_0}\ell(\boldsymbol{w})$ and $\nu \geq \mathbb{E}_{p_0}\ell(\boldsymbol{w})$ hold.

Proof. Setting $\nu \geq \frac{1}{2}\mathbb{E}_{p_0}\ell(\boldsymbol{w})$ and $\xi = \mathbb{E}_{p_0}\ell(\boldsymbol{w})$ in Lemma C.1 implies $\ell(\boldsymbol{w};\boldsymbol{x},y) - \mathbb{E}_{p_0}\ell(\boldsymbol{w};\boldsymbol{x},y) + 2\nu > 0$, which, in turn, means

$$\begin{aligned}\frac{\mathrm{d}q_{\boldsymbol{w}}}{\mathrm{d}p_0}(\boldsymbol{x},y) &= \frac{\max\{\ell(\boldsymbol{w};\boldsymbol{x},y) - \mathbb{E}_{p_0}\ell(\boldsymbol{w};\boldsymbol{x},y) + 2\nu, 0\}}{2\nu} \\ &= \frac{\ell(\boldsymbol{w};\boldsymbol{x},y) - \mathbb{E}_{p_0}\ell(\boldsymbol{w};\boldsymbol{x},y) + 2\nu}{2\nu}.\end{aligned}$$

The empirical version follows analogously.

To establish the last claim, we show that $8\beta^2\sqrt{6B}\sqrt{\widehat{\mathrm{OPT}}_{(2)}}/c_1 \geq \mathbb{E}_{\widehat{p}_0}\ell(\boldsymbol{w})$. By Corollary C.9, it holds that

$$\sqrt{\widehat{\mathrm{OPT}}_{(2)}} \geq \widehat{\mathrm{OPT}} = \mathbb{E}_{\widehat{p}^*}(\sigma(\boldsymbol{w}^* \cdot \boldsymbol{x}) - y)^2 \geq \mathbb{E}_{\widehat{p}^*}(\sigma(\boldsymbol{w}^* \cdot \boldsymbol{x}) - y)^2 - \nu\chi^2(\widehat{p}^*, \widehat{p}_0) = L(\boldsymbol{w}^*, \widehat{p}^*).$$

By definition of $\widehat{p}^*$, we have $L(\boldsymbol{w}^*, \widehat{p}^*) \geq L(\boldsymbol{w}^*, \widehat{p}_0) = \mathbb{E}_{\widehat{p}_0}(\sigma(\boldsymbol{w}^* \cdot \boldsymbol{x}) - y)^2 - \nu\chi^2(\widehat{p}_0, \widehat{p}_0) = \mathbb{E}_{\widehat{p}_0}(\sigma(\boldsymbol{w}^* \cdot \boldsymbol{x}) - y)^2$. Combining, we obtain $8\beta^2\sqrt{6B}\sqrt{\widehat{\mathrm{OPT}}_{(2)}}/c_1 \geq \mathbb{E}_{\widehat{p}_0}\ell(\boldsymbol{w})8\beta^2\sqrt{6B}/c_1$. We conclude by observing $8\beta^2\sqrt{6B}/c_1 \geq 1$. □

Another consequence of Corollary C.2 is a closed form expression for the risk, as a variance-regularized loss, similar to [DN19; Lam13].

Corollary C.3. *Fix an arbitrary distribution p_0. Recall the risk defined in Definition 1.2,*

$$R(\boldsymbol{w}; p_0) := \max_{p \ll p_0} \mathbb{E}_{(\boldsymbol{x},y)\sim p}\ell(\boldsymbol{w};\boldsymbol{x},y) - \nu\chi^2(p, p_0).$$

If $\nu \geq \frac{1}{2}\mathbb{E}_{\widehat{p}_0}\ell(\boldsymbol{w})$, it holds that

$$\begin{aligned}\chi^2(q_{\boldsymbol{w}}, p_0) &= \frac{\mathbb{E}_{p_0}[\ell^2(\boldsymbol{w})] - (\mathbb{E}_{p_0}[\ell(\boldsymbol{w})])^2}{4\nu^2} \\ R(\boldsymbol{w}; p_0) &= \mathbb{E}_{p_0}[\ell(\boldsymbol{w})] + \frac{\mathbb{E}_{p_0}[\ell^2(\boldsymbol{w})] - (\mathbb{E}_{p_0}[\ell(\boldsymbol{w})])^2}{4\nu}.\end{aligned}$$

Proof. Both of these follow from Corollary C.2. To see the first equality holds, observe that,

$$\begin{aligned}\chi^2(q_{\boldsymbol{w}}, p_0) &= \mathbb{E}_{p_0}\left(\frac{\mathrm{d}q_{\boldsymbol{w}}}{\mathrm{d}p_0} - 1\right)^2 \\ &= \mathbb{E}_{p_0}\frac{(\ell(\boldsymbol{w};\boldsymbol{x},y) - \mathbb{E}_{p_0}\ell(\boldsymbol{w};\boldsymbol{x},y))^2}{4\nu^2} \\ &= \frac{\mathbb{E}_{p_0}[\ell^2(\boldsymbol{w})] - (\mathbb{E}_{p_0}[\ell(\boldsymbol{w})])^2}{4\nu^2}.\end{aligned}$$

The second equality follows by a similar substitution.

Setting $\mathrm{d}q_{\boldsymbol{w}}$ as per Corollary C.2, we get

$$
\begin{aligned}
R(\boldsymbol{w}; p_0) &= \mathbb{E}_{p_0}\left[\left(\frac{\mathrm{d}q_{\boldsymbol{w}}}{\mathrm{d}p_0}\right)\ell(\boldsymbol{w})\right] - \nu\chi^2(q_{\boldsymbol{w}}, p_0) \\
&= \mathbb{E}_{p_0}\left[\ell(\boldsymbol{w})\left(1 + \frac{\ell(\boldsymbol{w}) - \mathbb{E}_{p_0}\ell(\boldsymbol{w})}{2\nu}\right)\right] - \frac{\mathbb{E}_{p_0}[\ell^2(\boldsymbol{w})] - (\mathbb{E}_{p_0}[\ell(\boldsymbol{w})])^2}{4\nu} \\
&= \mathbb{E}_{p_0}[\ell(\boldsymbol{w})] + \frac{\mathbb{E}_{p_0}[\ell(\boldsymbol{w})^2] - (\mathbb{E}_{p_0}[\ell(\boldsymbol{w})])^2}{2\nu} - \frac{\mathbb{E}_{p_0}[\ell^2(\boldsymbol{w})] - (\mathbb{E}_{p_0}[\ell(\boldsymbol{w})])^2}{4\nu} \\
&= \mathbb{E}_{p_0}[\ell(\boldsymbol{w})] + \frac{\mathbb{E}_{p_0}[\ell^2(\boldsymbol{w})] - (\mathbb{E}_{p_0}[\ell(\boldsymbol{w})])^2}{4\nu}.
\end{aligned}
$$

□

Finally, an important consequence of Lemma C.1 is that it is possible to efficiently compute the risk of a given vector $\boldsymbol{w}$ with respect to $\widehat{p}_0$. We use this to compare the risk of our final output with the risk that is achieved by the zero vector.

C.2 Concentration

The expression we get in Corollary C.2 for p^* in terms of p_0 allows us to translate concentration properties of p_0 to p^*. We first state and prove a helper lemma, Lemma C.4, that shows $\mathbb{E}_{p_0}\ell(\boldsymbol{w}^*) \approx \mathbb{E}_{\widehat{p}_0}\ell(\boldsymbol{w}^*)$. Note that this is for the reference distribution p_0, and not the target distribution p^*.

For ease of notation, we define $U := 2\beta^2W^2(S^2 + C_M^2B^2\log^2(WB\beta/\epsilon))$, which is the upper bound for the loss value in Equation (20) throughout this section.

Lemma C.4. *Suppose p_0 satisfies Assumption 2.3. Then for any fixed $\boldsymbol{w} \in \mathcal{B}(W)$ and all $t > 0$, it holds that*

$$|\mathbb{E}_{p_0}\ell(\boldsymbol{w}) - \mathbb{E}_{\widehat{p}_0(N)}\ell(\boldsymbol{w})| \le t$$

with probability at least $1 - 2\exp\Big(\frac{-t^2N}{8(\beta^2W^2(S^2+C_M^2B^2\log^2(WB\beta/\epsilon)))^2}\Big)$. In particular, the above inequality holds for $\boldsymbol{w}^*$.

Proof. By Equation (20), $\forall i \in [N]$, $0 \le (\sigma(\boldsymbol{w}\cdot\boldsymbol{x}_i) - y_i)^2 \le U$. Hoeffding's inequality (Fact B.2) now implies that, for all $t > 0$,

$$\Pr\Big[\sum_{i=1}^{N}\frac{1}{N}\ell(\boldsymbol{w}^*; \boldsymbol{x}_i, y_i) - \mathbb{E}_{p_0}\ell(\boldsymbol{w}^*) \ge t\Big] \le 2\exp\Big(\frac{-t^2N}{2U^2}\Big).$$

Rearranging and plugging the definition of U, we get the lemma. □

We now use Lemma C.4 show that bounded Lipschitz functions concentrate with respect to p^*.

Lemma C.5. *Fix $\zeta > 0$. Let $h = h(\boldsymbol{z}; \boldsymbol{x}, y) : \mathcal{B}(\zeta)\times\mathbb{R}^d\times\mathbb{R} \to \mathbb{R}$ be a measurable function with respect to $\boldsymbol{x}, y$ that satisfies the condition that $|h(\boldsymbol{z};\cdot;\cdot)| \le b$ almost surely. Then, for $N = O_{B,S,\beta}\Big(\frac{b^2}{t^2}\Big(1 + \frac{W^4\log^4(W/\epsilon)}{\nu^2}\Big)\log(1/\delta)\Big)$ samples drawn from the reference distribution p_0 to construct $\widehat{p}_0(N)$, for any fixed $\boldsymbol{z} \in \mathcal{B}(\zeta)$, with probability at least $1 - 4\delta$, it holds that*

$$|\mathbb{E}_{(\boldsymbol{x},y)\sim\widehat{p}^*}[h(\boldsymbol{z};\boldsymbol{x},y)] - \mathbb{E}_{(\boldsymbol{x},y)\sim p^*}[h(\boldsymbol{z};\boldsymbol{x},y)]| \le t.$$

Moreover, suppose $\boldsymbol{z} \mapsto h(\boldsymbol{z};\boldsymbol{x},y)$ is a-Lipschitz. Then, for

$$N = O_{B,S,\beta}\Big(\frac{b^2}{t^2}\Big(1 + \frac{W^4\log^4(W/\epsilon)}{\nu^2}\Big)(d\log(\zeta a/t) + \log(1/\delta))\Big)$$

with probability at least $1 - 4\delta$, it holds that for all $\boldsymbol{z} \in \mathcal{B}(\zeta)$,

$$|\mathbb{E}_{(\boldsymbol{x},y)\sim\widehat{p}^*}[h(\boldsymbol{z};\boldsymbol{x},y)] - \mathbb{E}_{(\boldsymbol{x},y)\sim p^*}[h(\boldsymbol{z};\boldsymbol{x},y)]| \le t.$$

Proof. We use Lemma C.1 to change the distribution with respect to which we are taking the expectation,

$$\mathbb{E}_{\widehat{p}^*}[h(\boldsymbol{z})] = \mathbb{E}_{\widehat{p}_0}\Big[h(\boldsymbol{z};\boldsymbol{x},y)\frac{\ell(\boldsymbol{w}^*;\boldsymbol{x},y) - \mathbb{E}_{\widehat{p}_0}\ell(\boldsymbol{w}^*;\boldsymbol{x},y) + 2\nu}{2\nu}\Big].$$

Lemma C.4 now implies that with probability $1 - 2\exp(-2Nt^2/(bU/2\nu)^2)$,

$$\mathbb{E}_{\widehat{p}^*}[h(\boldsymbol{z})] = \mathbb{E}_{\widehat{p}_0}\Big[h(\boldsymbol{z};\boldsymbol{x},y)\frac{\ell(\boldsymbol{w}^*;\boldsymbol{x},y) - \mathbb{E}_{p_0}\ell(\boldsymbol{w}^*;\boldsymbol{x},y) + 2\nu}{2\nu}\Big] \pm \frac{t}{4}.$$

We now show that the expectation on the right hand side concentrates. To this end, we will use Hoeffding's inequality (Fact B.2). To apply this, we will need a bound on the quantity in the expectation. We bound this via an application of Equation (20) and the fact that $|h| \leq b$ to get,

$$\Big|h(\boldsymbol{z};\boldsymbol{x},y)\frac{\ell(\boldsymbol{w}^*) - \mathbb{E}_{p_0}\ell(\boldsymbol{w}^*) + 2\nu}{2\nu}\Big| \leq b\left(1 + \frac{U}{2\nu}\right).$$

This means, with probability at least $1 - 2\exp(-2t^2N/(b^2(1 + U/2\nu)^2))$,

$$\Big|\mathbb{E}_{p_0}\Big[h(\boldsymbol{z})\frac{\ell(\boldsymbol{w}^*) - \mathbb{E}_{p_0}\ell(\boldsymbol{w}^*) + 2\nu}{2\nu}\Big] - E_{\widehat{p}_0}\Big[h(\boldsymbol{z})\frac{\ell(\boldsymbol{w}^*) - \mathbb{E}_{p_0}\ell(\boldsymbol{w}^*) + 2\nu}{2\nu}\Big]\Big| \leq \frac{t}{2}. \tag{23}$$

Since $\boldsymbol{w} \mapsto h(\boldsymbol{w})$ is a-Lipschitz, a standard net argument over $\exp(O(d\log(\zeta a/t)))$ vectors yields: with probability at least $1 - 2\exp(O(d\log(\zeta a/t) - t^2N/(b^2(1 + U/2\nu)^2)))$, it holds that for all $\boldsymbol{z} \in \mathscr{B}(\zeta)$,

$$\Big|\mathbb{E}_{p_0}\Big[h(\boldsymbol{z})\frac{\ell(\boldsymbol{w}^*) - \mathbb{E}_{p_0}\ell(\boldsymbol{w}^*) + 2\nu}{2\nu}\Big] - E_{\widehat{p}_0}\Big[h(\boldsymbol{z})\frac{\ell(\boldsymbol{w}^*) - \mathbb{E}_{p_0}\ell(\boldsymbol{w}^*) + 2\nu}{2\nu}\Big]\Big| \leq t. \tag{24}$$

Putting things together, we see that if we choose

$$N = \Omega\Big(\frac{b^2}{t^2}\Big(1 + \frac{U^2}{\nu^2}\Big)(d\log(\zeta a/t) + \log(1/\delta))\Big),$$

with probability at least $1 - 4\delta$, for all $\boldsymbol{z} \in \mathscr{B}(\zeta)$,

$$|\mathbb{E}_{\widehat{p}^*}[h(\boldsymbol{z})] - \mathbb{E}_{p^*}[h(\boldsymbol{z})]| \leq t.$$

□

C.3 Sharpness and Optimal Loss Value

Finally, as a consequence of Lemma C.5, we can derive that $\widehat{p}^*$ satisfies sharpness, $\widehat{\mathrm{OPT}} \approx \mathrm{OPT}$ and $\widehat{\mathrm{OPT}}_{(2)} \approx \mathrm{OPT}_{(2)}$.

Lemma C.6 (Shaprness for $\widehat{p}^*$)**.** *Suppose Assumptions 2.1 to 2.3 are satisfied, then for large enough N:*

$$N = \widetilde{O}_{B,S,\beta,\alpha,\gamma,\lambda}\Big(\frac{W^4}{\epsilon^2}\Big(1 + \frac{W^4\log^4(1/\epsilon)}{\nu^2}\Big)(d + \log(1/\delta))\Big),$$

with probability at least $1 - 4\delta$, for all $\boldsymbol{w} \in \mathscr{B}(2\|\boldsymbol{w}^*\|)$ with $\|\boldsymbol{w} - \boldsymbol{w}^*\| \geq \sqrt{\epsilon}$ and $\boldsymbol{u} \in \mathscr{B}(1)$,

$$\mathbb{E}_{\boldsymbol{x}\sim\widehat{p}^*_{\boldsymbol{x}}}[(\sigma(\boldsymbol{w}\cdot\boldsymbol{x}) - \sigma(\boldsymbol{w}^*\cdot\boldsymbol{x}))(\boldsymbol{w}\cdot\boldsymbol{x} - \boldsymbol{w}^*\cdot\boldsymbol{x})] \geq (c_0/2)\|\boldsymbol{w} - \boldsymbol{w}^*\|_2^2 \tag{25}$$

$$\mathbb{E}_{\boldsymbol{x}\sim\widehat{p}^*_{\boldsymbol{x}}}[(\boldsymbol{x}\cdot\boldsymbol{u})^\tau] \leq 6B \quad \textit{for } \tau = 2, 4. \tag{26}$$

Proof. Fact 2.4 shows that p^* the conditions above (with different constants). We need to translate these to $\widehat{p}^*$. For each of the inequalities above, we will do this via an application of Lemma C.5.

Proof of Equation (25)*:* Set h in Lemma C.5 to be $h(\boldsymbol{w};\boldsymbol{x},y) := (\sigma(\boldsymbol{w}\cdot\boldsymbol{x}) - \sigma(\boldsymbol{w}^*\cdot\boldsymbol{x}))(\boldsymbol{w}\cdot\boldsymbol{x} - \boldsymbol{w}^*\cdot\boldsymbol{x})$. We proceed to set the constants a and b that used in Lemma C.5. Equation (18) implies that for all $(\boldsymbol{x}, y)$ in the support of $\widehat{p}_0$, $|h(\boldsymbol{w};\boldsymbol{x},y)| \leq 4\beta W^2S^2 =: b$. Also, $\boldsymbol{w} \mapsto h(\boldsymbol{w})$ is $a := 2WS^2(\beta + 1)\sqrt{d}$-Lipschitz as a consequence of Equations (18) and (19).

Lemma C.5 now gives us that for $N = \widetilde{O}_{B,S,\beta}\Big(\frac{W^4}{t^2}\Big(1 + \frac{W^4 \log^4(1/\epsilon)}{\nu^2}\Big)(d + \log(1/\delta))\Big)$ with probability at least $1 - 4\delta$, for all $\boldsymbol{w} \in \mathcal{B}(2\|\boldsymbol{w}^*\|)$,

$$\mathbb{E}_{\boldsymbol{x}\sim\widehat{p}^*_{\boldsymbol{x}}}[(\sigma(\boldsymbol{w}\cdot\boldsymbol{x}) - \sigma(\boldsymbol{w}^*\cdot\boldsymbol{x}))(\boldsymbol{w}\cdot\boldsymbol{x} - \boldsymbol{w}^*\cdot\boldsymbol{x})] \geq c_0\|\boldsymbol{w} - \boldsymbol{w}^*\|_2^2 - t.$$

Using the fact that $\|\boldsymbol{w} - \boldsymbol{w}^*\|_2 \geq \sqrt{\epsilon}$, we set $t = c_0\epsilon/2$, giving us the sample complexity

$$N = \widetilde{O}_{B,S,\beta,\alpha,\gamma,\lambda}\Big(\frac{W^4}{\epsilon^2}\Big(1 + \frac{W^4 \log^4(1/\epsilon)}{\nu^2}\Big)(d + \log(1/\delta))\Big). \tag{27}$$

Proof of Equation (26)*:* This follows analogously to the proof above. Set $h(\boldsymbol{u}; \boldsymbol{x}, y) = (\boldsymbol{x}\cdot\boldsymbol{u})^\tau$ in Lemma C.5 for $\tau = 2, 4$ and we proceed to calculate constants a, b. By Equations (18) and (19), it holds that $h(\boldsymbol{u}) \leq S^4 =: b$ and $\boldsymbol{u} \mapsto h(\boldsymbol{u})$ is $a := 4S^4\sqrt{d}$-Lipschitz. Setting $t = B$, by Lemma C.5, for $N = \widetilde{O}_{B,S,\beta}\Big(\Big(1 + \frac{W^4 \log^4(W/\epsilon)}{\nu^2}\Big)(d + \log(1/\delta))\Big)$ the conclusion follows. Note that this is dominated by Equation (27). □

We now show that $\mathrm{OPT} \approx \widehat{\mathrm{OPT}}$.

Lemma C.7. *Suppose Assumptions 2.1 to 2.3 are satisfied and the sample size N is large enough and* $N = \widetilde{O}_{B,S,\beta}\Big(\frac{W^4 \log^4(1/\epsilon)}{t^2}\Big(1 + \frac{W^4 \log^4(1/\epsilon)}{\nu^2}\Big)\log(1/\delta)\Big)$*. Then for any fixed* $\boldsymbol{w} \in \mathcal{B}(W)$ *and all* $t > 0$*, it holds that*

$$|\mathbb{E}_{p^*}\ell(\boldsymbol{w}; \boldsymbol{x}, y) - \mathbb{E}_{\widehat{p}^*}\ell(\boldsymbol{w}; \boldsymbol{x}, y)| \leq t$$

with probability at least $1 - 4\delta$*. In particular, the above inequality holds for* $\boldsymbol{w}^*$*, i.e.,* $|\mathrm{OPT} - \widehat{\mathrm{OPT}}| \leq t$*.*

Proof. Set $b := \beta^2 W^2(2S^2 + 2C_M^2 B^2 \log^2(WB\beta/\epsilon) \geq \|\ell(\boldsymbol{w})\|_2$ in Lemma C.5, where the inequality is a consequence of Lemma B.3. Then, with $N = \widetilde{O}_{B,S,\beta}\Big(\frac{W^4 \log^4(1/\epsilon)}{t^2}\Big(1 + \frac{W^4 \log^4(1/\epsilon)}{\nu^2}\Big)\log(1/\delta)\Big)$ samples, with probability $1 - 4\delta$, $|\mathbb{E}_{p^*}\ell(\boldsymbol{w}; \boldsymbol{x}, y) - \mathbb{E}_{\widehat{p}^*}\ell(\boldsymbol{w}; \boldsymbol{x}, y)| \leq t$. □

Finally, we show $\mathrm{OPT}_{(2)} \approx \widehat{\mathrm{OPT}}_{(2)}$.

Lemma C.8. *Suppose Assumptions 2.1 to 2.3 are satisfied and the sample size N is large enough and* $N = \widetilde{O}_{B,S,\beta}\Big(\frac{W^8 \log^8(1/\epsilon)}{t^2}\Big(1 + \frac{W^4 \log^4(1/\epsilon)}{\nu^2}\Big)\log(1/\delta)\Big)$*. Then for any fixed* $\boldsymbol{w} \in \mathcal{B}(W)$ *and all* $t > 0$*, it holds that*

$$|\mathbb{E}_{p^*}\ell^2(\boldsymbol{w}; \boldsymbol{x}, y) - \mathbb{E}_{\widehat{p}^*}\ell^2(\boldsymbol{w}; \boldsymbol{x}, y)| \leq t$$

with probability at least $1 - 4\delta$*. In particular, the above inequality holds for* $\boldsymbol{w}^*$*, i.e.,* $|\mathrm{OPT}_{(2)} - \widehat{\mathrm{OPT}}_{(2)}| \leq t$*.*

Proof. Analogous to the previous proof, observe that $\|\ell^2(\boldsymbol{w})\|_2 \leq 8\beta^4 W^4(S^4 + C_M^4 B^4 \log^4(WB\beta/\epsilon) =: b$. By Lemma C.5, with $N = \widetilde{O}_{B,S,\beta}\Big(\frac{W^8 \log^8(1/\epsilon)}{t^2}\Big(1 + \frac{W^4 \log^4(1/\epsilon)}{\nu^2}\Big)\log(1/\delta)\Big)$, with probability $1 - 4\delta$, it holds that $|\mathbb{E}_{p^*}\ell^2(\boldsymbol{w}; \boldsymbol{x}, y) - \mathbb{E}_{\widehat{p}^*}\ell^2(\boldsymbol{w}; \boldsymbol{x}, y)| \leq t$. □

We capture properties of OPT and $\mathrm{OPT}_{(2)}$ in the following corollary.

Corollary C.9 (Properties of $\mathrm{OPT}, \mathrm{OPT}_{(2)}$)**.** *Suppose Assumptions 2.1 to 2.3 are satisfied and the sample size N is large enough and* $N = \widetilde{O}_{B,S,\beta}\Big(\frac{W^8 \log^8(1/\epsilon)}{t^2}\Big(1 + \frac{W^4 \log^4(1/\epsilon)}{\nu^2}\Big)\log(1/\delta)\Big)$*. Then for all* $t > 0$*, the following hold:*

1. *With probability* $1 - 4\delta$*,* $|\mathrm{OPT}_{(2)} - \widehat{\mathrm{OPT}}_{(2)}| \leq t$*.*
2. *With probability* $1 - 4\delta$*,* $|\mathrm{OPT} - \widehat{\mathrm{OPT}}| \leq t$*.*

3. $\mathrm{OPT} \leq \sqrt{\mathrm{OPT}_{(2)}}$ *and* $\widehat{\mathrm{OPT}} \leq \sqrt{\widehat{\mathrm{OPT}}_{(2)}}$.

Proof. The first two items follow immediately from Lemma C.8 and Lemma C.7. The third item is a consequence of Cauchy-Schwarz. □

D Gap Upper Bound

To prove Lemma 3.3, we need to construct an upper bound on the gap $\mathrm{Gap}(\boldsymbol{w}_i, \widehat{p}_i) = L(\boldsymbol{w}_i, \widehat{p}^*) - L(\boldsymbol{w}^*, \widehat{p}_i)$. To achieve this, we establish an upper bound on $L(\boldsymbol{w}_i, \widehat{p}^*)$, which motivates the update rule for $\widehat{p}_i$. We also establish a lower bound on $L(\boldsymbol{w}^*, \widehat{p}_i)$, which guides the update rule for $\boldsymbol{w}_i$ and the construction of $\boldsymbol{g}_i$. Note that the construction of the lower bound is more challenging here, due to the nonconvexity of the square loss. This is where most of the (non-standard) technical work happens. To simplify the notation, we use $\phi(\widehat{p}) := \chi^2(\widehat{p}, \widehat{p}_0)$ throughout this section.

Upper bound on $L(\boldsymbol{w}_i, \widehat{p}^*)$. We begin the analysis with the construction of the upper bound, which is used for defining the dual updates. Most of this construction follows a similar argument as used in other primal-dual methods such as [Dia+22c; SWD21].

Lemma D.1 (Upper Bound on $a_i L(\boldsymbol{w}_i, \widehat{p}^*)$)**.** *Let $\widehat{p}_i$ evolve as outlined in Line 7. Then, for all $i \geq 1$,*

$$a_i L(\boldsymbol{w}_i, \widehat{p}^*) \leq a_i L(\boldsymbol{w}_i, \widehat{p}_i) + (\nu_0 + \nu A_{i-1}) D_\phi(\widehat{p}^*, \widehat{p}_{i-1}) - (\nu_0 + \nu A_i) D_\phi(\widehat{p}^*, \widehat{p}_i)$$
$$- (\nu_0 + \nu A_{i-1}) D_\phi(\widehat{p}_i, \widehat{p}_{i-1}).$$

Proof. Recall that $\phi(\widehat{p}) := \chi^2(\widehat{p}, \widehat{p}_0)$. Observe that $L(\boldsymbol{w}_i, \widehat{p}^*)$ as a function of $\widehat{p}^*$ is linear minus the nonlinearity $\nu\phi$. We could directly maximize this function and define $\widehat{p}_i$ correspondingly (which would lead to a valid upper bound); however, such an approach appears insufficient for obtaining our results. Instead, adding and subtracting $(\nu_0 + \nu A_{i-1}) D_\phi(\widehat{p}^*, \widehat{p}_{i-1})$, we have

$$\begin{aligned} a_i L(\boldsymbol{w}_i, \widehat{p}^*) &= a_i L(\boldsymbol{w}_i, \widehat{p}^*) - (\nu_0 + \nu A_{i-1}) D_\phi(\widehat{p}^*, \widehat{p}_{i-1}) + (\nu_0 + \nu A_{i-1}) D_\phi(\widehat{p}^*, \widehat{p}_{i-1}) \\ &= h(\widehat{p}^*) + (\nu_0 + \nu A_{i-1}) D_\phi(\widehat{p}^*, \widehat{p}_{i-1}), \end{aligned} \tag{28}$$

where we define, for notational convenience:

$$h(\widehat{p}) := a_i L(\boldsymbol{w}_i, \widehat{p}) - (\nu_0 + \nu A_{i-1}) D_\phi(\widehat{p}, \widehat{p}_{i-1}).$$

Observe that by the definition of $\widehat{p}_i$, $h(\widehat{p})$ is maximized by $\widehat{p}_i$. Hence, using the definition of a Bregman divergence, we have that

$$\begin{aligned} h(\widehat{p}^*) &= h(\widehat{p}_i) + \langle \nabla h(\widehat{p}_i), \widehat{p}^* - \widehat{p}_i \rangle + D_h(\widehat{p}^*, \widehat{p}_i) \\ &\leq h(\widehat{p}_i) - (\nu_0 + \nu A_i) D_\phi(\widehat{p}^*, \widehat{p}_i), \end{aligned}$$

where in the inequality we used that $\langle \nabla h(\widehat{p}_i), \widehat{p}^* - \widehat{p}_i \rangle \leq 0$ (as $\widehat{p}_i$ maximizes h) and $D_h(\widehat{p}^*, \widehat{p}_i) = -(\nu_0 + \nu A_i) D_\phi(\widehat{p}^*, \widehat{p}_i)$ (as $h(\widehat{p})$ can be expressed as $-(\nu_0 + \nu A_i)\phi(\widehat{p})$ plus terms that are either linear in $\widehat{p}$ or independent of it. See Fact 2.7). Combining with Equation (28) and the definition of h and simplifying, the claimed bound follows. □

An important feature of Lemma D.1 is that the first two Bregman divergence terms usefully telescope, while the last one is negative and can be used in controlling the error terms arising from the algorithmic choices.

Lower bound on $L(\boldsymbol{w}^*, \widehat{p}_i)$. The more technical part of our analysis concerns the construction of a lower bound on $L(\boldsymbol{w}^*, \widehat{p}_i)$, which leads to update rule for $\boldsymbol{w}^*$. In standard, Chambolle-Pock-style primal-dual algorithms [ACW22; CP11; SWD21], where the coupling $L(\boldsymbol{w}, \widehat{p})$ between the primal and the dual is *bilinear*, the lower bound would be constructed using an analogue of the upper bound, with a small difference to correct for the fact that $\boldsymbol{w}_i$ is updated before $\widehat{p}_i$ and so one cannot use information about $\widehat{p}_i$ in the $\boldsymbol{w}_i$ update. This is done using an extrapolation idea, which replaces $\widehat{p}_i$ with an extrapolated value from prior two iterations and controls for the introduced error.

In our case, however, the coupling is not only nonlinear, but also *nonconvex* because $\ell(\boldsymbol{w}; \boldsymbol{x}, y) = (\sigma(\boldsymbol{w} \cdot \boldsymbol{x}) - y)^2$ is nonconvex. Nonlinearity is an issue because if we were to follow an analogue of the construction from Lemma D.1, we would need to assume that we can efficiently minimize over $\boldsymbol{w}$ the sum of $L(\boldsymbol{w}, \widehat{p})$ and a convex function (e.g., a quadratic), which translates into proximal point updates for the L_2^2 loss for which efficient computation is generally unclear. Nonlinearity alone (but assuming convexity) has been handled in the very recent prior work [MDH24], where this issue is addressed using convexity of the nonlinear function to bound it below by its linear approximation around $\boldsymbol{w}_i$. Unfortunately, as mentioned before, this approach cannot apply here as we do not have convexity. Instead, we use a rather intricate argument that relies on monotonicity and Lipschitzness properties of the activation σ and structural properties of the problem which only hold with respect to the target distribution $\widehat{p}^*$ (and the empirical target distribution p^*, due to our results from Lemma 2.5. Handling these issues related to nonconvexity of the loss in the construction of the upper bound is precisely what forces us to choose chi-square as the measure of divergence between distributions; see Lemma D.3 and the discussion therein.

Proposition D.2. *Consider the sequence $\{\boldsymbol{w}_i\}_i$ evolving as per Line 6. Under the setting in which Lemma 2.5 holds, we have for all $i \geq 1$,*

$$\begin{aligned}a_i L(\boldsymbol{w}^*, \widehat{p}_i) \geq\; & L(\boldsymbol{w}_i, \widehat{p}_i) - a_i E_i - (\nu_0 + \nu A_{i-2}) D_\phi(\widehat{p}_{i-1}, \widehat{p}_{i-2}) \\ & + \frac{1 + 0.5c_1 A_{i-1}}{2}\|\boldsymbol{w}^* - \boldsymbol{w}_i\|_2^2 - \frac{1 + 0.5c_1 A_{i-1}}{2}\|\boldsymbol{w}^* - \boldsymbol{w}_{i-1}\|_2^2 \\ & + \frac{1 + 0.5c_1 A_{i-1}}{4}\|\boldsymbol{w}_i - \boldsymbol{w}_{i-1}\|_2^2 - \frac{1 + 0.5c_1 A_{i-2}}{4}\|\boldsymbol{w}_{i-1} - \boldsymbol{w}_{i-2}\|_2^2 \\ & + a_i \langle \mathbb{E}_{\widehat{p}_i}[\boldsymbol{v}(\boldsymbol{w}_i; \boldsymbol{x}, y)] - \mathbb{E}_{\widehat{p}_{i-1}}[\boldsymbol{v}(\boldsymbol{w}_{i-1}; \boldsymbol{x}, y)], \boldsymbol{w}^* - \boldsymbol{w}_i\rangle \\ & - a_{i-1}\langle \mathbb{E}_{\widehat{p}_{i-1}}[\boldsymbol{v}(\boldsymbol{w}_{i-1}; \boldsymbol{x}, y)] - \mathbb{E}_{\widehat{p}_{i-2}}[\boldsymbol{v}(\boldsymbol{w}_{i-2}; \boldsymbol{x}, y)], \boldsymbol{w}^* - \boldsymbol{w}_{i-1}\rangle,\end{aligned}$$

where is E_i is defined by Equation (37).

Proof. From the definition of $L(\boldsymbol{w}^*, \widehat{p}_i)$, we have:

$$L(\boldsymbol{w}^*, \widehat{p}_i) = \mathbb{E}_{\widehat{p}_i}[(\sigma(\boldsymbol{w}^* \cdot \boldsymbol{x}) - y)^2] - \nu D(\widehat{p}_i, \widehat{p}_0).$$

Writing $(\sigma(\boldsymbol{w}^* \cdot \boldsymbol{x}) - y)^2 = ((\sigma(\boldsymbol{w}^* \cdot \boldsymbol{x}) - \sigma(\boldsymbol{w}_i \cdot \boldsymbol{x})) + (\sigma(\boldsymbol{w}_i \cdot \boldsymbol{x}) - y))^2$ and expanding the square, we have

$$\begin{aligned}L(\boldsymbol{w}^*, \widehat{p}_i) &= \mathbb{E}_{\widehat{p}_i}[(\sigma(\boldsymbol{w}_i \cdot \boldsymbol{x}) - y)^2] - \nu D(\widehat{p}_i, \widehat{p}_0) + \mathbb{E}_{\widehat{p}_i}[(\sigma(\boldsymbol{w}^* \cdot \boldsymbol{x}) - \sigma(\boldsymbol{w}_i \cdot \boldsymbol{x}))^2] \\ &\quad + \mathbb{E}_{\widehat{p}_i}[2(\sigma(\boldsymbol{w}_i \cdot \boldsymbol{x}) - y)(\sigma(\boldsymbol{w}^* \cdot \boldsymbol{x}) - \sigma(\boldsymbol{w}_i \cdot \boldsymbol{x}))] \\ &= L(\boldsymbol{w}_i, \widehat{p}_i) + S_i, \end{aligned} \tag{29}$$

where for notational convenience we define

$$S_i := \mathbb{E}_{\widehat{p}_i}[(\sigma(\boldsymbol{w}^* \cdot \boldsymbol{x}) - \sigma(\boldsymbol{w}_i \cdot \boldsymbol{x}))^2] + \mathbb{E}_{\widehat{p}_i}[2(\sigma(\boldsymbol{w}_i \cdot \boldsymbol{x}) - y)(\sigma(\boldsymbol{w}^* \cdot \boldsymbol{x}) - \sigma(\boldsymbol{w}_i \cdot \boldsymbol{x}))]. \tag{30}$$

Observe that $L(\boldsymbol{w}_i, \widehat{p}_i)$ on the right-hand side also appears in the upper bound on $L(\boldsymbol{w}_i, \widehat{p}^*)$ in Lemma D.1 and so it will get cancelled out when $L(\boldsymbol{w}^*, \widehat{p}_i)$ is subtracted from $L(\boldsymbol{w}_i, \widehat{p}^*)$ in the gap computation. Thus, we only need to focus on bounding S_i. This requires a rather technical argument, which we defer to Lemma D.3 below. Instead, we call on Lemma D.3 to state that

$$S_i \geq \mathbb{E}_{\widehat{p}_i}[\langle \boldsymbol{v}(\boldsymbol{w}_i; \boldsymbol{x}, y), \boldsymbol{w}^* - \boldsymbol{w}_i\rangle] - E_i \tag{31}$$

and carry out the rest of the proof under this assumption (which is proved in Lemma D.3).

At this point, we have obtained a "linearization" that was needed to continue by mimicking the construction of the upper bound. However, $\boldsymbol{v}(\boldsymbol{w}_i; \boldsymbol{x}, y)$ depends on $\boldsymbol{w}_i$, and so trying to define $\boldsymbol{w}_i$ based on this quantity would lead to an implicitly defined update, which is generally not efficiently computable. Instead, here we use the idea of extrapolation: instead of defining a step w.r.t. $\boldsymbol{w}_i$, we replace $\mathbb{E}_{\widehat{p}_i}[\boldsymbol{v}(\boldsymbol{w}_i; \boldsymbol{x}, y)]$ by an "extrapolated gradient" defined by (cf. Line 5 in Algorithm 1):

$$\boldsymbol{g}_{i-1} = \mathbb{E}_{\widehat{p}_{i-1}}[\boldsymbol{v}(\boldsymbol{w}_{i-1}; \boldsymbol{x}, y)] + \frac{a_{i-1}}{a_i}(\mathbb{E}_{\widehat{p}_{i-1}}[\boldsymbol{v}(\boldsymbol{w}_{i-1}; \boldsymbol{x}, y)] - \mathbb{E}_{\widehat{p}_{i-2}}[\boldsymbol{v}(\boldsymbol{w}_{i-2}; \boldsymbol{x}, y)]).$$

Combining with the bound on S_i from Equation (31) and simplifying, we now have

$$\begin{aligned}a_i S_i \geq\; & a_i \langle \boldsymbol{g}_{i-1}, \boldsymbol{w}^* - \boldsymbol{w}_i\rangle - a_i E_i \\ & + a_i \langle \mathbb{E}_{\widehat{p}_i}[\boldsymbol{v}(\boldsymbol{w}_i; \boldsymbol{x}, y) - \boldsymbol{g}_{i-1}], \boldsymbol{w}^* - \boldsymbol{w}_i\rangle.\end{aligned} \tag{32}$$

Let $\psi(\boldsymbol{w}) = a_i \langle \boldsymbol{g}_{i-1}, \boldsymbol{w} \rangle + \frac{1+0.5c_1 A_{i-1}}{2} \|\boldsymbol{w} - \boldsymbol{w}_{i-1}\|_2^2$ and observe that (by Line 6 in Algorithm 1) $\boldsymbol{w}_i = \arg\min_{\boldsymbol{w} \in \mathcal{B}(W)} \psi(\boldsymbol{w})$. Then, by a similar argument as in the proof of Lemma D.1, since ψ is minimized by $\boldsymbol{w}_i$ and is a quadratic function in $\boldsymbol{w}_i$, we have

$$\begin{aligned} a_i \langle \boldsymbol{g}_{i-1}, \boldsymbol{w}^* - \boldsymbol{w}_i \rangle \geq \; & \frac{1+0.5c_1 A_{i-1}}{2} \|\boldsymbol{w}^* - \boldsymbol{w}_i\|_2^2 - \frac{1+0.5c_1 A_{i-1}}{2} \|\boldsymbol{w}^* - \boldsymbol{w}_{i-1}\|_2^2 \\ & + \frac{1+0.5c_1 A_{i-1}}{2} \|\boldsymbol{w}_i - \boldsymbol{w}_{i-1}\|_2^2. \end{aligned} \tag{33}$$

On the other hand, by the definition of $\boldsymbol{g}_i$, we have

$$\begin{aligned} & a_i \langle \mathbb{E}_{\widehat{p}_i}[\boldsymbol{v}(\boldsymbol{w}_i; \boldsymbol{x}, y)] - \boldsymbol{g}_{i-1}, \boldsymbol{w}^* - \boldsymbol{w}_i \rangle \\ = \; & a_i \langle \mathbb{E}_{\widehat{p}_i}[\boldsymbol{v}(\boldsymbol{w}_i; \boldsymbol{x}, y)] - \mathbb{E}_{\widehat{p}_{i-1}}[\boldsymbol{v}(\boldsymbol{w}_{i-1}; \boldsymbol{x}, y)], \boldsymbol{w}^* - \boldsymbol{w}_i \rangle \\ & - a_{i-1} \langle \mathbb{E}_{\widehat{p}_{i-1}}[\boldsymbol{v}(\boldsymbol{w}_{i-1}; \boldsymbol{x}, y)] - \mathbb{E}_{\widehat{p}_{i-2}}[\boldsymbol{v}(\boldsymbol{w}_{i-2}; \boldsymbol{x}, y)], \boldsymbol{w}^* - \boldsymbol{w}_{i-1} \rangle \\ & + a_{i-1} \langle \mathbb{E}_{\widehat{p}_{i-1}}[\boldsymbol{v}(\boldsymbol{w}_{i-1}; \boldsymbol{x}, y)] - \mathbb{E}_{\widehat{p}_{i-2}}[\boldsymbol{v}(\boldsymbol{w}_{i-2}; \boldsymbol{x}, y)], \boldsymbol{w}_i - \boldsymbol{w}_{i-1} \rangle \end{aligned} \tag{34}$$

The first two terms on the right-hand side of Equation (34) telescope, so we focus on bounding the last term. We do so using Young's inequality (Fact B.1) followed by κ-Lipschitzness of $\boldsymbol{v}$, which leads to

$$\begin{aligned} & - a_{i-1} \langle \mathbb{E}_{\widehat{p}_{i-1}}[\boldsymbol{v}(\boldsymbol{w}_{i-1}; \boldsymbol{x}, y)] - \mathbb{E}_{\widehat{p}_{i-2}}[\boldsymbol{v}(\boldsymbol{w}_{i-2}; \boldsymbol{x}, y)], \boldsymbol{w}_i - \boldsymbol{w}_{i-1} \rangle \\ \leq \; & \frac{{a_{i-1}}^2}{1+0.5c_1 A_{i-1}} \|\mathbb{E}_{\widehat{p}_{i-1}}[\boldsymbol{v}(\boldsymbol{w}_{i-1}; \boldsymbol{x}, y)] - \mathbb{E}_{\widehat{p}_{i-2}}[\boldsymbol{v}(\boldsymbol{w}_{i-2}; \boldsymbol{x}, y)]\|_2^2 + \frac{1+0.5c_1 A_{i-1}}{4} \|\boldsymbol{w}_i - \boldsymbol{w}_{i-1}\|_2^2 \\ \overset{(i)}{\leq} \; & \frac{2{a_{i-1}}^2 \kappa^2}{1+0.5c_1 A_{i-1}} \|\boldsymbol{w}_{i-1} - \boldsymbol{w}_{i-2}\|_2^2 + \frac{2{a_{i-1}}^2 G^2}{1+0.5c_1 A_{i-1}} D_\phi(\widehat{p}_{i-1}, \widehat{p}_{i-2}) + \frac{1+0.5c_1 A_{i-1}}{4} \|\boldsymbol{w}_i - \boldsymbol{w}_{i-1}\|_2^2 \\ \overset{(ii)}{\leq} \; & \frac{1+0.5c_1 A_{i-2}}{4} \|\boldsymbol{w}_{i-1} - \boldsymbol{w}_{i-2}\|_2^2 + \frac{1+0.5c_1 A_{i-1}}{4} \|\boldsymbol{w}_i - \boldsymbol{w}_{i-1}\|_2^2 + (\nu_0 + \nu A_{i-2}) D_\phi(\widehat{p}_{i-1}, \widehat{p}_{i-2}), \end{aligned} \tag{35}$$

where in (ii) we used $\frac{2{a_{i-1}}^2 \kappa^2}{1+0.5c_1 A_{i-1}} \leq \frac{1+0.5c_1 A_{i-2}}{4}$ and $\frac{2{a_{i-1}}^2 G^2}{1+0.5c_1 A_{i-1}} \leq \nu_0 + \nu A_{i-2}$, which both hold by the choice of the step size, while (i) follows by boundedness and κ-Lipschitzness of $\boldsymbol{v}$ and Corollary B.5, using

$$\begin{aligned} & \|\mathbb{E}_{\widehat{p}_{i-1}}[\boldsymbol{v}(\boldsymbol{w}_{i-1}; \boldsymbol{x}, y)] - \mathbb{E}_{\widehat{p}_{i-2}}[\boldsymbol{v}(\boldsymbol{w}_{i-2}; \boldsymbol{x}, y)]\|_2^2 \\ = \; & \|\mathbb{E}_{\widehat{p}_{i-1}}[\boldsymbol{v}(\boldsymbol{w}_{i-1}; \boldsymbol{x}, y)] - \mathbb{E}_{\widehat{p}_{i-2}}[\boldsymbol{v}(\boldsymbol{w}_{i-1}; \boldsymbol{x}, y)] + \mathbb{E}_{\widehat{p}_{i-2}}[\boldsymbol{v}(\boldsymbol{w}_{i-1}; \boldsymbol{x}, y)] - \mathbb{E}_{\widehat{p}_{i-2}}[\boldsymbol{v}(\boldsymbol{w}_{i-2}; \boldsymbol{x}, y)]\|_2^2 \\ \leq \; & 2\|\mathbb{E}_{\widehat{p}_{i-1}}[\boldsymbol{v}(\boldsymbol{w}_{i-1}; \boldsymbol{x}, y)] - \mathbb{E}_{\widehat{p}_{i-2}}[\boldsymbol{v}(\boldsymbol{w}_{i-1}; \boldsymbol{x}, y)]\|_2^2 + 2\|\mathbb{E}_{\widehat{p}_{i-2}}[\boldsymbol{v}(\boldsymbol{w}_{i-1}; \boldsymbol{x}, y)] - \mathbb{E}_{\widehat{p}_{i-2}}[\boldsymbol{v}(\boldsymbol{w}_{i-2}; \boldsymbol{x}, y)]\|_2^2 \\ \leq \; & 2G^2 D_\phi(\widehat{p}_{i-1}, \widehat{p}_{i-2}) + 2\kappa^2 \|\boldsymbol{w}_{i-1} - \boldsymbol{w}_{i-2}\|_2^2. \end{aligned} \tag{36}$$

Combining Equations (32) to (35), we now have

$$\begin{aligned} a_i S_i \geq \; & - a_i E_i + \frac{1+0.5c_1 A_{i-1}}{2} \|\boldsymbol{w}^* - \boldsymbol{w}_i\|_2^2 - \frac{1+0.5c_1 A_{i-1}}{2} \|\boldsymbol{w}^* - \boldsymbol{w}_{i-1}\|_2^2 \\ & - (\nu_0 + \nu A_{i-2}) D_\phi(\widehat{p}_{i-1}, \widehat{p}_{i-2}) \\ & + \frac{1+0.5c_1 A_{i-1}}{4} \|\boldsymbol{w}_i - \boldsymbol{w}_{i-1}\|_2^2 - \frac{1+0.5c_1 A_{i-2}}{4} \|\boldsymbol{w}_{i-1} - \boldsymbol{w}_{i-2}\|_2^2 \\ & + a_i \langle \mathbb{E}_{\widehat{p}_i}[\boldsymbol{v}(\boldsymbol{w}_i; \boldsymbol{x}, y)] - \mathbb{E}_{\widehat{p}_i}[\boldsymbol{v}(\boldsymbol{w}_{i-1}; \boldsymbol{x}, y)], \boldsymbol{w}^* - \boldsymbol{w}_i \rangle \\ & - a_{i-1} \langle \mathbb{E}_{\widehat{p}_{i-1}}[\boldsymbol{v}(\boldsymbol{w}_{i-1}; \boldsymbol{x}, y)] - \mathbb{E}_{\widehat{p}_{i-1}}[\boldsymbol{v}(\boldsymbol{w}_{i-2}; \boldsymbol{x}, y)], \boldsymbol{w}^* - \boldsymbol{w}_{i-1} \rangle. \end{aligned}$$

To complete the proof, it remains to combine the last inequality with Equation (29). □

Lemma D.3. *Let S_i be defined by Equation* (30). *Then, under the setting of Proposition D.2, we have*

$$S_i \geq \mathbb{E}_{\widehat{p}_i}[\langle \boldsymbol{v}(\boldsymbol{w}; \boldsymbol{x}, y), \boldsymbol{w}^* - \boldsymbol{w}_i \rangle] - E_i,$$

where

$$E_i = \frac{c_1}{4} \|\boldsymbol{w}^* - \boldsymbol{w}_i\|_2^2 + \frac{8\beta^2 \sqrt{6B} \sqrt{\widehat{\mathrm{OPT}}_{(2)}}}{c_1} \chi^2(\widehat{p}_i, \widehat{p}^*) + \frac{48\beta^2 B \widehat{\mathrm{OPT}}}{c_1}. \tag{37}$$

Proof. Define the event $\mathscr{G} = \{(\boldsymbol{x}, y) : \sigma(\boldsymbol{w}_i \cdot \boldsymbol{x}) - y \geq 0\}$. Then,

$$\begin{aligned}
&\mathbb{E}_{\widehat{p}_i}[2(\sigma(\boldsymbol{w}_i \cdot \boldsymbol{x}) - y)(\sigma(\boldsymbol{w}^* \cdot \boldsymbol{x}) - \sigma(\boldsymbol{w}_i \cdot \boldsymbol{x}))] \\
&= \mathbb{E}_{\widehat{p}_i}[2(\sigma(\boldsymbol{w}_i \cdot \boldsymbol{x}) - y)(\sigma(\boldsymbol{w}^* \cdot \boldsymbol{x}) - \sigma(\boldsymbol{w}_i \cdot \boldsymbol{x}))\mathbb{I}_{\mathscr{G}} + 2(\sigma(\boldsymbol{w}_i \cdot \boldsymbol{x}) - y)(\sigma(\boldsymbol{w}^* \cdot \boldsymbol{x}) - \sigma(\boldsymbol{w}_i \cdot \boldsymbol{x}))\mathbb{I}_{\mathscr{G}^c}] \\
&\geq \mathbb{E}_{\widehat{p}_i}[\mathbb{I}_{\mathscr{G}} 2(\sigma(\boldsymbol{w}_i \cdot \boldsymbol{x}) - y)\sigma'(\boldsymbol{w}_i \cdot \boldsymbol{x})(\boldsymbol{w}^* \cdot \boldsymbol{x} - \boldsymbol{w}_i \cdot \boldsymbol{x})] \\
&\quad + \mathbb{E}_{\widehat{p}_i}[\mathbb{I}_{\mathscr{G}^c} 2(\sigma(\boldsymbol{w}_i \cdot \boldsymbol{x}) - y)\sigma'(\boldsymbol{w}^* \cdot \boldsymbol{x})(\boldsymbol{w}^* \cdot \boldsymbol{x} - \boldsymbol{w}_i \cdot \boldsymbol{x})],
\end{aligned}$$

where the last inequality uses convexity of $\sigma(\cdot)$ to bound the term that involves $\mathbb{I}_{\mathscr{G}}$ and concavity of $-\sigma(\cdot)$ to bound the term that involves $\mathbb{I}_{\mathscr{G}^c}$ and where σ' denotes any subderivative of σ (guaranteed to exist, due to convexity and Lipschitzness).

Recall that $\boldsymbol{v}(\boldsymbol{w}_i; \boldsymbol{x}, y) = 2\beta(\sigma(\boldsymbol{w}_i \cdot \boldsymbol{x}) - y)\boldsymbol{x}$. Using that $\sigma'(t) = \beta + (\sigma'(t) - \beta)$ for all t and combining with the inequality above, we see

$$\begin{aligned}
&\mathbb{E}_{\widehat{p}_i}[2(\sigma(\boldsymbol{w}_i \cdot \boldsymbol{x}) - y)(\sigma(\boldsymbol{w}^* \cdot \boldsymbol{x}) - \sigma(\boldsymbol{w}_i \cdot \boldsymbol{x}))] \\
&\geq \mathbb{E}_{\widehat{p}_i}[\langle \boldsymbol{v}(\boldsymbol{w}_i; \boldsymbol{x}, y), \boldsymbol{w}^* - \boldsymbol{w}_i \rangle] \\
&\quad + 2\mathbb{E}_{\widehat{p}_i}[\mathbb{I}_{\mathscr{G}}(\sigma(\boldsymbol{w}_i \cdot \boldsymbol{x}) - y)(\sigma'(\boldsymbol{w}_i \cdot \boldsymbol{x}) - \beta)(\boldsymbol{w}^* \cdot \boldsymbol{x} - \boldsymbol{w}_i \cdot \boldsymbol{x})] && (38) \\
&\quad + 2\mathbb{E}_{\widehat{p}_i}[\mathbb{I}_{\mathscr{G}^c}(\sigma(\boldsymbol{w}_i \cdot \boldsymbol{x}) - y)(\sigma'(\boldsymbol{w}^* \cdot \boldsymbol{x}) - \beta)(\boldsymbol{w}^* \cdot \boldsymbol{x} - \boldsymbol{w}_i \cdot \boldsymbol{x})], && (39)
\end{aligned}$$

and so to prove the lemma we only need to focus on bounding the terms in the last two lines.

Recall that σ is assumed to be monotonically increasing and β-Lipschitz, and so $0 \leq \sigma'(\boldsymbol{w}^* \cdot \boldsymbol{x}) \leq \beta$. Thus, we have

$$\begin{aligned}
&(\sigma(\boldsymbol{w}_i \cdot \boldsymbol{x}) - y)(\sigma'(\boldsymbol{w}^* \cdot \boldsymbol{x}) - \beta)(\boldsymbol{w}^* \cdot \boldsymbol{x} - \boldsymbol{w}_i \cdot \boldsymbol{x}) \\
&= (\sigma(\boldsymbol{w}^* \cdot \boldsymbol{x}) - y)(\sigma'(\boldsymbol{w}^* \cdot \boldsymbol{x}) - \beta)(\boldsymbol{w}^* \cdot \boldsymbol{x} - \boldsymbol{w}_i \cdot \boldsymbol{x}) \\
&\quad + (\sigma(\boldsymbol{w}_i \cdot \boldsymbol{x}) - \sigma(\boldsymbol{w}^* \cdot \boldsymbol{x}))(\sigma'(\boldsymbol{w}^* \cdot \boldsymbol{x}) - \beta)(\boldsymbol{w}^* \cdot \boldsymbol{x} - \boldsymbol{w}_i \cdot \boldsymbol{x}) \\
&\geq (\sigma(\boldsymbol{w}^* \cdot \boldsymbol{x}) - y)(\sigma'(\boldsymbol{w}^* \cdot \boldsymbol{x}) - \beta)(\boldsymbol{w}^* \cdot \boldsymbol{x} - \boldsymbol{w}_i \cdot \boldsymbol{x}), && (40)
\end{aligned}$$

where we have used $\sigma'(\boldsymbol{w}^* \cdot \boldsymbol{x}) - \beta \leq 0$ (by Lipschitzness) and $(\sigma(\boldsymbol{w}_i \cdot \boldsymbol{x}) - \sigma(\boldsymbol{w}^* \cdot \boldsymbol{x}))(\boldsymbol{w}^* \cdot \boldsymbol{x} - \boldsymbol{w}_i \cdot \boldsymbol{x}) \leq 0$ (by monotonicity of σ). By the same argument,

$$\begin{aligned}
&(\sigma(\boldsymbol{w}_i \cdot \boldsymbol{x}) - y)(\sigma'(\boldsymbol{w}_i \cdot \boldsymbol{x}) - \beta)(\boldsymbol{w}^* \cdot \boldsymbol{x} - \boldsymbol{w}_i \cdot \boldsymbol{x}) \\
&\geq (\sigma(\boldsymbol{w}^* \cdot \boldsymbol{x}) - y)(\sigma'(\boldsymbol{w}_i \cdot \boldsymbol{x}) - \beta)(\boldsymbol{w}^* \cdot \boldsymbol{x} - \boldsymbol{w}_i \cdot \boldsymbol{x}).
\end{aligned}$$

To complete the proof of the lemma, it remains to bound the expectation of the term in Equation (40). We proceed using that $|\sigma'(\boldsymbol{w} \cdot \boldsymbol{x}) - \beta| \leq \beta, \forall \boldsymbol{w}$, and thus for $\boldsymbol{w} \in \{\boldsymbol{w}^*, \boldsymbol{w}_i\}$:

$$\begin{aligned}
&|(\sigma(\boldsymbol{w}^* \cdot \boldsymbol{x}) - y)(\sigma'(\boldsymbol{w} \cdot \boldsymbol{x}) - \beta)(\boldsymbol{w}^* \cdot \boldsymbol{x} - \boldsymbol{w}_i \cdot \boldsymbol{x})| \\
&\leq \beta|\sigma(\boldsymbol{w}^* \cdot \boldsymbol{x}) - y||\boldsymbol{w}^* \cdot \boldsymbol{x} - \boldsymbol{w}_i \cdot \boldsymbol{x}|.
\end{aligned}$$

Taking the expectation on both sides, and combining with Equation (40), we further have

$$\begin{aligned}
&- 2\mathbb{E}_{\widehat{p}_i}[\mathbb{I}_{\mathscr{G}}(\sigma(\boldsymbol{w}_i \cdot \boldsymbol{x}) - y)(\sigma'(\boldsymbol{w}_i \cdot \boldsymbol{x}) - \beta)(\boldsymbol{w}^* \cdot \boldsymbol{x} - \boldsymbol{w}_i \cdot \boldsymbol{x})] \\
&- 2\mathbb{E}_{\widehat{p}_i}[\mathbb{I}_{\mathscr{G}^c}(\sigma(\boldsymbol{w}_i \cdot \boldsymbol{x}) - y)(\sigma'(\boldsymbol{w}^* \cdot \boldsymbol{x}) - \beta)(\boldsymbol{w}^* \cdot \boldsymbol{x} - \boldsymbol{w}_i \cdot \boldsymbol{x})] \\
&\leq 2\beta\, \mathbb{E}_{\widehat{p}_i}[(\mathbb{I}_{\mathscr{G}} + \mathbb{I}_{\mathscr{G}^c})|\sigma(\boldsymbol{w}^* \cdot \boldsymbol{x}) - y||\boldsymbol{w}^* \cdot \boldsymbol{x} - \boldsymbol{w}_i \cdot \boldsymbol{x}|] \\
&= 2\beta\, \mathbb{E}_{\widehat{p}_i}[|\sigma(\boldsymbol{w}^* \cdot \boldsymbol{x}) - y||\boldsymbol{w}^* \cdot \boldsymbol{x} - \boldsymbol{w}_i \cdot \boldsymbol{x}|] \\
&= 2\beta \int |\sigma(\boldsymbol{w}^* \cdot \boldsymbol{x}) - y||\boldsymbol{w}^* \cdot \boldsymbol{x} - \boldsymbol{w}_i \cdot \boldsymbol{x}| \mathrm{d}\widehat{p}_i \\
&= 2\beta \int |\sigma(\boldsymbol{w}^* \cdot \boldsymbol{x}) - y||\boldsymbol{w}^* \cdot \boldsymbol{x} - \boldsymbol{w}_i \cdot \boldsymbol{x}| \mathrm{d}\widehat{p}^* \\
&\quad + 2\beta \int |\sigma(\boldsymbol{w}^* \cdot \boldsymbol{x}) - y||\boldsymbol{w}^* \cdot \boldsymbol{x} - \boldsymbol{w}_i \cdot \boldsymbol{x}|(\mathrm{d}\widehat{p}_i - \mathrm{d}\widehat{p}^*). && (41)
\end{aligned}$$

In the last equality, the first integral is just the expectation with respect to $\widehat{p}^*$, and thus using Cauchy-Schwarz inequality, the definition of OPT, and Lemma C.6, the first term in Equation (41) can be bounded by

$$\begin{aligned}
&\int |\sigma(\boldsymbol{w}^* \cdot \boldsymbol{x}) - y||\boldsymbol{w}^* \cdot \boldsymbol{x} - \boldsymbol{w}_i \cdot \boldsymbol{x}| \mathrm{d}\widehat{p}^* \\
&\leq \sqrt{\mathbb{E}_{\widehat{p}^*}[(\sigma(\boldsymbol{w}^* \cdot \boldsymbol{x}) - y)^2] \mathbb{E}_{\widehat{p}^*}[(\boldsymbol{w}^* \cdot \boldsymbol{x} - \boldsymbol{w}_i \cdot \boldsymbol{x})^2]} \\
&\leq \sqrt{\widehat{\mathrm{OPT}}} \sqrt{6B} \|\boldsymbol{w}^* - \boldsymbol{w}_i\|_2 \\
&\leq \frac{24\beta B \widehat{\mathrm{OPT}}}{c_1} + \frac{c_1}{16\beta} \|\boldsymbol{w}^* - \boldsymbol{w}_i\|_2^2, \qquad (42)
\end{aligned}$$

where the last line is by Young's inequality and the second last line uses Lemma 2.5.

For the remaining integral in Equation (41), using the definition of chi-square divergence and Cauchy-Schwarz inequality, we have

$$\begin{aligned}
&\int |\sigma(\boldsymbol{w}^* \cdot \boldsymbol{x}) - y||\boldsymbol{w}^* \cdot \boldsymbol{x} - \boldsymbol{w}_i \cdot \boldsymbol{x}| (\mathrm{d}\widehat{p}_i - \mathrm{d}\widehat{p}^*) \\
= &\int |\sigma(\boldsymbol{w}^* \cdot \boldsymbol{x}) - y||\boldsymbol{w}^* \cdot \boldsymbol{x} - \boldsymbol{w}_i \cdot \boldsymbol{x}| \frac{(\mathrm{d}\widehat{p}_i - \mathrm{d}\widehat{p}^*)}{\sqrt{\mathrm{d}\widehat{p}^*}} \sqrt{\mathrm{d}\widehat{p}^*} \\
\overset{(i)}{\leq} &\sqrt{\chi^2(\widehat{p}_i, \widehat{p}^*) \mathbb{E}_{\widehat{p}^*}[(\sigma(\boldsymbol{w}^* \cdot \boldsymbol{x}) - y)^2 (\boldsymbol{w}^* \cdot \boldsymbol{x} - \boldsymbol{w}_i \cdot \boldsymbol{x})^2]} \\
\overset{(ii)}{\leq} &\chi^2(\widehat{p}_i, \widehat{p}^*)^{1/2} \mathbb{E}_{\widehat{p}^*}[(\sigma(\boldsymbol{w}^* \cdot \boldsymbol{x}) - y)^4]^{1/4} \mathbb{E}_{\widehat{p}^*}[(\boldsymbol{w}^* \cdot \boldsymbol{x} - \boldsymbol{w}_i \cdot \boldsymbol{x})^4]^{1/4} \\
\overset{(iii)}{\leq} &\chi^2(\widehat{p}_i, \widehat{p}^*)^{1/2} \widehat{\mathrm{OPT}}_{(2)}^{1/4} (6B \|\boldsymbol{w}^* - \boldsymbol{w}_i\|_2^4)^{1/4} \\
\overset{(iv)}{\leq} &\frac{4\beta\sqrt{6B}\sqrt{\widehat{\mathrm{OPT}}_{(2)}}}{c_1} \chi^2(\widehat{p}_i, \widehat{p}^*) + \frac{c_1}{16\beta} \|\boldsymbol{w}^* - \boldsymbol{w}_i\|_2^2, \qquad (43)
\end{aligned}$$

where (i) is by Cauchy-Schwarz, (ii) is by Cauchy-Schwarz again, (iii) is by the definition of $\widehat{\mathrm{OPT}}_{(2)}$ and Lemma C.6, and (iv) is by Young's inequality.

To complete the proof, it remains to plug Equations (41) to (43) back into Equation (39), and simplify. □

Gap upper bound proof of Lemma 3.3. Combining the upper and lower bounds from Lemma D.1 and Proposition D.2, we are now ready to prove Lemma 3.3, which we restate below.

Lemma 3.3 (Gap Upper Bound)**.** *Let $\boldsymbol{w}_i, \widehat{p}_i, a_i, A_i$ evolve according to Algorithm 1, where we take, by convention, $a_{-1} = A_{-1} = a_0 = A_0 = 0$ and $\boldsymbol{w}_{-1} = \boldsymbol{w}_0$, $\widehat{p}_{-1} = \widehat{p}_0$. Assuming Lemma 2.5 applies, then, for all $k \geq 1$, $\sum_{i=1}^k a_i \mathrm{Gap}(\boldsymbol{w}_i, \widehat{p}_i)$ is bounded above by*

$$\begin{aligned}
&\frac{1}{2}\|\boldsymbol{w}^* - \boldsymbol{w}_0\|_2^2 + \nu_0 D_\phi(\widehat{p}^*, \widehat{p}_0) - \frac{1 + 0.5c_1 A_k}{2} \|\boldsymbol{w}^* - \boldsymbol{w}_k\|_2^2 - (\nu_0 + \nu A_k) D_\phi(\widehat{p}^*, \widehat{p}_k) \\
&+ \sum_{i=1}^k a_i \frac{c_1}{4} \|\boldsymbol{w}^* - \boldsymbol{w}_i\|_2^2 + \frac{8\beta^2\sqrt{6B}\sqrt{\widehat{\mathrm{OPT}}_{(2)}}}{c_1} \sum_{i=1}^k a_i \chi^2(\widehat{p}_i, \widehat{p}^*) + \frac{48\beta^2 B \widehat{\mathrm{OPT}} A_k}{c_1}.
\end{aligned}$$

Proof. Combining the upper bound on $a_i L(\boldsymbol{w}_i, \widehat{p}^*)$ from Lemma D.1 with the lower bound on $a_i L(\boldsymbol{w}^*, \widehat{p}_i)$ from Proposition D.2 and recalling that $\mathrm{Gap}(\boldsymbol{w}_i, \widehat{p}_i) = L(\boldsymbol{w}_i, \widehat{p}^*) - L(\boldsymbol{w}^*, \widehat{p}_i)$ and

$A_i = A_{i-1} + a_i$, we have

$$
\begin{aligned}
a_i \mathrm{Gap}(\boldsymbol{w}_i, \widehat{\boldsymbol{p}}_i) \le a_i E_i& \\
&+ (\nu_0 + \nu A_{i-1}) D_\phi(\widehat{\boldsymbol{p}}^*, \widehat{\boldsymbol{p}}_{i-1}) - (\nu_0 + \nu A_i) D_\phi(\widehat{\boldsymbol{p}}^*, \widehat{\boldsymbol{p}}_i) \\
&+ (\nu_0 + \nu A_{i-2}) D_\phi(\widehat{\boldsymbol{p}}_{i-1}, \widehat{\boldsymbol{p}}_{i-2}) - (\nu_0 + \nu A_{i-1}) D_\phi(\widehat{\boldsymbol{p}}_i, \widehat{\boldsymbol{p}}_{i-1}) \\
&+ \frac{1 + 0.5 c_1 A_{i-1}}{2} \|\boldsymbol{w}^* - \boldsymbol{w}_{i-1}\|_2^2 - \frac{1 + 0.5 c_1 A_i}{2} \|\boldsymbol{w}^* - \boldsymbol{w}_i\|_2^2 \\
&+ \frac{1 + 0.5 c_1 A_{i-2}}{4} \|\boldsymbol{w}_{i-1} - \boldsymbol{w}_{i-2}\|_2^2 - \frac{1 + 0.5 c_1 A_{i-1}}{4} \|\boldsymbol{w}_i - \boldsymbol{w}_{i-1}\|_2^2 \\
&+ a_{i-1} \langle \mathbb{E}_{\widehat{\boldsymbol{p}}_{i-1}}[\boldsymbol{v}(\boldsymbol{w}_{i-1}; \boldsymbol{x}, y)] - \mathbb{E}_{\widehat{\boldsymbol{p}}_{i-2}}[\boldsymbol{v}(\boldsymbol{w}_{i-2}; \boldsymbol{x}, y)], \boldsymbol{w}^* - \boldsymbol{w}_{i-1} \rangle \\
&- a_i \langle \mathbb{E}_{\widehat{\boldsymbol{p}}_i}[\boldsymbol{v}(\boldsymbol{w}_i; \boldsymbol{x}, y)] - \mathbb{E}_{\widehat{\boldsymbol{p}}_{i-1}}[\boldsymbol{v}(\boldsymbol{w}_{i-1}; \boldsymbol{x}, y)], \boldsymbol{w}^* - \boldsymbol{w}_i \rangle.
\end{aligned}
$$

Observe that except for the first line on the right-hand side of the above inequality, all remaining terms telescope. Summing over $i = 1, 2, \ldots, k$ and recalling that, by convention, $a_0 = A_0 = a_{-1} = A_{-1} = 0$, $\boldsymbol{w}_{-1} = \boldsymbol{w}_0$, and $\widehat{\boldsymbol{p}}_{-1} = \widehat{\boldsymbol{p}}_0$, we have

$$
\begin{aligned}
\sum_{i=1}^k a_i \mathrm{Gap}(\boldsymbol{w}_i, \widehat{\boldsymbol{p}}_i) \le{}& \sum_{i=1}^k a_i E_i + \frac{1}{2} \|\boldsymbol{w}^* - \boldsymbol{w}_0\|_2^2 + \nu_0 D_\phi(\widehat{\boldsymbol{p}}^*, \widehat{\boldsymbol{p}}_0) \\
&- \frac{1 + 0.5 c_1 A_k}{2} \|\boldsymbol{w}^* - \boldsymbol{w}_k\|_2^2 - (\nu_0 + A_k) D_\phi(\widehat{\boldsymbol{p}}^*, \widehat{\boldsymbol{p}}_k) \\
&- a_k \langle \mathbb{E}_{\widehat{\boldsymbol{p}}_k}[\boldsymbol{v}(\boldsymbol{w}_k; \boldsymbol{x}, y)] - \mathbb{E}_{\widehat{\boldsymbol{p}}_{k-1}}[\boldsymbol{v}(\boldsymbol{w}_{k-1}; \boldsymbol{x}, y)], \boldsymbol{w}^* - \boldsymbol{w}_k \rangle \\
&- \frac{1 + 0.5 c_1 A_{k-1}}{4} \|\boldsymbol{w}_k - \boldsymbol{w}_{k-1}\|_2^2 - (\nu_0 + \nu A_{k-1}) D_\phi(\widehat{\boldsymbol{p}}_k, \widehat{\boldsymbol{p}}_{k-1}).
\end{aligned}
\tag{44}
$$

To complete the proof, it remains to bound $a_k |\langle \mathbb{E}_{\widehat{\boldsymbol{p}}_k}[\boldsymbol{v}(\boldsymbol{w}_k; \boldsymbol{x}, y)] - \mathbb{E}_{\widehat{\boldsymbol{p}}_{k-1}}[\boldsymbol{v}(\boldsymbol{w}_{k-1}; \boldsymbol{x}, y)], \boldsymbol{w}^* - \boldsymbol{w}_k \rangle|$, which is done similarly as in the proof of Proposition D.2. In particular,

$$
\begin{aligned}
&a_k |\langle \mathbb{E}_{\widehat{\boldsymbol{p}}_k}[\boldsymbol{v}(\boldsymbol{w}_k; \boldsymbol{x}, y)] - \mathbb{E}_{\widehat{\boldsymbol{p}}_{k-1}}[\boldsymbol{v}(\boldsymbol{w}_{k-1}; \boldsymbol{x}, y)], \boldsymbol{w}^* - \boldsymbol{w}_k \rangle| \\
\overset{(i)}{\le}{}& \frac{{a_k}^2}{1 + 0.5 c_1 A_k} \|\mathbb{E}_{\widehat{\boldsymbol{p}}_k}[\boldsymbol{v}(\boldsymbol{w}_k; \boldsymbol{x}, y)] - \mathbb{E}_{\widehat{\boldsymbol{p}}_{k-1}}[\boldsymbol{v}(\boldsymbol{w}_{k-1}; \boldsymbol{x}, y)]\|_2^2 + \frac{1 + 0.5 c_1 A_k}{4} \|\boldsymbol{w}^* - \boldsymbol{w}_k\|_2^2 \\
\overset{(ii)}{\le}{}& \frac{{a_k}^2}{1 + 0.5 c_1 A_k} \Big(2G^2 D_\phi(\widehat{\boldsymbol{p}}_k, \widehat{\boldsymbol{p}}_{k-1}) + 2\kappa^2 \|\boldsymbol{w}_k - \boldsymbol{w}_{k-1}\|_2^2 \Big) + \frac{1 + 0.5 c_1 A_k}{4} \|\boldsymbol{w}^* - \boldsymbol{w}_k\|_2^2 \\
\overset{(iii)}{\le}{}& (\nu_0 + \nu A_{k-1}) D_\phi(\widehat{\boldsymbol{p}}_k, \widehat{\boldsymbol{p}}_{k-1}) + \frac{1 + 0.5 c_1 A_{k-1}}{4} \|\boldsymbol{w}_k - \boldsymbol{w}_{k-1}\|_2^2 + \frac{1 + 0.5 c_1 A_k}{4} \|\boldsymbol{w}^* - \boldsymbol{w}_k\|_2^2,
\end{aligned}
\tag{45}
$$

where (i) is by Young's inequality and (ii) is by Equation (36), and (iii) is by $\frac{2G^2 {a_k}^2}{1+0.5 c_1 A_k} \le \nu_0 + \nu A_{k-1}$ and $\frac{2\kappa^2 {a_k}^2}{1+0.5 c_1 A_k} \le \frac{1+0.5 c_1 A_{k-1}}{4}$, which both hold by the choice of the step sizes in Algorithm 1.

To complete the proof, it remains to plug Equation (45) back into Equation (44), use the definition of E_i from Equation (37), and simplify. □

E Omitted Proofs in Main Theorem

Claim 3.5. *For all iterations $k \ge 0$, $\|\boldsymbol{w}_k\|_2 \le 2\|\boldsymbol{w}^*\|_2$.*

Proof of Claim 3.5. It trivially holds that $\mathbf{0} = \boldsymbol{w}_0 \in \mathscr{B}(2\|\boldsymbol{w}^*\|_2)$. Suppose $\|\boldsymbol{w}_i\|_2 \le 2\|\boldsymbol{w}^*\|_2$ for all iterations $i \le t$ where $t \ge 0$. Then

$$
\begin{aligned}
&-\frac{12\beta^2 B}{c_1}\widehat{\mathrm{OPT}}A_k + \sum_{i=1}^{k} a_i \frac{c_1}{2}\|\boldsymbol{w}_i - \boldsymbol{w}^*\|_2^2 + \sum_{i=1}^{k} \nu a_i D_\phi(\widehat{p}^*, \widehat{p}_i) + a_{k+1}\mathrm{Gap}(\boldsymbol{w}_{k+1}, \widehat{p}_{k+1})\\
&\le \sum_{i=1}^{k+1} a_i \mathrm{Gap}(\boldsymbol{w}_i, \widehat{p}_i)\\
&\le \frac{1}{2}\|\boldsymbol{w}^* - \boldsymbol{w}_0\|_2^2 + \nu_0 D_\phi(\widehat{p}^*, \widehat{p}_0) - \frac{1+0.5c_1A_{k+1}}{2}\|\boldsymbol{w}^* - \boldsymbol{w}_{k+1}\|_2^2 - (\nu_0 + \nu A_{k+1})D_\phi(\widehat{p}^*, \widehat{p}_{k+1})\\
&\quad + \sum_{i=1}^{k+1} a_i \frac{c_1}{4}\|\boldsymbol{w}^* - \boldsymbol{w}_i\|_2^2 + \frac{8\beta^2\sqrt{6B}\sqrt{\widehat{\mathrm{OPT}}_{(2)}}}{c_1}\sum_{i=1}^{k+1} a_i \chi^2(\widehat{p}_i, \widehat{p}^*) + \frac{48\beta^2 B\widehat{\mathrm{OPT}}A_{k+1}}{c_1},
\end{aligned}
$$

where we used the gap upper bound Lemma 3.3 again as it does not require $\boldsymbol{w} \in \mathscr{B}(\|\boldsymbol{w}^*\|_2)$. We proceed to deduce a different lower bound for $\mathrm{Gap}(\boldsymbol{w}_{k+1}, \widehat{p}_{k+1})$. Similar to Lemma 3.2, we break into two terms $L(\boldsymbol{w}^*, \widehat{p}_{k+1}) - (-L(\boldsymbol{w}^*, \widehat{p}^*)) \ge \nu D_\phi(\widehat{p}^*, \widehat{p}_{k+1})$ and $L(\boldsymbol{w}_{k+1}, \widehat{p}^*) - L(\boldsymbol{w}^*, \widehat{p}^*) = \mathbb{E}_{\widehat{p}^*}[(\sigma(\boldsymbol{w}_{k+1} \cdot \boldsymbol{x}) - y)^2 - (\sigma(\boldsymbol{w}^* \cdot \boldsymbol{x}) - y)^2] \ge -\widehat{\mathrm{OPT}}$, where the first term is bounded the same way as in Lemma 3.2. Hence, $\mathrm{Gap}(\boldsymbol{w}_{k+1}, \widehat{p}_{k+1}) \ge -\widehat{\mathrm{OPT}}$. Therefore, we simplify as before

$$
\begin{aligned}
&\frac{1+0.5c_1A_{k+1}}{2}\|\boldsymbol{w}^* - \boldsymbol{w}_{k+1}\|_2^2 + (\nu_0 + \nu A_{k+1})D_\phi(\widehat{p}^*, \widehat{p}_{k+1}) - a_i\frac{c_1}{4}\|\boldsymbol{w}^* - \boldsymbol{w}_{k+1}\|_2^2\\
&\le \frac{1}{2}\|\boldsymbol{w}^* - \boldsymbol{w}_0\|_2^2 + \nu_0 D_\phi(\widehat{p}^*, \widehat{p}_0) + \frac{12\beta^2 B}{c_1}\widehat{\mathrm{OPT}}A_k + a_{k+1}\widehat{\mathrm{OPT}} + \frac{48\beta^2 B\widehat{\mathrm{OPT}}A_{k+1}}{c_1},
\end{aligned}
$$

which implies by nonnegativity of Bregman divergence that:

$$
\frac{2+c_1A_k}{4}\|\boldsymbol{w}^* - \boldsymbol{w}_{k+1}\|_2^2 \le \frac{1}{2}\|\boldsymbol{w}^* - \boldsymbol{w}_0\|_2^2 + \nu_0 D_\phi(\widehat{p}^*, \widehat{p}_0) + \Big(\frac{60\beta^2 B}{c_1}A_k + a_{k+1}\Big(1 + \frac{48\beta^2 B}{c_1}\Big)\Big)\widehat{\mathrm{OPT}}.
$$

Rearranging and using $2 + c_1A_k \ge 2$,

$$
\|\boldsymbol{w}^* - \boldsymbol{w}_{k+1}\|_2^2 \le \|\boldsymbol{w}^* - \boldsymbol{w}_0\|_2^2 + 2\nu_0 D_\phi(\widehat{p}^*, \widehat{p}_0) + \Big(\frac{240\beta^2 B}{c_1^2} + \frac{4a_{k+1}}{2+c_1A_k}\Big(1 + \frac{48\beta^2 B}{c_1}\Big)\Big)\widehat{\mathrm{OPT}},
$$

Our choice of stepsizes a_i implies $\frac{a_{k+1}}{2+c_1A_k} \le 1/\max\{\kappa, G\} \le 1$, hence

$$
\|\boldsymbol{w}^* - \boldsymbol{w}_{k+1}\|_2^2 \le \|\boldsymbol{w}^* - \boldsymbol{w}_0\|_2^2 + 2\nu_0 D_\phi(\widehat{p}^*, \widehat{p}_0) + \Big(\frac{288\beta^2 B}{c_1^2} + \frac{1}{\max\{\kappa, G\}}\Big)\widehat{\mathrm{OPT}},
$$

Claim E.1. *For $\nu \ge 8\beta^2\sqrt{6B}\sqrt{\widehat{\mathrm{OPT}}_{(2)}}/c_1$, it holds that*

$$
\chi^2(\widehat{p}^*, \widehat{p}_0) = \frac{\mathrm{Var}_{\widehat{p}_0}(\ell(\boldsymbol{w}^*))}{4\nu^2} \le \frac{\widehat{\mathrm{OPT}}_{(2)}}{2\nu^2} \le c_1/(1536\beta^4 B).
$$

Similarly, for $\nu \ge \mathbb{E}_{p_0}\ell(\boldsymbol{w}^*), 8\beta^2\sqrt{6B}\sqrt{\mathrm{OPT}_{(2)}}/c_1$, it holds that

$$
\chi^2(p^*, p_0) = \frac{\mathrm{Var}_{p_0}(\ell(\boldsymbol{w}^*))}{4\nu^2} \le \frac{\mathrm{OPT}_{(2)}}{2\nu^2} \le c_1/(1536\beta^4 B).
$$

Proof of Claim E.1. By Corollary C.3, $\chi^2(\widehat{p}^*, \widehat{p}_0) = \frac{\mathbb{E}_{\widehat{p}_0}[\ell^2(\boldsymbol{w}^*)] - (\mathbb{E}_{\widehat{p}_0}[\ell(\boldsymbol{w}^*)])^2}{4\nu^2} \le \frac{\mathbb{E}_{\widehat{p}_0}[\ell^2(\boldsymbol{w}^*)]}{4\nu^2} \le \frac{\widehat{\mathrm{OPT}}_{(2)}}{2\nu^2} \le c_1/(1536\beta^4 B)$, where the second last inequality uses $\widehat{p}^* \ge \widehat{p}_0/2$ from Corollary C.2 and the last inequality comes from lower bound on ν in the assumption.

The population version follows analogously. □

Since $D_\phi(\widehat{p}^*, \widehat{p}_0) = \chi^2(\widehat{p}^*, \widehat{p}_0)$, by choosing $\nu_0 = 768\beta^4 B\epsilon/c_1$, we ensure $2\nu_0 D_\phi(\widehat{p}^*, \widehat{p}_0) \le \epsilon$.

By choosing ν_0 small enough and initialization $\boldsymbol{w}_0 = \mathbf{0}$, it holds that

$$\|\boldsymbol{w}^* - \boldsymbol{w}_{k+1}\|_2^2 \le \|\boldsymbol{w}^*\|_2^2 + \epsilon + \Big(\frac{288\beta^2 B}{c_1^2} + \frac{1}{\max\{\kappa, G\}}\Big)\widehat{\mathrm{OPT}}.$$

We may assume without loss of generality that $\frac{1}{\max\{\kappa,G\}} \ll \frac{288\beta^2 B}{c_1^2}$ because both κ and G is $O(d)$ but the right hand side is an absolute constant. We may also assume without loss of generality that $\frac{300\beta^2 B}{c_1^2}\widehat{\mathrm{OPT}} + \epsilon \le 2\|\boldsymbol{w}^*\|_2^2$, thus completing the induction step $\|\boldsymbol{w}^* - \boldsymbol{w}_{k+1}\|_2 \le 2\|\boldsymbol{w}^*\|_2$. The reason for the last no loss of generality is the following: otherwise, we can compare, per Claim E.2, the empirical risk of the output from our algorithm and of $\hat{\boldsymbol{w}} = \mathbf{0}$ and output the solution with the lower risk to get an $O(\mathrm{OPT}) + \epsilon$ solution. □

Claim E.2 (Zero-Tester)**.** *In the setting of Theorem 3.1, it is possible to efficiently check if $R(\mathbf{0}; \widehat{\mathsf{p}}_0) > R(\widehat{\boldsymbol{w}}; \widehat{\mathsf{p}}_0)$ or not; where $\widehat{\boldsymbol{w}}$ is the output of Algorithm 1.*

Proof. Observe that $L(\boldsymbol{w}, \widehat{\mathsf{p}}) = \sum_{i=1}^N \widehat{\mathsf{p}}_i(\sigma(\boldsymbol{w}\cdot\boldsymbol{x}) - y)^2 - \nu\chi^2(\widehat{\mathsf{p}}, \widehat{\mathsf{p}}_0)$ is $1/\nu$-strongly concave in $(\widehat{\mathsf{p}}^{(1)}, \ldots, \widehat{\mathsf{p}}^{(N)})$. Now, since $R(\boldsymbol{w}; \widehat{\mathsf{p}}_0) = \max_{\widehat{\mathsf{p}}} L(\boldsymbol{w}, \widehat{\mathsf{p}})$, we can estimate the risk at any given $\boldsymbol{w}$ using standard maximization techniques (such as gradient descent).

To test which risk is larger, we estimate $R(\mathbf{0}; \widehat{\mathsf{p}}_0)$ and $R(\widehat{\boldsymbol{w}}; \widehat{\mathsf{p}}_0)$ to a necessary accuracy and then compare.

□

F Parameter Estimation to Loss and Risk Approximation

Theorem 3.1 shows that Algorithm 1 recovers a vector $\widehat{\boldsymbol{w}}$ such that $\|\widehat{\boldsymbol{w}} - \boldsymbol{w}^*\|_2 \le \sqrt{\mathrm{OPT}} + \sqrt{\epsilon}$, where $\mathrm{OPT} := \mathbb{E}_{\mathsf{p}^*}(\sigma(\boldsymbol{w}^*\cdot\boldsymbol{x}) - y)^2$.

F.1 Loss Approximation

In this section we show that this implies that the neuron we recover achieves a constant factor approximation to the optimal squared loss.

Lemma F.1. *Let p^* satisfy Assumption 2.1 and Assumption 2.2. Suppose $(\widehat{\boldsymbol{w}}, \widehat{\mathsf{p}})$ is the solution returned by Algorithm 1 when given $N =$ samples drawn from p_0. Then, $\mathbb{E}_{\mathsf{p}^*}(\sigma(\widehat{\boldsymbol{w}}\cdot\boldsymbol{x}) - y)^2 \le O_{\beta,B}(\mathrm{OPT}) + \epsilon$.*

Proof. Recall that σ is β-Lipschitz, and Fact 2.4 gives us $\mathbb{E}_{\mathsf{p}^*}(\boldsymbol{u}\cdot\boldsymbol{x})^2 \le 5B$ for all unit vectors $\boldsymbol{u}$. These together imply,

$$\begin{aligned}
\mathbb{E}_{\mathsf{p}^*}(\sigma(\widehat{\boldsymbol{w}}\cdot\boldsymbol{x}) - y)^2 &\le 2\mathbb{E}_{\mathsf{p}^*}(\sigma(\boldsymbol{w}^*\cdot\boldsymbol{x}) - y)^2 + 2\mathbb{E}_{\mathsf{p}^*}(\sigma(\boldsymbol{w}^*\cdot\boldsymbol{x}) - \sigma(\widehat{\boldsymbol{w}}\cdot\boldsymbol{x}))^2 \\
&\le 2\,\mathrm{OPT} + 2\beta^2\|\widehat{\boldsymbol{w}} - \boldsymbol{w}^*\|_2^2\,\mathbb{E}_{\mathsf{p}^*}\left(\frac{\widehat{\boldsymbol{w}} - \boldsymbol{w}^*}{\|\widehat{\boldsymbol{w}} - \boldsymbol{w}^*\|}\cdot\boldsymbol{x}\right)^2 \\
&\le 2\,\mathrm{OPT} + 2\beta^2(2C_3^2\,\mathrm{OPT} + 2\epsilon)\,5B \\
&\le (2 + 20B\beta^2C_3^2)\ \mathrm{OPT} + 10\beta^2B\epsilon.
\end{aligned}$$

□

F.2 Risk Approximation

Fix $\hat{\boldsymbol{w}}$ as output by Algorithm 1 and $\boldsymbol{w}^*$ as defined in Definition 1.2. Since we are bounding the population risk throughout this subsection, we write $R(\boldsymbol{w}) = R(\boldsymbol{w}; \mathsf{p}_0)$ in short. The goal of this subsection is to show

$$R(\hat{\boldsymbol{w}}) - R(\boldsymbol{w}^*) \le O(\mathrm{OPT}) + \epsilon.$$

We first introduce some convex analysis results that we rely on in this subsection:

Fact F.2 (Strong convexity of chi-square divergence)**.** *Consider the space $\mathcal{P}(p_0) = \{p : p \ll p_0\}$. For $p \in \mathcal{P}(p_0)$, we denote by $\frac{\mathrm{d}p}{\mathrm{d}p_0}$ the Radon–Nikodym derivative of p with respect to p_0, and we define $\|p\|_{p_0}^* = \sqrt{\int (\frac{\mathrm{d}p}{\mathrm{d}p_0})^2 \mathrm{d}p_0}$. Then $p \mapsto \chi^2(p, p_0)$ is 2-strongly convex with respect to $\|\cdot\|_{p_0}^*$.*

Fact F.3. *Consider the space $\mathcal{P}(p_0) = \{p : p \ll p_0\}$. Denote by $\langle \cdot, \cdot \rangle_{p_0}$ the inner product $\langle \ell_1, \ell_2 \rangle_{p_0} = \int \ell_1 \ell_2 \mathrm{d}p_0$ and denote by $\|\cdot\|_{p_0}$ the correponding norm. Then $\|\cdot\|_{p_0}$ is the dual norm of $\|\cdot\|_{p_0}^*$ defined in Fact F.2.*

Definition F.4 (Convex conjugate)**.** Given a convex function defined on a vector space $\mathbb{E}$ denoted by $f : \mathbb{E} \to \mathbb{R}$, its convex conjugate is defined as:

$$f^*(y) = \sup_{x \in \mathbb{E}} (\langle y, x \rangle - f(x))$$

for all $y \in \mathbb{E}^*$ where $\mathbb{E}^*$ is the dual space of $\mathbb{E}$ and $\langle y, x \rangle$ denotes the inner product.

Fact F.5 (Conjugate Correspondence Theorem, [Bec17, Theorem 5.26])**.** *Let $\nu > 0$. If $f : \mathbb{E} \to \mathbb{R}$ is a ν-strongly convex continuous function, then its convex conjugate $f^* : \mathbb{E}^* \to \mathbb{R}$ is $\frac{1}{\nu}$-smooth.*

We are then able to state and prove the key technical corollary in this subsection:

Corollary F.6. *For any p_0-measurable function $\ell : \mathbb{R}^d \times \mathbb{R} \to \mathbb{R}$, let $\mathcal{R}(\ell) = \max_{p \ll p_0} \mathbb{E}_{p_0} \ell - \nu \chi^2(p, p_0)$. The function $\mathcal{R}(\cdot)$ is $1/(2\nu)$-smooth with respect to the norm $\|\cdot\|_{p_0}$ defined in Fact F.3.*

Proof. Observe by definition of the convex conjugate that $\mathcal{R}(\cdot)$ is the convex conjugate of the function $\nu\chi^2(\cdot, p_0)$. Since the function $\nu\chi^2(\cdot, p_0)$ is 2ν-strongly convex with respect to the norm $\|\cdot\|_{p_0}^*$ by Fact F.2, it follows from Fact F.5 that $\mathcal{R}(\cdot)$ is $1/(2\nu)$-smooth with respect to the norm $\|\cdot\|_{p_0}$. □

For ease of presentation, we define the following quantities: let $\ell^*(\boldsymbol{x}, y) = (\sigma(\boldsymbol{w}^* \cdot \boldsymbol{x}) - y)^2$ and $\hat{\ell}(\boldsymbol{x}, y) = (\sigma(\hat{\boldsymbol{w}} \cdot \boldsymbol{x}) - y)^2$. We first compute $\nabla_\ell \mathcal{R}(\ell^*)$ by conjugate subgradient theorem.

Fact F.7 (Conjugate Subgradient Theorem [Bec17, Theorem 4.20])**.** *Let $f : \mathbb{E} \to \mathbb{R}$ be convex and continuous. The following claims are equivalent for any $x \in \mathbb{E}$ and $y \in \mathbb{E}^*$:*

1. $\langle y, x \rangle = f(x) + f^*(y)$
2. $y \in \partial f(x)$
3. $x \in \partial f^*(y)$

Corollary F.8. *Let p^* be as defined in Definition 1.2. Then $p^* \in \partial_\ell \mathcal{R}(\ell^*)$.*

Proof. We verify that $\mathcal{R}(\ell^*) = \max_{p \ll p_0} \mathbb{E}_{p_0}[\sigma(\boldsymbol{w}^* \cdot \boldsymbol{x} - y)^2] - \nu\chi^2(p, p_0) = \mathbb{E}_{p^*}[\sigma(\boldsymbol{w}^* \cdot \boldsymbol{x} - y)^2] - \nu\chi^2(p^*, p_0) = \mathbb{E}_{p^*}\ell^* - \nu\chi^2(p^*, p_0) = \langle p^*, \ell^* \rangle - \nu\chi^2(p^*, p_0)$, where the second equality is the definition of p^* and the third equality is the definition of ℓ^*. By Fact F.7 and observing that $\mathcal{R}(\cdot)$ is the convex conjugate of the function $\nu\chi^2(\cdot, p_0)$, we have $p^* \in \partial_\ell \mathcal{R}(\ell^*)$. □

Theorem F.9. *Suppose Corollary C.2 holds for both $\boldsymbol{w}^*$ and $\hat{\boldsymbol{w}}$ with respect to the population distribution. Then*

$$R(\hat{\boldsymbol{w}}; p_0) - R(\boldsymbol{w}^*; p_0) \le C_4(\mathrm{OPT} + \epsilon),$$

where $C_4 = 1 + 2(10B\beta^2 + c_1)C_3 + c_1\sqrt{5B}\beta^2 C_3^2$. In particular, Corollary C.2 holds for both $\boldsymbol{w}^*$ and $\hat{\boldsymbol{w}}$ is satisfied under the assumptions in Theorem 3.1.

Proof. By the definition of smoothness, it holds that for any $p \in \partial_\ell \mathcal{R}(\ell^*)$, we have

$$R(\hat{\boldsymbol{w}}; p_0) - R(\boldsymbol{w}^*; p_0) = \mathcal{R}(\hat{\ell}) - \mathcal{R}(\ell^*) \le \langle p, \hat{\ell} - \ell^* \rangle + \frac{1}{2\nu}\|\hat{\ell} - \ell^*\|_{p_0}^2$$

Hence it follows from Corollary F.8 that

$$\begin{aligned} R(\hat{\boldsymbol{w}}; p_0) - R(\boldsymbol{w}^*; p_0) &\le \langle p^*, \hat{\ell} - \ell^* \rangle + \frac{1}{2\nu}\|\hat{\ell} - \ell^*\|_{p_0}^2 \\ &= \mathbb{E}_{p^*}[(\sigma(\hat{\boldsymbol{w}} \cdot \boldsymbol{x}) - y)^2 - (\sigma(\boldsymbol{w}^* \cdot \boldsymbol{x}) - y)^2] + \frac{1}{2\nu}\mathbb{E}_{p_0}[((\sigma(\hat{\boldsymbol{w}} \cdot \boldsymbol{x}) - y)^2 - (\sigma(\boldsymbol{w}^* \cdot \boldsymbol{x}) - y)^2)^2]. \end{aligned}$$

We use the following shorthand for ease of presentation.

$$\Psi(\boldsymbol{x}, y) = \sigma(\boldsymbol{w}^* \cdot \boldsymbol{x}) - y,$$
$$\Delta(\boldsymbol{x}, y) = \sigma(\hat{\boldsymbol{w}} \cdot \boldsymbol{x}) - \sigma(\boldsymbol{w}^* \cdot \boldsymbol{x}).$$

Then

$$\begin{aligned}
&R(\hat{\boldsymbol{w}}; p_0) - R(\boldsymbol{w}^*; p_0) \le \langle p^*, \hat{\ell} - \ell^* \rangle + \frac{1}{2\nu} \|\hat{\ell} - \ell^*\|_{p_0}^2 \\
&= \mathbb{E}_{p^*}[(\Delta + \Psi)^2 - \Psi^2] + \frac{1}{2\nu} \mathbb{E}_{p_0}[((\Delta + \Psi)^2 - \Psi^2)^2] \\
&= \mathbb{E}_{p^*}[\Delta^2 + 2\Delta\Psi] + \frac{1}{2\nu} \mathbb{E}_{p_0}[\Delta^2(\Delta + 2\Psi)^2] \\
&\le \mathbb{E}_{p^*}[\Delta^2 + 2\Delta\Psi] + \frac{1}{\nu} \mathbb{E}_{p_0}[\Delta^2(\Delta^2 + 4\Psi^2)],
\end{aligned}$$

where the last inequality is the standard inequality $(a + b)^2 \le 2a^2 + 2b^2$.

Recall in Fact 2.4 that $\mathbb{E}_{p^*}[\Delta\Psi] \ge c_0 \|\hat{\boldsymbol{w}} - \boldsymbol{w}^*\|_2^2$ and $\mathbb{E}_{p^*}[(\boldsymbol{x} \cdot (\hat{\boldsymbol{w}} - \boldsymbol{w}^*))^\tau] \le 5B \|\hat{\boldsymbol{w}} - \boldsymbol{w}^*\|_2^\tau$.

From the second- and fourth-moment bounds, we have $\mathbb{E}_{p^*}[\Delta^\tau] = \mathbb{E}_{p^*}[(\sigma(\boldsymbol{x} \cdot \hat{\boldsymbol{w}}) - \sigma(\boldsymbol{x} \cdot \boldsymbol{w}^*))^\tau] \le \beta^\tau \mathbb{E}_{p^*}[(\boldsymbol{x} \cdot (\hat{\boldsymbol{w}} - \boldsymbol{w}^*))^\tau] \le 5B\beta^\tau \|\hat{\boldsymbol{w}} - \boldsymbol{w}^*\|_2^\tau$ for $\tau = 2, 4$, where the second last inequality follows from β-Lipschitzness of $\sigma(\cdot)$. Taking $\tau = 2$ gives us a bound for $\mathbb{E}_{p^*}[\Delta^2]$.

For $\mathbb{E}_{p^*}[\Delta\Psi]$, it follows from Cauchy-Schwarz that $\mathbb{E}_{p^*}[\Delta\Psi] \le \sqrt{\mathbb{E}_{p^*}[\Delta^2] \mathbb{E}_{p^*}[\Psi^2]} \le \sqrt{5B\beta^2 \operatorname{OPT}} \|\hat{\boldsymbol{w}} - \boldsymbol{w}^*\|_2$.

By Corollary C.2, we have $\mathbb{E}_{p_0}[\Delta^4] \le 2\mathbb{E}_{p^*}[\Delta^4] \le 5B\beta^4 \|\hat{\boldsymbol{w}} - \boldsymbol{w}^*\|_2^4$.

Finally, similarly by Corollary C.2, it follows additionally from Cauchy-Schwarz that $\mathbb{E}_{p_0}[\Delta^2\Psi^2] \le 2\mathbb{E}_{p^*}[\Delta^2\Psi^2] \le 2\sqrt{\mathbb{E}_{p_0}[\Delta^4]\mathbb{E}_{p^*}[\Psi^4]} \le 2\|\hat{\boldsymbol{w}} - \boldsymbol{w}^*\|_2^2 \sqrt{5B\beta^4 \operatorname{OPT}_{(2)}}$. By Theorem 3.1, we have $\nu \ge 8\beta^2\sqrt{6B}\sqrt{\operatorname{OPT}_{(2)} + \epsilon}/c_1$ by assumption, hence $4\mathbb{E}_{p_0}[\Delta^2\Psi^2]/\nu \le c_1 \|\hat{\boldsymbol{w}} - \boldsymbol{w}^*\|_2^2$.

Combining the above four bounds and the guarantee of Theorem 3.1 that $\|\hat{\boldsymbol{w}} - \boldsymbol{w}^*\|_2^2 \le 2C_3 \operatorname{OPT} + 2\epsilon$, we have

$$\begin{aligned}
&R(\hat{\boldsymbol{w}}; p_0) - R(\boldsymbol{w}^*; p_0) \\
&\le 5B\beta^2 \|\hat{\boldsymbol{w}} - \boldsymbol{w}^*\|_2^2 + 2\sqrt{5B\beta^2 \operatorname{OPT}} \|\hat{\boldsymbol{w}} - \boldsymbol{w}^*\|_2 + c_1 \|\hat{\boldsymbol{w}} - \boldsymbol{w}^*\|_2^2 + 5B\beta^4 \|\hat{\boldsymbol{w}} - \boldsymbol{w}^*\|_2^4/\nu \\
&\le \operatorname{OPT} + (10B\beta^2 + c_1 + 5B\beta^4 \|\hat{\boldsymbol{w}} - \boldsymbol{w}^*\|_2^2/\nu) \|\hat{\boldsymbol{w}} - \boldsymbol{w}^*\|_2^2 \\
&\le \operatorname{OPT} + 2(10B\beta^2 + c_1)(C_3 \operatorname{OPT} + \epsilon) + 40B\beta^4(C_3^2 \operatorname{OPT}^2 + \epsilon^2)/\nu.
\end{aligned}$$

By Corollary C.9, $\operatorname{OPT} \le \sqrt{\operatorname{OPT}_{(2)}}$, hence $\nu \ge 8\beta^2\sqrt{6B}\sqrt{\operatorname{OPT}_{(2)} + \epsilon}/c_1 \ge 8\beta^2\sqrt{6B}\max\{\operatorname{OPT}, \sqrt{\epsilon}\}/c_1$, hence

$$\begin{aligned}
&R(\hat{\boldsymbol{w}}; p_0) - R(\boldsymbol{w}^*; p_0) \\
&\le \operatorname{OPT} + 2(10B\beta^2 + c_1)(C_3 \operatorname{OPT} + \epsilon) + c_1\sqrt{5B}\beta^2(C_3^2 \operatorname{OPT}^2 + \epsilon^2)/\max\{\operatorname{OPT}, \sqrt{\epsilon}\} \\
&\le \operatorname{OPT} + 2(10B\beta^2 + c_1)(C_3 \operatorname{OPT} + \epsilon) + c_1\sqrt{5B}\beta^2(C_3^2 \operatorname{OPT} + \epsilon^{1.5}) \\
&= (1 + 2(10B\beta^2 + c_1)C_3 + c_1\sqrt{5B}\beta^2 C_3^2) \operatorname{OPT} + (2(10B\beta^2 + c_1) + c_1\sqrt{5B}\beta^2)\epsilon.
\end{aligned}$$

□

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction clearly outline the open problem we provide the first results for, clearly state the assumptions involved (cf. Section 1.1 — Problem Setup and Section 1.2 — Main Result), and give a detailed technical overview of our algorithm (cf. Section 1.3 — Technical Overview).

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The limitations are clearly stated in the statements of each theorem and are discussed in the introduction of the paper and description of the key lemma (Lemma 3.4).

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: Each theorem statement provides all the assumptions and we provide complete proofs for all statements that are either in the main body of the paper or in the appendix.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [NA]

Justification: The paper is theoretical in nature and does not include experiments.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [NA]

   Justification: The paper is theoretical in nature and does not include experiments.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [NA]

   Justification: The paper is theoretical in nature and does not include experiments.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [NA]

   Justification: The paper is theoretical in nature and does not include experiments.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [NA]

   Justification: The paper is theoretical in nature and does not include experiments.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: Our research conforms in every respect with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [NA]

    Justification: The work is theoretical and we do not see any major or immediate implications

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: The work is theoretical.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [NA]

    Justification: This paper does not use existing assets.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: This paper does not release new assets.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: This paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: This paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.