Robust Learning of a Group DRO Neuron

Guyang Cao
UW-Madison
Madison, WI, USA

Shuyao Li
UW-Madison
Madison, WI, USA

Sushrut Karmalkar
Microsoft Research
Cambridge, UK

Jelena Diakonikolas
UW-Madison
Madison, WI, USA

Abstract

We study the problem of learning a single neuron under standard squared loss in the presence of arbitrary label noise and group-level distributional shifts, for a broad family of covariate distributions. Our goal is to identify a "best-fit" neuron parameterized by $\boldsymbol{w}_*$ that performs well under the most challenging reweighting of the groups. Specifically, we address a Group Distributionally Robust Optimization problem: given sample access to K distinct distributions $p_{[1]}, \ldots, p_{[K]}$, we seek to approximate $\boldsymbol{w}_*$ that minimizes the worst-case objective over convex combinations of group distributions $\boldsymbol{\lambda} \in \Delta_K$, where the objective is $\sum_{i\in[K]} \lambda_{[i]} \mathbb{E}_{(\boldsymbol{x},y)\sim p_{[i]}}(\sigma(\boldsymbol{w}\cdot\boldsymbol{x}) - y)^2 - \nu d_f(\boldsymbol{\lambda}, \frac{1}{K}\mathbf{1})$ and d_f is an f-divergence that imposes (optional) penalty on deviations from uniform group weights, scaled by a parameter $\nu \geq 0$. We develop a computationally efficient primal-dual algorithm that outputs a vector $\widehat{\boldsymbol{w}}$ that is constant-factor competitive with $\boldsymbol{w}_*$ under the worst-case group weighting. Our analytical framework tackles the inherent nonconvexity of the loss function, providing robust learning guarantees in the face of arbitrary label corruptions and group-specific distributional shifts. The implementation of the dual extrapolation update motivated by our algorithmic framework shows promise on LLM pre-training benchmarks.

1 INTRODUCTION

The challenge of ensuring model robustness against distributional shifts is a central theme in modern machine learning. Group Distributionally Robust Optimization (Group DRO) has emerged as a principled framework to address such challenges, particularly in settings with heterogeneous data from distinct subpopulations (or groups) (Hashimoto et al., 2018; Oren et al., 2019; Sagawa et al., 2020). The objective of Group DRO is to learn a model that minimizes its loss under the worst-case reweighting of these groups, thereby guarding against poor performance on any single subpopulation. Group DRO has enjoyed significant empirical success in large-scale applications; for instance, Xie et al. (2023) and Xia et al. (2024) leverage dynamic reweighting of data domains to improve the performance of large language models. Despite these practical advances, most of the existing theory for DRO is limited to convex optimization problems. This leaves a critical gap in our understanding of the nonconvex landscapes that characterize deep learning, where these methods are most impactful.

Alas, even *without* distributional robustness, learning guarantees for nonconvex models are nontrivial. A canonical example is the classical problem of learning a single neuron (Rosenblatt, 1958; Nelder and Wedderburn, 1972): namely, a function of the form $\sigma(\boldsymbol{w}_*^\top \boldsymbol{x})$, where σ is a known activation function (e.g., ReLU: $\sigma(t) = \max\{0, t\}$), $\boldsymbol{w}_* \in \mathbb{R}^d$ is the unknown parameter vector, and $\boldsymbol{x} \in \mathbb{R}^d$ is the data vector; the goal of a learner is to minimize the mean squared loss $\mathscr{L}_2(\boldsymbol{w}) := \mathbb{E}_{(\boldsymbol{x},y)\sim\mathscr{D}}[(\sigma(\boldsymbol{w}^\top\boldsymbol{x}) - y)^2]$ over a centered Euclidean ball of given radius W, denoted by $\mathscr{B}(W)$.

This problem is already computationally challenging in general because of the inherent nonconvexity of the squared loss for common activations like ReLU. Without any distributional assumptions imposed on the labeled examples and for standard activations like the sigmoid and ReLU, robust learning is NP-hard even if we only require constant-factor approximation for the minimum mean squared loss (Sima, 2002; Manurangsi and Reichman, 2018). Under more structured conditions, tractability can be recovered: if the labels are *realizable* (i.e., if for labeled pairs $(\boldsymbol{x}, y)$, $y = \sigma(\boldsymbol{w}_*^\top \boldsymbol{x})$ for some fixed activation σ and parameter vector $\boldsymbol{w}_*$) or exhibit zero-mean, bounded-variance noise, fairly mild assumptions on the $\boldsymbol{x}$-marginal distribution and activation σ suffice for minimizing the mean squared loss to error $\epsilon > 0$ in polynomial time (in $d, 1/\epsilon$ and other problem parameters) (Kalai and Sastry, 2009; Kakade et al., 2011; Soltanolkotabi, 2017).

Proceedings of the 29th International Conference on Artificial Intelligence and Statistics (AISTATS) 2026, Tangier, Morocco. PMLR: Volume 300.

However, once arbitrary label noise is introduced—a realistic setting accounting for possible model misspecification and non-structured noise classically studied in the context of agnostic learning (Haussler, 1992; Kearns et al., 1992)—the problem again becomes intractable: even in the Gaussian setting, achieving additive error $\epsilon > 0$ requires $d^{\mathrm{poly}(1/\epsilon)}$ time, ruling out any polynomial-time algorithm (Goel et al., 2019; Diakonikolas et al., 2020a; Goel et al., 2020; Diakonikolas et al., 2021b; Diakonikolas et al., 2023). These hardness results highlight the necessity of designing efficient constant-factor approximation algorithms under structural assumptions imposed on the $\boldsymbol{x}$-marginal distribution and the class of activation functions.

Our work addresses the problem of learning a single neuron in a setting that combines two significant challenges: adversarial label noise and group-level distributional shifts. We consider a scenario where training data originate from K groups, each associated with an unknown distribution $p_{[1]}, \ldots, p_{[K]}$, while the group memberships of examples are known. The reference distribution, representing an unperturbed state, is an equal mixture $p_0 = \frac{1}{K}\sum_{i=1}^{K} p_{[i]}$. Our aim is to determine a weight vector $\boldsymbol{w} \in \mathscr{B}(W)$ parameterizing a neuron whose loss is robust to the worst-case reweighting of the groups. We formalize this by seeking a $\boldsymbol{w}$ minimizing the squared loss under a worst-case convex combination of group distributions, where deviations from a uniform weighting are penalized by an f-divergence. This leads to the following min-max problem:

$$\min_{\boldsymbol{w}\in\mathscr{B}(W)} \max_{\boldsymbol{\lambda}\in\Delta_K} \sum_{i=1}^{K} \lambda_{[i]}\, \mathbb{E}_{p_{[i]}}\big(\sigma(\boldsymbol{w}^\top \boldsymbol{x}) - y\big)^2 - \nu\, d_f\big(\boldsymbol{\lambda}, \tfrac{1}{K}\mathbf{1}\big).$$

Here, $\sigma : \mathbb{R} \to \mathbb{R}$ denotes a known activation function such as ReLU (see Definition 1.1 for the definition of activations captured by our results), $\nu \geq 0$ is a regularization parameter controlling the robustness trade-off, and d_f is a strongly convex f-divergence that quantifies the discrepancy between the learned group-weight vector $\boldsymbol{\lambda}$ and a uniform weighting. This formulation captures both *adversarial label noise* (within groups) and *distributional shifts* (across groups).

For this setting, we develop the first computationally efficient algorithm with provable guarantees. Our algorithm is primal-dual and it outputs an estimated parameter $\widehat{\boldsymbol{w}}$ that is competitive with the optimal parameter vector $\boldsymbol{w}_*$. Specifically, for the worst-case reweighting $\boldsymbol{\lambda}^*$ corresponding to $\boldsymbol{w}_*$, our estimate satisfies, for a constant $C > 1$,

$$\sum_{i\in[K]} \lambda^*_{[i]} \mathbb{E}_{p_{[i]}}(\sigma(\widehat{\boldsymbol{w}} \cdot \boldsymbol{x}) - y)^2 \leq C \max_{i\in[K]} \mathbb{E}_{p_{[i]}}(\sigma(\boldsymbol{w}_* \cdot \boldsymbol{x}) - y)^2 + \epsilon.$$

We recall here that, as discussed earlier in the introduction, even for $K = 1$ (no group shifts), obtaining error with $C = 1$ is not possible for polynomial-time algorithms in the considered setting. Our analytical framework handles the structured nonconvexity of the problem while utilizing memory-efficient dual-side extrapolation, making our approach well-suited for large-scale applications.

We remark here that the only prior work that handles both adversarial label noise and distributional shifts is the recent work Li et al. (2024). Compared to Li et al. (2024), our work applies to a different type of distributional shifts (group shifts in place of distributional ambiguity across all examples, which is better aligned with recent applications like Xie et al. (2023) and Xia et al. (2024)); it removes higher-moment loss assumptions and applies even when there is no penalization ($\nu = 0$); it is not restricted to χ^2-divergence; and it employs extrapolation (momentum) on the *dual* instead of the *primal* side, which has implications on ease of implementation as discussed later in the paper.

In the rest of the section, we introduce the necessary background to formally state our main result and provide a technical overview. A detailed discussion of related literature is provided in Appendix A.

1.1 Problem Setup

For two discrete probability vectors $\boldsymbol{\lambda}, \boldsymbol{\lambda}' \in \Delta_K$, the f-divergence is defined as $d_f(\boldsymbol{\lambda}, \boldsymbol{\lambda}') = \sum_{i=1}^{K} \lambda'_{[i]} f\big(\frac{\lambda_{[i]}}{\lambda'_{[i]}}\big)$, where $f : [0, \infty) \to \mathbb{R}_+$ is a convex function satisfying $f(1) = 0$. We focus on f-divergences for which the mapping $\boldsymbol{\lambda} \mapsto d_f(\boldsymbol{\lambda}, \boldsymbol{\lambda}_0)$ is *strongly convex* with respect to a relevant norm over the simplex Δ_K; these include some of the most commonly used examples like the χ^2-divergence and Kullback-Leibler (KL) divergence.

Same as Li et al. (2024), our analysis applies to a broad class of activations that are convex and (α, β)-unbounded.

Definition 1.1 (Unbounded Activation (Diakonikolas et al., 2022a))**.** A function $\sigma : \mathbb{R} \to \mathbb{R}$ is (α, β)-*unbounded* if it is non-decreasing and satisfies: (i) σ is β-Lipschitz continuous; (ii) for some $\alpha > 0$, $\sigma(t_1) - \sigma(t_2) \geq \alpha\,(t_1 - t_2)$ holds for all $t_1 \geq t_2 \geq 0$; (iii) $\sigma(0) = 0$.

To make this challenging nonconvex problem tractable, we build upon the approach of prior work on robust learning of a neuron (Wang et al., 2023; Li et al., 2024)—by imposing structural properties on the $\boldsymbol{x}$-marginal distributions across groups. The first of those assumptions is subexponential concentration of 1D projections:

Assumption 1.2 (Sub-Exponential Tails)**.** *There is a constant $B > 0$ such that for every $i \in [K]$ and every unit vector $\boldsymbol{u} \in \mathscr{B}(1)$, for all $r \geq 1$,*

$$\Pr_{\boldsymbol{x}\sim p_{\boldsymbol{x}[i]}}\big(|\boldsymbol{u} \cdot \boldsymbol{x}| \geq r\big) \leq \exp(-r/B).$$

Additionally, similar to Wang et al. (2023) and Li et al. (2024), we impose the following margin condition.

Assumption 1.3 (Uniform Margin)**.** *There exist constants $\zeta, \gamma \in (0,1]$ such that for every group $i \in [K]$ and every vector $\boldsymbol{w} \in \mathbb{R}^d$,*

$$\mathbb{E}_{\boldsymbol{x}\sim p_{\boldsymbol{x}[i]}}\big[\boldsymbol{x}\boldsymbol{x}^\top \mathbb{I}\{\boldsymbol{w}\cdot\boldsymbol{x} \geq \gamma\|\boldsymbol{w}\|_2\}\big] \succeq \zeta \boldsymbol{I}_d. \tag{1}$$

This means that the covariance matrix within the "margin region" $\{\boldsymbol{x} : \boldsymbol{w}\cdot\boldsymbol{x} \geq \gamma\|\boldsymbol{w}\|_2\}$ is uniformly well-conditioned across all groups and weight vectors $\boldsymbol{w}$. We remark here that (1) is required for *every* weight vector $\boldsymbol{w}$, making it a stronger condition than its counterparts in Wang et al. (2023) and Li et al. (2024), which only enforced such a condition for the target parameter $\boldsymbol{w}_*$. Nevertheless, this stronger condition still captures all well-concentrated distributions discussed in Wang et al. (2023) including well-behaved distributions from Diakonikolas et al. (2022a) (which include log-concave and s-concave distributions), discrete Gaussians, and the uniform distribution on $\{-1,0,1\}^d$, because the definitions of those distributions do not involve $\boldsymbol{w}_*$.

With these components, we can formally define our regularized Group DRO problem.

Definition 1.4 (Loss, Risk, and OPT_m)**.** Let $\mathscr{B}(W) = \{\boldsymbol{w} \in \mathbb{R}^d : \|\boldsymbol{w}\|_2 \leq W\}$ be the feasible set for weight vectors and let $p_{[1]}, \ldots, p_{[K]}$ be the K group distributions. For any $\boldsymbol{w} \in \mathscr{B}(W)$ and $\boldsymbol{\lambda} \in \Delta_K$, the regularized loss is:

$$\bar{L}(\boldsymbol{w}, \boldsymbol{\lambda}) = \sum_{i=1}^K \lambda_{[i]}\, \mathbb{E}_{(\boldsymbol{x},y)\sim p_{[i]}}\big(\sigma(\boldsymbol{w}\cdot\boldsymbol{x}) - y\big)^2 - \nu\, d_f\big(\boldsymbol{\lambda}, \tfrac{1}{K}\mathbf{1}\big).$$

The DRO risk for a given $\boldsymbol{w}$ is the maximum loss over all possible group weightings:

$$R(\boldsymbol{w}) = \max_{\boldsymbol{\lambda}\in\Delta_K} \bar{L}(\boldsymbol{w}, \boldsymbol{\lambda}). \tag{2}$$

Let $\boldsymbol{w}_* = \arg\min_{\boldsymbol{w}\in\mathscr{B}(W)} R(\boldsymbol{w})$ be the optimal weight vector and $\boldsymbol{\lambda}^*$ be its corresponding worst-case group weights. Our performance benchmark is the unregularized loss of $\boldsymbol{w}_*$ on its single worst-performing group:

$$\mathrm{OPT}_\mathrm{m} = \max_{i\in[K]} \mathbb{E}_{(\boldsymbol{x},y)\sim p_{[i]}}\big(\sigma(\boldsymbol{w}_*\cdot\boldsymbol{x}) - y\big)^2.$$

Problem 1.5 (Robust Learning a Group DRO Neuron)**.** Given a convex (α,β)-unbounded activation σ, error parameters $\epsilon, \delta \in (0,1)$, regularization parameter $\nu \geq 0$, weight radius $W > 0$, and sample access to labeled examples from each of the K group distributions $p_{[1]}, \ldots, p_{[K]}$, the goal is to output a parameter vector $\widehat{\boldsymbol{w}} \in \mathscr{B}(W)$ such that, with probability at least $1-\delta$, $\|\widehat{\boldsymbol{w}} - \boldsymbol{w}_*\|_2^2 \leq C\ \mathrm{OPT}_\mathrm{m} + \epsilon$ for a universal constant $C > 1$.

As we later argue (see Equation (10)), $\|\widehat{\boldsymbol{w}} - \boldsymbol{w}_*\|_2^2 \leq C\ \mathrm{OPT}_\mathrm{m} + \epsilon$ in turn implies a loss value bound of $\sum_{i\in[K]} \lambda^*_{[i]} \mathbb{E}_{p_{[i]}}(\sigma(\widehat{\boldsymbol{w}}\cdot\boldsymbol{x}) - y)^2 \leq C'\,\mathrm{OPT}_\mathrm{m} + \epsilon$, where $\boldsymbol{w}_*$, $\boldsymbol{\lambda}^*$, and OPT_m are from Definition 1.4, and $C' > 1$ is a universal constant.

1.2 Main Result

As stated earlier, our main contribution is an efficient algorithm for robust learning of a single neuron under adversarial label noise and group distributional shifts. Below, notation $\tilde{O}_c(\cdot)$ hides poly-logarithmic dependence in the argument and polynomial dependence in parameters c.

Theorem 1.6 (Informal; see Theorem 3.1)**.** *Suppose the underlying group distributions $p_{[i]}, i \in [K]$ satisfy appropriate margin and concentration properties (Assumptions 1.2 and 1.3) and the learner is provided with $\tilde{O}_{\beta,\nu,W}(\log K \log(1/\delta) d/\epsilon^2)$ samples from each of the $p_{[i]}, i \in [K]$. Then Algorithm 1, after $\tilde{O}_W(\min\{\log(1/\epsilon)\sqrt{1/\nu}, \sqrt{K}/\epsilon\})$ iterations, each running in near-linear time in the sample size, returns $\hat{\boldsymbol{w}} \in \mathscr{B}(W)$ that, with probability at least $1-\delta$, satisfies $\|\hat{\boldsymbol{w}} - \boldsymbol{w}_*\|_2^2 \leq C\ \mathrm{OPT}_\mathrm{m} + \epsilon$, where $C > 1$ is an absolute constant independent of d, K, W, ϵ, ν.*

Notably, our analysis shows that the algorithm attains sample complexity $\tilde{O}(Kd/\epsilon^2)$, which matches the known optimal rate for convex unregularized Group DRO problems (Soma et al., 2022; Zhang et al., 2023) up to log factors.

1.3 Technical Overview

Using standard uniform convergence arguments, we reduce the problem to solving its empirical version with a sufficiently large per-group sample size. The resulting problem is a challenging non-bilinearly coupled nonconvex-concave saddle-point problem over the empirical mixtures:

$$\min_{\boldsymbol{w}\in\mathscr{B}(W)} \max_{\widehat{\boldsymbol{\lambda}}\in\Delta_K} L(\boldsymbol{w}, \widehat{\boldsymbol{\lambda}}),$$

where $L(\boldsymbol{w}, \widehat{\boldsymbol{\lambda}}) = \sum_{i=1}^K \widehat{\lambda}_{[i]}\, \mathbb{E}_{(\boldsymbol{x},y)\sim\widehat{p}_{[i]}}\big(\sigma(\boldsymbol{w}\cdot\boldsymbol{x}) - y\big)^2 - \nu\, d_f\big(\widehat{\boldsymbol{\lambda}}, \frac{1}{K}\mathbf{1}\big)$ and $\widehat{p}_{[i]}$ represents the empirical distribution formed by samples drawn from group i.

Even for convex-concave objectives, the nonlinear-linear coupling between the primal and the dual as in our setting makes the analysis of primal-dual-style methods challenging and is still actively explored in current research; see Mehta et al. (2025) and references therein. On the other hand, existing approaches to min-max problems with a nonconvex objective over the primal variables like Lin et al. (2020), Zhang et al. (2020), Rafique et al. (2022), and Li et al. (2025) only offer stationarity guarantees, which are known to be insufficient for formal learning guarantees of a neuron even in much less challenging settings without distributional shifts (Yehudai and Shamir, 2020).

The most closely related work to ours, Li et al. (2024), handles a similar but distinct problem. They showed how the structured nonconvexity of learning a single neuron can be leveraged to obtain a provably convergent primal-dual algorithm for standard χ^2-divergence-based DRO. As in their

paper, to analyze progress towards the solution, we define a *gap function* that measures the suboptimality of an iterate pair $(\boldsymbol{w}_t, \widehat{\boldsymbol{\lambda}}_t)$ relative to a hybrid reference point $(\boldsymbol{w}_*, \widehat{\boldsymbol{\lambda}}^*)$, and track progress in the gap function per iteration involving one primal and one dual update.

A key technical challenge lies in bounding the primal gap due to the nonconvexity of the loss. Here, the analysis in Li et al. (2024) is insufficient for our goals. The analysis is limited to DRO with a χ^2 penalty, requires a large regularization parameter ν, and imposes fourth-moment assumptions on the loss. Our work extends this to the more practical *group* DRO setting, accommodates both KL and χ^2 divergences, and removes these restrictive assumptions.

From a practical standpoint, the primal extrapolation used in Li et al. (2024) is also undesirable. Primal extrapolation is memory-intensive, because the algorithm has to remember two previous primal variables $\boldsymbol{w} \in \mathbb{R}^d$ of dimension d to compute the extrapolation step. It is also unclear how to implement primal extrapolation together with widely used off-the-shelf solvers such as Adam (Kingma and Ba, 2015). Our algorithm instead performs extrapolation on the dual variable $\boldsymbol{\lambda} \in \mathbb{R}^K$. This is more efficient as typically $K \ll d$.

However, dual-side extrapolation requires a different analysis. A key difficulty is that the extrapolated dual vector is not guaranteed to have nonnegative entries. In typical analyses involving extrapolation steps, extrapolation is either performed before further bounding the objective (see, e.g., Chambolle and Pock (2011) and Kotsalis et al. (2022)) or it is done on both the primal and the dual side (Mehta et al., 2025), which in effect allows for the induced error terms to be telescoped or canceled out. Observe that the coupled term in the objective $L(\boldsymbol{w}, \widehat{\boldsymbol{\lambda}})$ is of the form $\widehat{\boldsymbol{\lambda}}^\top F(\boldsymbol{w})$, for a vector-valued mapping F whose each coordinate $(F(\boldsymbol{w}))_i$ is a nonconvex function of $\boldsymbol{w}$. Even if each $(F(\boldsymbol{w}))_i$ were convex (so that the mapping $\widehat{\boldsymbol{\lambda}}^\top F(\boldsymbol{w})$ is convex for feasible $\widehat{\boldsymbol{\lambda}}$), once we replace $\widehat{\boldsymbol{\lambda}}$ with an extrapolated vector $\bar{\boldsymbol{\lambda}}$ which is no longer guaranteed to be nonnegative, the mapping $\bar{\boldsymbol{\lambda}}^\top F(\boldsymbol{w})$ would no longer be guaranteed to be convex and so typical inequalities involving convexity would no longer apply to bounding $\bar{\boldsymbol{\lambda}}^\top F(\boldsymbol{w})$ below, which is needed in the analysis. We overcome this issue by first bounding $\widehat{\boldsymbol{\lambda}}^\top F(\boldsymbol{w})$ (using linearization, discussed next) and then applying extrapolation. This introduces nontrivial error terms that we bound by repeatedly leveraging the structural properties of the considered problem (see Appendix E.2).

Additionally, since in our case $(F(\boldsymbol{w}))_i$ is nonconvex, we need an appropriate strategy for bounding below $\widehat{\boldsymbol{\lambda}}^\top F(\boldsymbol{w})$. As mentioned earlier, it is possible to use the results from Li et al. (2024) leveraging structured nonconvexity of the neuron mean squared loss for this purpose. However, as already discussed, this would lead to a requirement for much stronger assumptions. Instead, we prove a key technical result ("linearization," Lemma 3.5), which leverages the specific group-wise structure of the problem to usefully bound below the mean squared loss function. This result has two significant advantages: (1) It avoids the complex, higher-order error terms that appear in prior work, which depend on the divergence between dual iterates. This simplifies the overall analysis considerably and allows us to address alternative divergences like KL. (2) It removes the requirement for a non-trivial lower bound on the regularization parameter ν. This allows our framework to handle the $\nu \to 0$ regime, smoothly connecting our robust formulation to classical, non-regularized Group DRO.

By combining this improved bound with sharpness properties of the loss function on the target mixture distribution p^*, we then carefully control the accumulated error terms and prove that the iterates converge to a solution competitive with the "best-fit" neuron parameterized by $\boldsymbol{w}_*$.

2 PRELIMINARIES

For a positive integer N, $[N] := \{1, \ldots, N\}$. If $\mathscr{E}$ is a subset of some ambient universe then $\mathscr{E}^c$ denotes its complement, and $\mathbb{I}_{\mathscr{E}}(x) = 1_{\{x \in \mathscr{E}\}}$ is its indicator function. For vectors $\boldsymbol{x}, \widehat{\boldsymbol{x}} \in \mathbb{R}^d$, $\langle \boldsymbol{x}, \widehat{\boldsymbol{x}} \rangle = \boldsymbol{x} \cdot \widehat{\boldsymbol{x}} = \boldsymbol{x}^\top \widehat{\boldsymbol{x}}$ is their inner product, and $\|\boldsymbol{x}\|_2$ is the ℓ_2 norm; we write $\boldsymbol{x} \le \widehat{\boldsymbol{x}}$ to mean $x^{(j)} \le \hat{x}^{(j)}$ coordinate-wise. $\boldsymbol{I}_d$ denotes the $d \times d$ identity matrix. $A \succeq B$ means that $A - B$ is positive semidefinite. The iteration index is denoted by t. The group weight vector at iteration t is $\boldsymbol{\lambda}_t = [\lambda_{t[1]}, \ldots, \lambda_{t[K]}]^\top \in \Delta_K$, where $\Delta_K := \{\boldsymbol{\lambda} \in \mathbb{R}^K : \sum_{i=1}^K \lambda_{t[i]} = 1, \lambda_{t[i]} \ge 0 \text{ for all } i \in [K]\}$ is the probability simplex in $\mathbb{R}^K$. For two distributions p and p', we use $p \ll p'$ to denote that p is absolutely continuous with respect to p', which means that for all measurable sets A, $p'(A) = 0$ implies $p(A) = 0$. For probability measures $p \ll p'$, we write $\frac{\mathrm{d}p}{\mathrm{d}p'}$ for the Radon-Nikodym derivative, and use $\chi^2(p, p') := \int (\frac{\mathrm{d}p}{\mathrm{d}p'} - 1)^2 \mathrm{d}p'$ and $\mathrm{KL}(p, p') := \int \log(\frac{\mathrm{d}p}{\mathrm{d}p'}) \mathrm{d}p$ to denote the chi-squared (χ^2) and KL divergences of p relative to p'.

We now state several facts used throughout our analysis. The population loss exhibits sharpness under our distributional assumptions (Wang et al., 2023):

Fact 2.1 (Population Sharpness and Moment Bounds (Wang et al., 2023))**.** *Let p satisfy Assumption 1.2. Define $c_0 := \frac{\gamma\zeta\alpha}{6B\log(20B/\zeta^2)}$. Then $\mathbb{E}_{\boldsymbol{x}\sim p_{\boldsymbol{x}[i]}}\big[\big(\sigma(\boldsymbol{w}\cdot\boldsymbol{x}) - \sigma(\boldsymbol{w}_*\cdot\boldsymbol{x})\big)(\boldsymbol{w}\cdot\boldsymbol{x} - \boldsymbol{w}_*\cdot\boldsymbol{x})\big] \ge c_0\|\boldsymbol{w} - \boldsymbol{w}_*\|_2^2$, and for any unit vector $\boldsymbol{u}$ and $\tau \in \{2, 4\}$, we have $\mathbb{E}_{\boldsymbol{x}\sim p_{\boldsymbol{x}[i]}}\big[(\boldsymbol{u}\cdot\boldsymbol{x})^\tau\big] \le 5B$.*

Uniform convergence arguments extend these bounds to empirical distributions (up to constants) when the per-group sample size N is sufficiently large:

Lemma 2.2 (Empirical Sharpness and Moment Bounds; Informal. See Lemma C.2)**.** *Under Assumptions 1.2 and 1.3, if the per-group sample size N/K is sufficiently*

large (dependent on $\beta, B, W, \nu, d, K, \delta$), then with high probability, for all groups $i \in [K]$, all $\boldsymbol{w} \in \mathcal{B}(3\|\boldsymbol{w}_*\|_2)$ with $\|\boldsymbol{w} - \boldsymbol{w}_*\|_2 \geq \sqrt{\epsilon}$, and any unit vector $\boldsymbol{u}$:

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

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

A direct consequence of Lemma 2.2, using Cauchy-Schwarz inequality and β-Lipschitzness of σ, is the following two-sided bound. For $c_1 := c_0^2/(24B)$ and any $\boldsymbol{w} \in \mathcal{B}(W)$:

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

Similar to prior work, we assume that the labels are bounded by a sufficiently large parameter $M = O(WB\beta\log(\beta BW/\epsilon))$. This assumption is without loss of generality (as established by the following fact) and can be ensured by simple pre-processing of labeled examples given to the algorithm. Thus, in the rest of the paper, we assume that the labels are bounded by M.

Fact 2.3 (Label Truncation (Wang et al., 2023))**.** *Under Assumptions 1.2 and 1.3, let $y' = \operatorname{sign}(y)\max\{|y|, M\}$ with $M = C_M WB\beta\log(\beta BW/\epsilon)$, for a sufficiently large constant C_M. Then for all $i \in [K]$ and all $\boldsymbol{w} \in \mathcal{B}(W)$, it holds that $\mathbb{E}_{p_{[i]}}(\sigma(\boldsymbol{w}\cdot\boldsymbol{x}) - y')^2 \leq \mathbb{E}_{p_{[i]}}(\sigma(\boldsymbol{w}\cdot\boldsymbol{x}) - y)^2 + \epsilon$.*

We recall the following first-order optimality condition:

Fact 2.4 (First-Order Optimality)**.** *If f is continuously differentiable on a closed convex set Ω and $\boldsymbol{x}^* \in \Omega$ is a local maximizer of f over Ω, then $\langle\nabla f(\boldsymbol{x}^*), \boldsymbol{y} - \boldsymbol{x}^*\rangle \leq 0$ for all $\boldsymbol{y} \in \Omega$. If f is also concave, this condition implies that $\boldsymbol{x}^*$ is a global maximizer.*

For a differentiable function $\phi : \mathbb{R}^N \to \mathbb{R}$, the Bregman divergence is $D_\phi(\boldsymbol{y}, \boldsymbol{x}) = \phi(\boldsymbol{y}) - \phi(\boldsymbol{x}) - \langle\nabla\phi(\boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x}\rangle$. Its invariance to affine addition is also useful:

Fact 2.5. *If $\psi(\boldsymbol{x}) = \phi(\boldsymbol{x}) + \langle\boldsymbol{a}, \boldsymbol{x}\rangle + b$, then $D_\psi(\boldsymbol{y}, \boldsymbol{x}) = D_\phi(\boldsymbol{y}, \boldsymbol{x})$ for all $\boldsymbol{x}, \boldsymbol{y}$.*

3 CONVERGENCE ANALYSIS

This section presents our primal-dual algorithm and establishes its convergence guarantees. For clarity, we first define the core notation used in the analysis. Let $\ell(\boldsymbol{w}; \boldsymbol{x}, y) := (\sigma(\boldsymbol{w}\cdot\boldsymbol{x}) - y)^2$ denote the squared loss. Since σ may not be differentiable, our algorithm relies on the *surrogate gradient* $\boldsymbol{v}(\boldsymbol{w}; \boldsymbol{x}, y) := 2\beta(\sigma(\boldsymbol{w}\cdot\boldsymbol{x}) - y)\boldsymbol{x}$, which serves as a well-behaved proxy for the true gradient, as in Kakade et al. (2011), Diakonikolas et al. (2020b), Wang et al. (2023), and Li et al. (2024). Our analysis measures performance against *empirical* benchmarks defined with respect to the *population-optimal* weight vector $\boldsymbol{w}_*$; this seeming mismatch is important to our convergence analysis. Define:

$$\widehat{\mathrm{OPT}_{\mathrm{m}}} := \max_{i\in[K]} \mathbb{E}_{(\boldsymbol{x},y)\sim\widehat{p}_{[i]}}[\ell(\boldsymbol{w}_*; \boldsymbol{x}, y)].$$

Let $\widehat{\boldsymbol{\lambda}}^* = \arg\max_{\widehat{\boldsymbol{\lambda}}\in\Delta_K} L(\boldsymbol{w}_*, \widehat{\boldsymbol{\lambda}})$ be the worst-case group weighting for $\boldsymbol{w}_*$ over the *empirical* distributions $\widehat{\boldsymbol{p}}_{[i]}$. We also write $\phi(\widehat{\boldsymbol{\lambda}}) := d_f(\widehat{\boldsymbol{\lambda}}, \frac{1}{K}\mathbf{1})$ for the divergence penalty.

Algorithm 1: A Primal-Dual Algorithm for Group DRO

Input: Sample sets $(\boldsymbol{x}_{[i]}^{(j)}, y_{[i]}^{(j)})_{j=1}^{N/K}$ for $i \in [K]$;
parameters $\nu \geq 0, W > 0, \epsilon > 0, \beta, B, c_1$.

Initialization: $\nu_0 = \epsilon/(4K)$,
$A_{-1} = a_{-1} = A_0 = a_0 = 0$, $\boldsymbol{w}_{-1} = \boldsymbol{w}_0 = \mathbf{0}$,
$\widehat{\boldsymbol{\lambda}}_{-1} = \widehat{\boldsymbol{\lambda}}_0 = \frac{\mathbf{1}}{K}$. Set constants C_4, C_W' based on (6).
for $t = 1, \ldots, n$ **do**

- $a_t = \min\{(1 + \frac{c_1}{8C_4})^{t-1}\frac{1}{4C_4}, \max\{(1 + \frac{\sqrt{c_1\nu}}{4\sqrt{2}C_W'})^{t-1}\frac{\sqrt{\nu_0}}{4C_W'}, \frac{c_1\nu_0}{(4\sqrt{2}C_W')^2}t\}\}$, $A_t = A_{t-1} + a_t$;
- $\boldsymbol{v}(\boldsymbol{w}; \boldsymbol{x}, y) = 2\beta(\sigma(\boldsymbol{w}\cdot\boldsymbol{x}) - y)\,\boldsymbol{x}$;
- $\bar{\boldsymbol{\lambda}}_{t-1} := \widehat{\boldsymbol{\lambda}}_{t-1} + \frac{a_{t-1}}{a_t}(\widehat{\boldsymbol{\lambda}}_{t-1} - \widehat{\boldsymbol{\lambda}}_{t-2})$;
- $\boldsymbol{w}_t := \arg\min_{\boldsymbol{w}} \{a_t \sum_{i=1}^K \bar{\lambda}_{t-1[i]}\mathbb{E}_{\widehat{p}_{[i]}}[\langle\boldsymbol{v}(\boldsymbol{w}_{t-1}; \boldsymbol{x}, y), \boldsymbol{w}\rangle] + \frac{1+0.5c_1A_t}{2}\|\boldsymbol{w} - \boldsymbol{w}_{t-1}\|_2^2\}$;
- $\widehat{\boldsymbol{\lambda}}_t := \arg\max_{\widehat{\boldsymbol{\lambda}}\in\Delta_K}\{a_t L(\boldsymbol{w}_t, \widehat{\boldsymbol{\lambda}}) - (\nu_0 + \nu A_{t-1})D_\phi(\widehat{\boldsymbol{\lambda}}, \widehat{\boldsymbol{\lambda}}_{t-1})\}$

Output: The final weight vector $\boldsymbol{w}_n$.

Our method, stated in Algorithm 1, is an iterative primal-dual algorithm that maintains iterates $(\boldsymbol{w}_t, \widehat{\boldsymbol{\lambda}}_t)$ and step sizes $a_t > 0$ with cumulative sums $A_t = \sum_{k=1}^t a_k$. Both updates are efficient, as they involve projections onto simple sets like the Euclidean ball $\mathcal{B}(W)$ and the probability simplex Δ_K, and/or minimization of KL divergence over the probability simplex, which is computable in closed form. A key algorithmic feature is the use of *extrapolation on the low-dimensional dual variable* $\widehat{\boldsymbol{\lambda}} \in \mathbb{R}^K$ to construct a gradient estimate for the primal update.

Before moving onto overviewing our convergence analysis, we state our main result.

3.1 Main Result

Our main result is summarized in the following theorem. We first define problem parameters (constants) that factor into the step size definition (consequently, iteration com-

plexity) and the constant factor approximation:

$$\begin{aligned} C_3 &:= 31\beta\sqrt{B}/c_1, \\ C_4 &:= 27c_1 + 2163\beta^4 B^2/c_1, \\ C'_W &:= 2\sqrt{3}\sqrt{6\beta^2 + C_M^2 B \log^2\Big(\frac{\beta BW}{\epsilon}\Big)}\beta W B. \end{aligned} \tag{6}$$

In (6), C_3 and C_4 can be treated as universal constants for typical single neuron learning problems. This is because, as argued in Wang et al. (2023), the sharpness constant c_1 is universal for any non-degenerate problem in the considered class. Similarly, for typical activations like ReLU, β is a constant (specifically, for the case of ReLU, $\alpha = \beta = 1$). The parameter B comes from one-dimensional concentration of projections of data vectors $\boldsymbol{x}$ (see Assumption 1.2); for all standard examples (like Gaussians, isotropic log-concaves, discrete Gaussians, uniform distribution on $\{-1, 0, 1\}^d$ discussed in Wang et al. (2023)), B is a small universal constant. The "constant" C'_W however cannot be treated as a universal constant as it depends on W. Fortunately, C'_W has no effect on the approximation ratio in the statement of Theorem 3.1—instead, it only affects the iteration complexity, which is polynomial in all problem parameters, as claimed in the introduction.

Theorem 3.1 (Main Theorem)**.** *Suppose the margin and concentration conditions (Assumptions 1.2 and 1.3) hold. Let c_1 be the sharpness constant from Lemma 2.2, and set $\nu_0 := \epsilon/(4K)$. If the total number of samples is $N = \tilde{O}_{\beta,B,\nu}\big(\frac{KW^4 d}{\epsilon^2}\log(\frac{1}{\delta})\big)$, then with probability at least $1-\delta$, the output $(\boldsymbol{w}_n, \widehat{\boldsymbol{\lambda}}_n)$ of Algorithm 1 satisfies the following inequality for all $n \geq 1$:*

$$\begin{aligned} &\frac{1 + 0.5c_1 A_n}{4}\|\boldsymbol{w}_n - \boldsymbol{w}_*\|_2^2 + (\nu_0 + \nu A_n) D_\phi(\widehat{\boldsymbol{\lambda}}_n, \widehat{\boldsymbol{\lambda}}^*) \\ &\leq D_0 + \frac{120\beta^2 B}{c_1} A_n(\mathrm{OPT_m} + \epsilon), \end{aligned} \tag{7}$$

where $D_0 = \frac{1}{2}\|\boldsymbol{w}_0 - \boldsymbol{w}_*\|_2^2 + \nu_0 D_\phi(\widehat{\boldsymbol{\lambda}}^*, \widehat{\boldsymbol{\lambda}}_0)$.

Furthermore, after a number of iterations n scaling as

$$\begin{aligned} n = O\Big(C'_W \min\{ &\tfrac{1}{\sqrt{c_1\nu}}\log\big(\tfrac{\|\boldsymbol{w}_*\|_2^2}{2\epsilon} + \tfrac{D_\phi(\widehat{\boldsymbol{\lambda}}^*, \widehat{\boldsymbol{\lambda}}_0)}{4K}\big), \\ &\sqrt{\tfrac{2K\|\boldsymbol{w}_*\|_2^2}{\epsilon^2} + \tfrac{D_\phi(\widehat{\boldsymbol{\lambda}}^*, \widehat{\boldsymbol{\lambda}}_0)}{\epsilon}}\,\}\Big) \end{aligned} \tag{8}$$

the output $\boldsymbol{w}_n$ is guaranteed to satisfy

$$\|\boldsymbol{w}_n - \boldsymbol{w}_*\|_2 \leq C_3\big(\sqrt{\mathrm{OPT_m}} + \sqrt{\epsilon}\big), \tag{9}$$

$$\begin{aligned} \mathbb{E}_{(\boldsymbol{x},y)\sim p^*}\big[\ell(\boldsymbol{w}_n; \boldsymbol{x}, y)\big] \leq \big(2 + 20B\beta^2 C_3^2\big)\ \mathrm{OPT_m} \\ + 20\beta^2 C_3^2 B\epsilon, \end{aligned} \tag{10}$$

where $p^* = \sum_{i=1}^K \lambda^*_{[i]} p_{[i]}$ is the worst-case population mixture and relevant constants are defined by (6)*.*

We remark here that we have made no attempt to optimize the constant factors in either the final approximation ratio or sample and iteration complexities—our focus was on establishing (any) constant factor approximation with a polynomial sample and computation algorithm, which already required highly technical arguments. We expect the absolute constant factors to be improvable.

3.2 Overview of the Analysis

The analysis centers on bounding the *primal-dual gap*-like function $\mathrm{Gap}(\boldsymbol{w}_t, \widehat{\boldsymbol{\lambda}}_t)$ defined with respect to the hybrid reference point $(\boldsymbol{w}_*, \widehat{\boldsymbol{\lambda}}^*)$ with population-optimal $\boldsymbol{w}_*$ and worst-case $\widehat{\boldsymbol{\lambda}}^*$ via

$$\mathrm{Gap}(\boldsymbol{w}_t, \widehat{\boldsymbol{\lambda}}_t) = L(\boldsymbol{w}_t, \widehat{\boldsymbol{\lambda}}^*) - L(\boldsymbol{w}_*, \widehat{\boldsymbol{\lambda}}_t).$$

Observe that $\mathrm{Gap}(\boldsymbol{w}_t, \widehat{\boldsymbol{\lambda}}_t)$ is the sum of the "primal gap" $L(\boldsymbol{w}_t, \widehat{\boldsymbol{\lambda}}^*) - L(\boldsymbol{w}_*, \widehat{\boldsymbol{\lambda}}^*)$ and the "dual gap" $L(\boldsymbol{w}_*, \widehat{\boldsymbol{\lambda}}^*) - L(\boldsymbol{w}_*, \widehat{\boldsymbol{\lambda}}_t)$. Because the loss is nonconvex in $\boldsymbol{w}$, the primal gap, and thus $\mathrm{Gap}(\boldsymbol{w}_t, \widehat{\boldsymbol{\lambda}}_t)$, is not necessarily nonnegative. Our strategy is to combine a lower bound on the gap derived from empirical sharpness with an upper bound derived from the analysis, which in turn motivates the algorithm's update rules.

We note here that while tracking similar gap functions is common to the analysis of primal-dual methods (see, e.g., Chambolle and Pock (2011) and Mehta et al. (2025)), here the central challenges come from the loss nonconvexity, which makes bounding the relevant quantities highly nontrivial. In the rest of this section, we first state (in Lemma 3.2) a lower bound on the gap function, which is crucial to being able to establish the contraction of distance to target solutions. This bound is derived following a similar argument to Li et al. (2024) and its proof is provided for completeness, in Appendix D.

The most challenging part of the analysis comes from bounding the gap function above, and, in particular, bounding below its dual gap component. The reason, as discussed in Section 1.3, is that the loss function is nonconvex and we perform extrapolation of the dual updates, to ensure our algorithm is compatible with large-scale implementations.

3.3 Gap Lower Bound

We begin the convergence analysis by establishing a lower bound on the gap function $\mathrm{Gap}(\boldsymbol{w}, \widehat{\boldsymbol{\lambda}})$.

Lemma 3.2 (Gap Lower Bound)**.** *Under Assumptions 1.2 and 1.3, if the per-group sample size N/K is sufficiently large (see Lemma C.2), for all $\boldsymbol{w} \in \mathcal{B}(3\|\boldsymbol{w}_*\|_2)$ and all $\widehat{\boldsymbol{\lambda}} \in \Delta_K$, we have $\mathrm{Gap}(\boldsymbol{w}, \widehat{\boldsymbol{\lambda}}) \geq -\frac{12\beta^2 B}{c_1}\widehat{\mathrm{OPT_m}} + \frac{c_1}{2}\|\boldsymbol{w} - \boldsymbol{w}_*\|_2^2 + \nu D_\phi(\widehat{\boldsymbol{\lambda}}^*, \widehat{\boldsymbol{\lambda}})$.*

The application of empirical sharpness, which is crucial to establishing the gap lower bound in Lemma 3.2, is predicated on the iterates $\boldsymbol{w}_t$ remaining in a neighborhood of a

target parameter vector $\boldsymbol{w}_*$. We establish this property by induction. The proof is provided in Appendix E.3.

Lemma 3.3. *For all iterations $t \geq 0$ of Algorithm 1, the iterates satisfy $\|\boldsymbol{w}_t\|_2 \leq 3\|\boldsymbol{w}_*\|_2$.*

3.4 Gap Upper Bound

Having obtained a lower bound on the gap function, we now derive a corresponding upper bound based on our algorithm updates. The core of the argument is to analyze the per-iteration progress of Algorithm 1 to construct a telescoping sum. The main technical result is the following proposition, with the full proof deferred to Appendix E.

Proposition 3.4 (Gap Upper Bound)**.** *Let the sequences $\{a_t\}$, $\{A_t\}$, $\{\boldsymbol{w}_t\}$, and $\{\widehat{\boldsymbol{\lambda}}_t\}$ be generated by Algorithm 1, where, by convention, $a_{-1} = a_0 = A_{-1} = A_0 = 0$, $\boldsymbol{w}_{-1} = \boldsymbol{w}_0 = \mathbf{0}$, and $\widehat{\boldsymbol{\lambda}}_{-1} = \widehat{\boldsymbol{\lambda}}_0 = \frac{1}{K}\mathbf{1}$. Under Assumptions 1.2 and 1.3, if the per-group sample size N/K is sufficiently large (see Lemma C.2), then for any $n \geq 1$:*

$$\begin{aligned}
&\sum_{t=1}^{n} a_t \mathrm{Gap}(\boldsymbol{w}_t, \widehat{\boldsymbol{\lambda}}_t) \\
&\leq \frac{1}{2}\|\boldsymbol{w}_* - \boldsymbol{w}_0\|_2^2 - \frac{1 + 0.5c_1A_n}{4}\|\boldsymbol{w}_* - \boldsymbol{w}_n\|_2^2 \\
&\quad + \nu_0 D_\phi(\widehat{\boldsymbol{\lambda}}^*, \widehat{\boldsymbol{\lambda}}_0) - \frac{1 + 0.5c_1A_n}{4}\|\boldsymbol{w}_n - \boldsymbol{w}_{n-1}\|_2^2 \\
&\quad - (\nu_0 + \nu A_n) D_\phi(\widehat{\boldsymbol{\lambda}}^*, \widehat{\boldsymbol{\lambda}}_n) + \frac{28\beta^2 B}{c_1} A_n \widehat{\mathrm{OPT}_{\mathrm{m}}}.
\end{aligned}$$

The proof of this result is highly technical, requiring careful handling of several nontrivial error terms induced by the extrapolation of the dual update and by the nonconvexity of the loss. For this reason, it is deferred to Appendix E. We however highlight a key structural result used for nontrivially bounding below the expected loss, in the following lemma. We also provide the full proof of the lemma below.

Lemma 3.5 (Linearization)**.** *For each group $i \in [K]$ and each iteration $t \in [n]$, the following bound holds:*

$$\begin{aligned}
&\mathbb{E}_{\widehat{p}_{[i]}}\big[2(\sigma(\boldsymbol{w}_t \cdot \boldsymbol{x}) - y)(\sigma(\boldsymbol{w}_* \cdot \boldsymbol{x}) - \sigma(\boldsymbol{w}_t \cdot \boldsymbol{x}))\big] \\
&\geq \mathbb{E}_{\widehat{p}_{[i]}}\big[\langle \boldsymbol{v}(\boldsymbol{w}_t; \boldsymbol{x}, y), \boldsymbol{w}_* - \boldsymbol{w}_t\rangle\big] - E_t,
\end{aligned}$$

where the error term is $E_t := \frac{24\beta^2 B\widehat{\mathrm{OPT}_{\mathrm{m}}}}{c_1} + \frac{c_1}{4}\|\boldsymbol{w}_* - \boldsymbol{w}_t\|_2^2$.

Proof. We split the expectation on the left hand side according to whether $\sigma(\boldsymbol{w}_t \cdot \boldsymbol{x}) - y \geq 0$. Define $e_t(\boldsymbol{x}, y) = \sigma(\boldsymbol{w}_t \cdot \boldsymbol{x}) - y$, $e_*(\boldsymbol{x}, y) = \sigma(\boldsymbol{w}_* \cdot \boldsymbol{x}) - y$, and $\mathcal{G} = \{(\boldsymbol{x}, y) \mid e_t(\boldsymbol{x}, y) \geq 0\}$. Then

$$\begin{aligned}
&\mathrm{LHS}_i := \mathbb{E}_{\widehat{p}_{[i]}}\big[2e_t(\boldsymbol{x}, y)\,(\sigma(\boldsymbol{w}_* \cdot \boldsymbol{x}) - \sigma(\boldsymbol{w}_t \cdot \boldsymbol{x}))\big] \\
&= \mathbb{E}_{\widehat{p}_{[i]}}\Big[2e_t(\boldsymbol{x}, y)\,(\sigma(\boldsymbol{w}_* \cdot \boldsymbol{x}) - \sigma(\boldsymbol{w}_t \cdot \boldsymbol{x}))\mathbb{I}_{\mathcal{G}}\Big] \\
&\quad + \mathbb{E}_{\widehat{p}_{[i]}}\Big[2e_t(\boldsymbol{x}, y)\,(\sigma(\boldsymbol{w}_* \cdot \boldsymbol{x}) - \sigma(\boldsymbol{w}_t \cdot \boldsymbol{x}))\mathbb{I}_{\mathcal{G}^c}\Big].
\end{aligned}$$

On $\mathcal{G}$, we apply convexity of σ at $\boldsymbol{w}_t \cdot \boldsymbol{x}$:

$$\sigma(\boldsymbol{w}_* \cdot \boldsymbol{x}) - \sigma(\boldsymbol{w}_t \cdot \boldsymbol{x}) \;\geq\; \sigma'(\boldsymbol{w}_t \cdot \boldsymbol{x})\,(\boldsymbol{w}_* \cdot \boldsymbol{x} - \boldsymbol{w}_t \cdot \boldsymbol{x}),$$

and on $\mathcal{G}^c$ we similarly use convexity at $\boldsymbol{w}_* \cdot \boldsymbol{x}$ after flipping signs, where σ' is the subderivative of σ which must exist due to convexity and Lipschitzness. Hence

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

Now recall $\boldsymbol{v}(\boldsymbol{w}_t; \boldsymbol{x}, y) = 2\beta e_t(\boldsymbol{x}, y)\boldsymbol{x}$. Adding and subtracting β in the derivatives, we rewrite the left hand side:

$$\begin{aligned}
&\mathbb{E}_{\widehat{p}_{[i]}}[2e_t(\boldsymbol{x}, y)(\sigma(\boldsymbol{w}_* \cdot \boldsymbol{x}) - \sigma(\boldsymbol{w}_t \cdot \boldsymbol{x}))] \\
&\geq \mathbb{E}_{\widehat{p}_{[i]}}[\langle \boldsymbol{v}(\boldsymbol{w}; \boldsymbol{x}, y), \boldsymbol{w}_* - \boldsymbol{w}_t\rangle] \\
&\quad + 2\mathbb{E}_{\widehat{p}_{[i]}}[e_t(\boldsymbol{x}, y)(\sigma'(\boldsymbol{w}_t \cdot \boldsymbol{x}) - \beta)(\boldsymbol{w}_* \cdot \boldsymbol{x} - \boldsymbol{w}_t \cdot \boldsymbol{x})\mathbb{I}_{\mathcal{G}}] \\
&\quad + 2\mathbb{E}_{\widehat{p}_{[i]}}[e_t(\boldsymbol{x}, y)(\sigma'(\boldsymbol{w}_* \cdot \boldsymbol{x}) - \beta)(\boldsymbol{w}_* \cdot \boldsymbol{x} - \boldsymbol{w}_t \cdot \boldsymbol{x})\mathbb{I}_{\mathcal{G}^c}].
\end{aligned} \tag{11}$$

where we consider the last two terms involving factors $(\sigma'(\cdot) - \beta)$ to be the "error terms" . Since σ is nondecreasing and β-Lipschitz, $-\beta \leq \sigma'(\cdot) - \beta \leq 0$. We show that both error terms are bounded in absolute value by $\beta\,\mathbb{E}_{\widehat{p}_{[i]}}\big[\,|e_*(\boldsymbol{x}, y)|\,|\boldsymbol{w}_* \cdot \boldsymbol{x} - \boldsymbol{w}_t \cdot \boldsymbol{x}|\big]$. On the event $\mathcal{G}$,

$$\begin{aligned}
&e_t(\boldsymbol{x}, y)(\sigma'(\boldsymbol{w}_t \cdot \boldsymbol{x}) - \beta)(\boldsymbol{w}_* \cdot \boldsymbol{x} - \boldsymbol{w}_t \cdot \boldsymbol{x}) \\
&= e_*(\boldsymbol{x}, y)(\sigma'(\boldsymbol{w}_t \cdot \boldsymbol{x}) - \beta)(\boldsymbol{w}_* \cdot \boldsymbol{x} - \boldsymbol{w}_t \cdot \boldsymbol{x}) \\
&+ (e_t(\boldsymbol{x}, y) - e_*(\boldsymbol{x}, y))(\sigma'(\boldsymbol{w}_t \cdot \boldsymbol{x}) - \beta)(\boldsymbol{w}_* \cdot \boldsymbol{x} - \boldsymbol{w}_t \cdot \boldsymbol{x}) \\
&\geq e_*(\boldsymbol{x}, y)(\sigma'(\boldsymbol{w}_t \cdot \boldsymbol{x}) - \beta)(\boldsymbol{w}_* \cdot \boldsymbol{x} - \boldsymbol{w}_t \cdot \boldsymbol{x}).
\end{aligned} \tag{12}$$

The last inequality holds because $(e_t(\boldsymbol{x}, y) - e_*(\boldsymbol{x}, y))(\boldsymbol{w}_* \cdot \boldsymbol{x} - \boldsymbol{w}_t \cdot \boldsymbol{x}) \leq 0$ by monotonicity of σ. Similarly, we have

$$\begin{aligned}
&e_t(\boldsymbol{x}, y)(\sigma'(\boldsymbol{w}_* \cdot \boldsymbol{x}) - \beta)(\boldsymbol{w}_* \cdot \boldsymbol{x} - \boldsymbol{w}_t \cdot \boldsymbol{x}) \\
&\geq e_*(\boldsymbol{x}, y)(\sigma'(\boldsymbol{w}_* \cdot \boldsymbol{x}) - \beta)(\boldsymbol{w}_* \cdot \boldsymbol{x} - \boldsymbol{w}_t \cdot \boldsymbol{x}).
\end{aligned} \tag{13}$$

Plugging Equation (12) and Equation (13) into Equation (11), we have

$$\begin{aligned}
&\mathbb{E}_{\widehat{p}_{[i]}}[2e_t(\boldsymbol{x}, y)(\sigma(\boldsymbol{w}_* \cdot \boldsymbol{x}) - \sigma(\boldsymbol{w}_t \cdot \boldsymbol{x}))] \\
&\geq \mathbb{E}_{\widehat{p}_{[i]}}[\langle \boldsymbol{v}(\boldsymbol{w}; \boldsymbol{x}, y), \boldsymbol{w}_* - \boldsymbol{w}_t\rangle] \\
&\quad + 2\mathbb{E}_{\widehat{p}_{[i]}}[e_*(\boldsymbol{x}, y)(\sigma'(\boldsymbol{w}_t \cdot \boldsymbol{x}) - \beta)(\boldsymbol{w}_* \cdot \boldsymbol{x} - \boldsymbol{w}_t \cdot \boldsymbol{x})\mathbb{I}_{\mathcal{G}}] \\
&\quad + 2\mathbb{E}_{\widehat{p}_{[i]}}[e_*(\boldsymbol{x}, y)(\sigma'(\boldsymbol{w}_* \cdot \boldsymbol{x}) - \beta)(\boldsymbol{w}_* \cdot \boldsymbol{x} - \boldsymbol{w}_t \cdot \boldsymbol{x})\mathbb{I}_{\mathcal{G}^c}].
\end{aligned} \tag{14}$$

Taking absolute values and using $-\beta \leq \sigma'(\cdot) - \beta \leq 0$, for any $\boldsymbol{w} \in \{\boldsymbol{w}_t, \boldsymbol{w}_*\}$, we have

$$\begin{aligned}
&-e_*(\boldsymbol{x}, y)(\sigma'(\boldsymbol{w} \cdot \boldsymbol{x}) - \beta)(\boldsymbol{w}_* \cdot \boldsymbol{x} - \boldsymbol{w}_t \cdot \boldsymbol{x}) \\
&\leq |e_*(\boldsymbol{x}, y)(\sigma'(\boldsymbol{w} \cdot \boldsymbol{x}) - \beta)(\boldsymbol{w}_* \cdot \boldsymbol{x} - \boldsymbol{w}_t \cdot \boldsymbol{x})| \\
&\leq \beta|e_*(\boldsymbol{x}, y)(\boldsymbol{w}_* \cdot \boldsymbol{x} - \boldsymbol{w}_t \cdot \boldsymbol{x})|,
\end{aligned} \tag{15}$$

Therefore, plugging Equation (15) into Equation (14),

$$\begin{aligned}
&\mathbb{E}_{\widehat{p}_{[i]}}[2e_t(\boldsymbol{x},y)(\sigma(\boldsymbol{w}_*\cdot\boldsymbol{x})-\sigma(\boldsymbol{w}_t\cdot\boldsymbol{x}))]\\
&\geq \mathbb{E}_{\widehat{p}_{[i]}}[\langle \boldsymbol{v}(\boldsymbol{w};\boldsymbol{x},y),\boldsymbol{w}_*-\boldsymbol{w}_t\rangle]\\
&-2\mathbb{E}_{\widehat{p}_{[i]}}[|e_*(\boldsymbol{x},y)(\sigma'(\boldsymbol{w}_t\cdot\boldsymbol{x})-\beta)(\boldsymbol{w}_*\cdot\boldsymbol{x}-\boldsymbol{w}_t\cdot\boldsymbol{x})|\mathbb{I}_{\mathcal{G}}]\\
&-2\mathbb{E}_{\widehat{p}_{[i]}}[|e_*(\boldsymbol{x},y)(\sigma'(\boldsymbol{w}_*\cdot\boldsymbol{x})-\beta)(\boldsymbol{w}_*\cdot\boldsymbol{x}-\boldsymbol{w}_t\cdot\boldsymbol{x})|\mathbb{I}_{\mathcal{G}^c}]\\
&\geq \mathbb{E}_{\widehat{p}_{[i]}}[\langle \boldsymbol{v}(\boldsymbol{w};\boldsymbol{x},y),\boldsymbol{w}_*-\boldsymbol{w}_t\rangle]\\
&\quad-2\beta\mathbb{E}_{\widehat{p}_{[i]}}[|e_*(\boldsymbol{x},y)(\boldsymbol{w}_*\cdot\boldsymbol{x}-\boldsymbol{w}_t\cdot\boldsymbol{x})|\mathbb{I}_{\mathcal{G}}]\\
&\quad-2\beta\mathbb{E}_{\widehat{p}_{[i]}}[|e_*(\boldsymbol{x},y)(\boldsymbol{w}_*\cdot\boldsymbol{x}-\boldsymbol{w}_t\cdot\boldsymbol{x})|\mathbb{I}_{\mathcal{G}^c}]\\
&\geq \mathbb{E}_{\widehat{p}_{[i]}}[\langle \boldsymbol{v}(\boldsymbol{w};\boldsymbol{x},y),\boldsymbol{w}_*-\boldsymbol{w}_t\rangle]\\
&\quad-2\beta\mathbb{E}_{\widehat{p}_{[i]}}[(\mathbb{I}_{\mathcal{G}}+\mathbb{I}_{\mathcal{G}^c})|e_*(\boldsymbol{x},y)(\boldsymbol{w}_*\cdot\boldsymbol{x}-\boldsymbol{w}_t\cdot\boldsymbol{x})|]\\
&= \mathbb{E}_{\widehat{p}_{[i]}}[\langle \boldsymbol{v}(\boldsymbol{w};\boldsymbol{x},y),\boldsymbol{w}_*-\boldsymbol{w}_t\rangle]\\
&\quad-2\beta\mathbb{E}_{\widehat{p}_{[i]}}[|e_*(\boldsymbol{x},y)(\boldsymbol{w}_*\cdot\boldsymbol{x}-\boldsymbol{w}_t\cdot\boldsymbol{x})|].
\end{aligned}$$

We apply Cauchy-Schwarz and Equation (4) to get,

$$\begin{aligned}
&\mathbb{E}_{\widehat{p}_{[i]}}\big[\,|e_*(\boldsymbol{x},y)|\,|\boldsymbol{w}_*\cdot\boldsymbol{x}-\boldsymbol{w}_t\cdot\boldsymbol{x}|\big]\\
&\leq \sqrt{\mathbb{E}_{\widehat{p}_{[i]}}[e_*(\boldsymbol{x},y)^2]}\,\sqrt{\mathbb{E}_{\widehat{p}_{[i]}}[(\boldsymbol{w}_*\cdot\boldsymbol{x}-\boldsymbol{w}_t\cdot\boldsymbol{x})^2]}\\
&\leq \sqrt{\widehat{\mathrm{OPT}_{\mathrm{m}}}}\,\sqrt{6B}\,\|\boldsymbol{w}_*-\boldsymbol{w}_t\|_2.
\end{aligned}$$

Young's inequality (Fact B.1) then gives

$$\begin{aligned}
&\sqrt{\widehat{\mathrm{OPT}_{\mathrm{m}}}}\,\sqrt{6B}\,\|\boldsymbol{w}_*-\boldsymbol{w}_t\|_2\\
&\leq \frac{12\beta B}{c_1}\,\widehat{\mathrm{OPT}_{\mathrm{m}}}+\frac{c_1}{8\beta}\|\boldsymbol{w}_*-\boldsymbol{w}_t\|_2^2.
\end{aligned}$$

Combining these estimates, for each $i\in[K]$,

$$\begin{aligned}
\mathrm{LHS}_i &\geq \mathbb{E}_{\widehat{p}_{[i]}}[\langle \boldsymbol{v}(\boldsymbol{w}_t;\boldsymbol{x},y),\,\boldsymbol{w}_*-\boldsymbol{w}_t\rangle]\\
&\quad-\Big(\frac{24\beta^2B}{c_1}\,\widehat{\mathrm{OPT}_{\mathrm{m}}}+\frac{c_1}{4}\|\boldsymbol{w}_*-\boldsymbol{w}_t\|_2^2\Big),
\end{aligned}$$

which completes the proof. □

3.5 Proof of Main Theorem

Proof of Theorem 3.1. The proof proceeds by combining the lower and upper bounds on the cumulative gap. Summing the per-iteration lower bound from Lemma 3.2 from $t=1$ to n and combining it with the upper bound from Proposition 3.4 gives the following inequality:

$$\begin{aligned}
&\frac{1+0.5\,c_1A_n}{4}\|\boldsymbol{w}_*-\boldsymbol{w}_n\|_2^2+\frac{c_1}{2}\sum_{t=1}^n a_t\|\boldsymbol{w}_t-\boldsymbol{w}_*\|_2^2\\
&+\nu\sum_{t=1}^n a_t\,D_\phi(\widehat{\boldsymbol{\lambda}}_t,\widehat{\boldsymbol{\lambda}}^*)+(\nu_0+\nu A_n)\,D_\phi(\widehat{\boldsymbol{\lambda}}^*,\widehat{\boldsymbol{\lambda}}_n)\\
&+\frac{1+0.5c_1A_n}{4}\|\boldsymbol{w}_n-\boldsymbol{w}_{n-1}\|_2^2\\
&\leq \tfrac{1}{2}\|\boldsymbol{w}_*-\boldsymbol{w}_0\|_2^2+\nu_0\,D_\phi(\widehat{\boldsymbol{\lambda}}^*,\widehat{\boldsymbol{\lambda}}_0)+\frac{40\,\beta^2B}{c_1}\,\widehat{\mathrm{OPT}_{\mathrm{m}}}\,A_n.
\end{aligned}$$

Dropping the nonnegative terms that involve $\|\boldsymbol{w}_t-\boldsymbol{w}_*\|_2^2$, $D_\phi(\widehat{\boldsymbol{\lambda}}_t,\widehat{\boldsymbol{\lambda}}^*)$, and $\|\boldsymbol{w}_n-\boldsymbol{w}_{n-1}\|_2^2$ from LHS of the inequality above and combining with the high-probability bound $\widehat{\mathrm{OPT}_{\mathrm{m}}}\leq 3(\mathrm{OPT}_{\mathrm{m}}+\epsilon)$ we establish in Lemma C.3, we arrive at the first result of the main theorem in Equation (7).

Next, we derive the explicit convergence guarantees. From (7), we can isolate the primal error:

$$\begin{aligned}
\|\boldsymbol{w}_n-\boldsymbol{w}_*\|_2^2 &\leq \frac{4D_0}{1+0.5c_1A_n}+\frac{160\beta^2BA_n}{c_1(1+0.5c_1A_n)}\widehat{\mathrm{OPT}_{\mathrm{m}}}\\
&\leq \frac{4D_0}{1+0.5c_1A_n}+\frac{320\beta^2B}{c_1^2}\widehat{\mathrm{OPT}_{\mathrm{m}}}.
\end{aligned}$$

The step-size schedule in Algorithm 1 is chosen so that for the number of iterations n specified in the theorem, we have $A_n=\Omega(c_1(W^2+K)/\epsilon)=\Omega(c_1D_0/\epsilon)$. This makes the first term $O(\epsilon)$. Substituting the bound on $\widehat{\mathrm{OPT}_{\mathrm{m}}}$ and taking the square root establishes the distance guarantee (9). Finally, to bound the risk (10), we use the decomposition $\ell(\boldsymbol{w}_n;\boldsymbol{x},y)\leq 2\ell(\boldsymbol{w}_*;\boldsymbol{x},y)+2(\sigma(\boldsymbol{w}_n\cdot\boldsymbol{x})-\sigma(\boldsymbol{w}_*\cdot\boldsymbol{x}))^2$, which follows by an application of Young's inequality. Taking the expectation over the worst-case population distribution p^* then leads to

$$\begin{aligned}
&\mathbb{E}_{p^*}[\ell(\boldsymbol{w}_n;\boldsymbol{x},y)]\\
&\leq 2\mathbb{E}_{p^*}[\ell(\boldsymbol{w}_*;\boldsymbol{x},y)]+2\mathbb{E}_{p^*}[(\sigma(\boldsymbol{w}_n\cdot\boldsymbol{x})-\sigma(\boldsymbol{w}_*\cdot\boldsymbol{x}))^2]\\
&\leq 2\,\mathrm{OPT}_{\mathrm{m}}\\
&\quad+2\beta^2\|\boldsymbol{w}_n-\boldsymbol{w}_*\|_2^2\cdot\max_{i\in[K]}\mathbb{E}_{p_{[i]}}\Big[\Big(\frac{\boldsymbol{w}_n-\boldsymbol{w}_*}{\|\boldsymbol{w}_n-\boldsymbol{w}_*\|_2}\cdot\boldsymbol{x}\Big)^2\Big]\\
&\leq 2\,\mathrm{OPT}_{\mathrm{m}}+2\beta^2\cdot 2C_3^2(\mathrm{OPT}_{\mathrm{m}}+\epsilon)\cdot 5B,
\end{aligned}$$

where the last line uses the previously established bound (9) on $\|\boldsymbol{w}_n-\boldsymbol{w}_*\|_2^2$ and the moment bound from Fact 2.1. Rearranging terms yields the final result (10). □

4 EXPERIMENTS

While our main contributions are theoretical, we provide illustrative experiments to demonstrate that our core algorithmic idea—regularized dual-extrapolated reweighting—can be applied to large-scale language model pre-training. These results are not intended as comprehensive empirical validation, but rather serve as a preliminary exploration of how our theoretically-motivated update rule performs in practice.

Setup We isolate the impact of our reweighting algorithm by integrating our primal-dual method with KL-divergence regularization (**PD-KL**) directly into the Sheared LLaMA framework (Xia et al., 2024). Starting from the `Sheared-LLaMA-1.3B` model—a 1.3B parameter version pruned from LLaMA2-7B (Xia et al., 2024)—we continue pre-training on RedPajama (Together

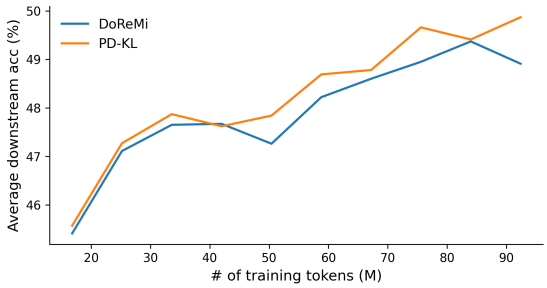

Figure 1: Compute-performance curve on *Sheared-LLaMA-1.3B*. Y-axis is the unweighted overall accuracy scores, X-axis is the number of tokens trained.

Computer, 2023). Our baseline is the dynamic batch loading algorithm from Xia et al. (2024), where we replace only its exponential ascent rule with our dual update (Algorithm 1) that incorporates KL regularization and extrapolation. All other training parameters remain identical to ensure fair comparison. More training details are included in Appendix F.

Performance is evaluated on the same suite of 11 downstream tasks used in Xia et al. (2024, Table 2), under their prescribed few-shot settings. We report unweighted average accuracy across tasks at multiple checkpoints. By design, this setup attributes any performance differences directly to the choice of on-the-fly domain reweighting strategy. Results are presented from 16.8M to 92.4M tokens.

4.1 Main results

We compare our PD-KL algorithm against DoReMi (Xie et al., 2023)—which is the training method used on the Sheared LLaMA model in Xia et al. (2024). As shown in Figure 1, our PD-KL algorithm demonstrates a consistent performance improvement over the baseline dynamic batch loading across the training trajectory. At most checkpoints, PD-KL achieves a higher average downstream accuracy varying from 0.04% to 0.96%. Moreover, at 33.6M tokens our PD-KL model reaches 47.87% average accuracy. DoReMi requires at least 1.5x more training time to achieve the same accuracy. This suggests that the principles of regularization and dual-side extrapolation, motivated by our theory, can offer practical benefits for training stability and downstream generalization even in highly complex models.

Per-task breakdown. For transparency, Table 1 lists every task used in evaluation, its metric, and per-task scores for both models. To remain faithful to the reference and avoid choice-induced bias, we keep the same per-task metric (e.g., `acc` vs. `acc_norm`) and the same shot setting as in Xia et al. (2024), with one exception: following community conventions in large language modeling, we evaluate **ARC-E** using `acc_norm` rather than `acc`.

Table 1: Per-task results (%) at 92.4M tokens.

Bucket	Task	Metric	DoReMi	PD-KL
Commonsense & RC	**ARC-E**	`acc_norm`	**50.00**	49.83
Commonsense & RC	**ARC-C(25)**	`acc_norm`	29.95	**30.38**
Commonsense & RC	**HellaSwag(10)**	`acc_norm`	**54.78**	54.62
Commonsense & RC	**PICA**	`acc`	70.78	**71.87**
Commonsense & RC	**SciQ**	`acc`	85.00	**85.90**
Commonsense & RC	**WinoGrande**	`acc`	54.22	**55.25**
Commonsense & RC	**WSC**	`acc`	36.54	36.54
Continued & LM	**BoolQ(32)**	`acc`	56.39	**63.64**
Continued & LM	**LogicQA**	`acc_norm`	**28.11**	27.80
Continued & LM	**LAMBADA**	`acc`	48.61	**50.63**
World Knowledge	**TruthfulQA(5)**	`acc`	**23.62**	22.15
Unweighted Mean			48.91	**49.87**

5 CONCLUSION

We studied the problem of robustly learning a single neuron in the group distributionally robust settings, where labels can be arbitrary and the goal is to be competitive with the "best-fit" model, as measured by the mean squared loss on the worst case group reweighting. The dual updates of our primal-dual approach can be seen as a novel group/domain reweighting. We hope that our work will encourage more research in this area. Possible future work includes extending either primal-dual updates or only dual updates to popular applications of language model pretraining and comparing whether it is competitive with current state-of-the-art domain reweighting algorithms. On the more theoretical side, it would be interesting to strengthen the error guarantee to a constant factor error in terms of OPT rather than $\mathrm{OPT}_{\mathrm{m}}$ (for $\nu \gg 0$, since $\mathrm{OPT} = \mathrm{OPT}_{\mathrm{m}}$ for $\nu = 0$) and to generalize the technical approach to other related settings involving nonconvex risk minimization.

Acknowledgments

Guyang Cao was supported in part by NSF CAREER Award CCF-2440563. Shuyao Li was supported in part by AFOSR Awards FA9550-21-1-0084 and FA9550-24-1-0076, and by NSF CAREER Award CCF-2440563. Jelena Diakonikolas was supported in part by the Air Force Office of Scientific Research under award number FA9550-24-1-0076, NSF CAREER Award CCF-2440563, and NSF MFAI Award DMS-2502282. Any opinions, findings, conclusions, or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the U.S. Department of Defense.

References

Agarwal, Alekh and Zhang, Tong (2022). "Minimax regret optimization for robust machine learning under distribution shift". In: *Conference on Learning Theory*. PMLR.

Ben-David, Shai et al. (2010). "A theory of learning from different domains". In: *Machine learning* 79.

Ben-Tal, Aharon, El Ghaoui, Laurent, and Nemirovski, Arkadi (2009). *Robust optimization*. Vol. 28. Princeton university press.

Bickel, Steffen, Brückner, Michael, and Scheffer, Tobias (2007). "Discriminative learning for differing training and test distributions". In: *Proceedings of the 24th international conference on Machine learning*.

Blanchet, Jose, Murthy, Karthyek, and Nguyen, Viet Anh (2021). "Statistical analysis of Wasserstein distributionally robust estimators". In: *Tutorials in Operations Research: Emerging Optimization Methods and Modeling Techniques with Applications*. INFORMS.

Chambolle, Antonin and Pock, Thomas (2011). "A first-order primal-dual algorithm for convex problems with applications to imaging". In: *Journal of mathematical imaging and vision* 40.

Chen, Ruidi and Paschalidis, Ioannis Ch (2018). "A robust learning approach for regression models based on distributionally robust optimization". In: *Journal of Machine Learning Research* 19.13.

Chen, Sitan et al. (2020). "Classification under misspecification: Halfspaces, generalized linear models, and evolvability". In: *Advances in Neural Information Processing Systems* 33.

Diakonikolas, I. et al. (2022a). "Learning a Single Neuron with Adversarial Label Noise via Gradient Descent". In: *Conference on Learning Theory (COLT)*, pp. 4313–4361.

Diakonikolas, Ilias, Gouleakis, Themis, and Tzamos, Christos (2019). "Distribution-independent pac learning of halfspaces with massart noise". In: *Advances in Neural Information Processing Systems* 32.

Diakonikolas, Ilias, Kane, Daniel, and Ren, Lisheng (2023). "Near-optimal cryptographic hardness of agnostically learning halfspaces and relu regression under gaussian marginals". In: *International Conference on Machine Learning*. PMLR.

Diakonikolas, Ilias, Kane, Daniel, and Zarifis, Nikos (2020a). "Near-Optimal SQ Lower Bounds for Agnostically Learning Halfspaces and ReLUs under Gaussian Marginals". In: *Annual Conference on Neural Information Processing Systems*.

Diakonikolas, Ilias, Park, Jongho, and Tzamos, Christos (2021a). "ReLU Regression with Massart Noise". In: *Advances in Neural Information Processing Systems* 34.

Diakonikolas, Ilias et al. (2020b). "Approximation Schemes for ReLU regression". In: *Conference on Learning Theory, COLT 2020*.

Diakonikolas, Ilias et al. (2021b). "The optimality of polynomial regression for agnostic learning under Gaussian marginals in the SQ model". In: *Conference on Learning Theory*. PMLR, pp. 1552–1584.

Diakonikolas, Ilias et al. (2022b). "Hardness of Learning a Single Neuron with Adversarial Label Noise". In: *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*.

Duchi, John Charles and Namkoong, Hongseok (2021). "Learning models with uniform performance via distributionally robust optimization". In: *The Annals of Statistics* 49.3.

Dwork, Cynthia et al. (2012). "Fairness through awareness". In: *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*.

Goel, Surbhi, Gollakota, Aravind, and Klivans, Adam (2020). "Statistical-Query Lower Bounds via Functional Gradients". In: *Annual Conference on Neural Information Processing Systems*.

Goel, Surbhi, Karmalkar, Sushrut, and Klivans, Adam R (2019). "Time/Accuracy Tradeoffs for Learning a ReLU with respect to Gaussian Marginals". In: *Advances in Neural Information Processing Systems 32*.

Haghtalab, Nika, Jordan, Michael I., and Zhao, Eric (2022). "On-Demand Sampling: Learning Optimally from Multiple Distributions". In: *Advances in Neural Information Processing Systems*. Vol. 35.

Hashimoto, Tatsunori et al. (2018). "Fairness without demographics in repeated loss minimization". In: *International Conference on Machine Learning*. PMLR.

Haussler, David (1992). "Decision theoretic generalizations of the PAC model for neural net and other learning applications". In: *Information and Computation* 100, pp. 78–150.

Huang, Jiayuan et al. (2006). "Correcting sample selection bias by unlabeled data". In: *Advances in neural information processing systems* 19.

Kakade, Sham M et al. (2011). "Efficient learning of generalized linear and single index models with isotonic regression". In: *Advances in Neural Information Processing Systems* 24.

Kalai, Adam Tauman and Sastry, Rajeev (2009). "The Isotron Algorithm: High-Dimensional Isotonic Regression." In: *COLT*.

Kalan, Seyed Morteza Mousavi, Soltanolkotabi, Mahdi, and Avestimehr, Salman (2019). "Fitting relus via sgd and quantized sgd". In: *2019 IEEE International Symposium on Information Theory (ISIT)*. IEEE.

Karmakar, Sayar, Mukherjee, Anirbit, and Muthukumar, Ramchandran (2021). "A Study of Neural Training with Iterative Non-Gradient Methods". In: *SSRN Electronic Journal*.

Kearns, Michael John, Schapire, Robert Elias, and Sellie, Linda Marie (1992). "Toward efficient agnostic learn-

ing". In: *Proceedings of the fifth annual workshop on Computational learning theory*.

Kingma, Diederik P and Ba, Jimmy (2015). "Adam: A Method for Stochastic Optimization". In: *International Conference on Learning Representations (ICLR)*.

Kotsalis, Georgios, Lan, Guanghui, and Li, Tianjiao (2022). "Simple and optimal methods for stochastic variational inequalities, I: operator extrapolation". In: *SIAM Journal on Optimization* 32.3, pp. 2041–2073.

Li, Jiajin, Zhu, Linglingzhi, and So, Anthony Man-Cho (2025). "Nonsmooth nonconvex–nonconcave minimax optimization: Primal–dual balancing and iteration complexity analysis". In: *Mathematical Programming*, pp. 1–51.

Li, Shuyao et al. (2024). "Learning a Single Neuron Robustly to Distributional Shifts and Adversarial Label Noise". In: *Advances in Neural Information Processing Systems*. Ed. by Globerson, A. et al. Vol. 37. Curran Associates, Inc., pp. 67383–67421.

Lin, Tianyi, Jin, Chi, and Jordan, Michael (2020). "On gradient descent ascent for nonconvex-concave minimax problems". In: *International conference on machine learning*. PMLR, pp. 6083–6093.

Mansour, Yishay, Mohri, Mehryar, and Rostamizadeh, Afshin (2009). "Domain Adaptation: Learning Bounds and Algorithms". In: *Proceedings of the 22nd Annual Conference on Learning Theory (COLT)*. Montréal, Canada.

Manurangsi, Pasin and Reichman, Daniel (2018). "The computational complexity of training relu (s)". In: *arXiv preprint arXiv:1810.04207*.

Mehta, Ronak, Diakonikolas, Jelena, and Harchaoui, Zaid (2025). "Min-Max Optimization with Dual-Linear Coupling". In: *arXiv preprint arXiv:2507.06328*.

Nelder, John Ashworth and Wedderburn, Robert William Macdonald (1972). "Generalized linear models". In: *Journal of the Royal Statistical Society: Series A (General)* 135.3.

Oren, Yonatan et al. (2019). "Distributionally Robust Language Modeling". In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Ed. by Inui, Kentaro et al. Hong Kong, China: Association for Computational Linguistics, pp. 4227–4237.

Pan, Sinno Jialin and Yang, Qiang (2009). "A survey on transfer learning". In: *IEEE Transactions on knowledge and data engineering* 22.10.

Patel, Vishal M et al. (2015). "Visual domain adaptation: A survey of recent advances". In: *IEEE signal processing magazine* 32.3.

Qi, Qi et al. (2021). "An online method for a class of distributionally robust optimization with non-convex objectives". In: *Advances in Neural Information Processing Systems* 34, pp. 10067–10080.

Rafique, Hassan et al. (2022). "Weakly-convex–concave min–max optimization: provable algorithms and applications in machine learning". In: *Optimization Methods and Software* 37.3, pp. 1087–1121.

Rosenblatt, Frank (1958). "The perceptron: a probabilistic model for information storage and organization in the brain". In: *Psychological Review* 65.6, pp. 386–408.

Sagawa, Shiori et al. (2020). "Distributionally Robust Neural Networks for Group Shifts: On the Importance of Regularization for Worst-Case Generalization". In: *International Conference on Learning Representations (ICLR)*.

Shapiro, Alexander (2017). "Distributionally robust stochastic programming". In: *SIAM Journal on Optimization* 27.4.

Shimodaira, Hidetoshi (2000). "Improving predictive inference under covariate shift by weighting the log-likelihood function". In: *Journal of statistical planning and inference* 90.2.

Sima, Jiri (2002). "Training a single sigmoidal neuron is hard". In: *Neural computation* 14.11.

Sinha, Aman, Namkoong, Hongseok, and Duchi, John (2018). "Certifying Some Distributional Robustness with Principled Adversarial Training". In: *International Conference on Learning Representations*.

Soltanolkotabi, Mahdi (2017). "Learning ReLUs via gradient descent". In: *Advances in neural information processing systems* 30.

Soma, Tasuku, Gatmiry, Khashayar, and Jegelka, Stefanie (2022). *Optimal algorithms for group distributionally robust optimization and beyond*.

Tan, Chuanqi et al. (2018). "A survey on deep transfer learning". In: *Artificial Neural Networks and Machine Learning–ICANN 2018*. Springer.

Together Computer (2023). *RedPajama: An Open Source Recipe to Reproduce LLaMA training dataset*. URL: `https://github.com/togethercomputer/RedPajama-Data` (Accessed 10/02/2025).

Wang, Jun-Kun, Abernethy, Jacob, and Levy, Kfir Y. (2024). "No-Regret Dynamics in the Fenchel Game: A Unified Framework for Algorithmic Convex Optimization". In: *Mathematical Programming* 205.1-2, pp. 203–268.

Wang, Puqian et al. (2023). "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, pp. 36541–36577.

Xia, Mengzhou et al. (2024). "Sheared LLaMA: Accelerating Language Model Pre-training via Structured Pruning". In: *International Conference on Learning Representations (ICLR), 2024*.

– *Sheared LLaMA: Accelerating Language Model Pre-training*. `https://github.com/princeton-nlp/LLM-Shearing/blob/`

`main / llmshearing / callbacks / dynamic_loading_callback.py`. (Accessed 10/02/2025).

Xie, Sang Michael et al. (2023). "Doremi: Optimizing data mixtures speeds up language model pretraining". In: *Advances in Neural Information Processing Systems* 36, pp. 69798–69818.

Xu, Ziyu et al. (2020). "Class-weighted classification: Trade-offs and robust approaches". In: *International conference on machine learning*. PMLR.

Yehudai, Gilad and Shamir, Ohad (2020). "Learning a single neuron with gradient methods". In: *Conference on Learning Theory*, pp. 3756–3786.

Zhang, Jiawei et al. (2020). "A single-loop smoothed gradient descent-ascent algorithm for nonconvex-concave min-max problems". In: *Advances in neural information processing systems* 33, pp. 7377–7389.

Zhang, Lijun et al. (2023). "Stochastic Approximation Approaches to Group Distributionally Robust Optimization". In: *Advances in Neural Information Processing Systems (NeurIPS)*.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes]
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Yes]

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes]
   (b) Complete proofs of all theoretical results. [Yes]
   (c) Clear explanations of any assumptions. [Yes]

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes]
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes]
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes]
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Yes]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [Yes]
   (b) The license information of the assets, if applicable. [Yes]
   (c) New assets either in the supplemental material or as a URL, if applicable. [Yes]
   (d) Information about consent from data providers/curators. [Yes]
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Yes]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. [Not Applicable]
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

Supplementary Material

Organization In Appendix A, we review additional related work. In Appendix B, we collect the definitions and standard results in probability and optimization that we use throughout the paper. In Appendix C, we prove that, with high probability, all the empirical expectations we need closely track their population counterparts. Finally, in Appendix E, we provide the full proof of the gap upper bound, Proposition 3.4.

A Additional Related Work

Robustly Learning Neurons Generalized linear models (a.k.a. "single neurons") have long been a cornerstone of statistics and machine learning (Nelder and Wedderburn, 1972), and early algorithms such as the Isotron (Kalai and Sastry, 2009) and Kakade-Shamir's efficient GLM learner (Kakade et al., 2011) provided the first guarantees for learning a single neuron in the presence of mild, structured label noise. More recent work has focused on ReLU activations under progressively stronger noise models: for realizable or random additive noise, both gradient-based and spectral methods can recover or approximate the true neuron (Soltanolkotabi, 2017; Kalan et al., 2019; Yehudai and Shamir, 2020); in the fully agnostic setting, achieving purely additive error $\epsilon > 0$ is intractable even for Gaussian inputs (Diakonikolas et al., 2020a; Goel et al., 2020; Diakonikolas et al., 2022b), though constant-factor approximations are possible under log-concave or other structured marginals (Diakonikolas et al., 2020b; Diakonikolas et al., 2022a; Wang et al., 2023). Extensions to semi-random or Massart-type noise require further specialized techniques (Diakonikolas et al., 2021a; Karmakar et al., 2021; Chen et al., 2020; Diakonikolas et al., 2019) beyond the scope of the present work.

In this work, we consider the strictly harder regime where an adversary may not only corrupt labels within each of K groups, but also shift the covariate distribution across groups. We formulate this as a Group DRO problem over $p_{[1]}, \ldots, p_{[K]}$, seeking $\boldsymbol{w}$ that minimizes the worst-case weighted squared loss plus an f-divergence penalty on the group weights.

Distributionally Robust Optimization Covariate shift and related distributional mismatches have been studied extensively in the past—in covariate-shift correction (Shimodaira, 2000; Huang et al., 2006; Bickel et al., 2007), label-proportion changes (Dwork et al., 2012; Xu et al., 2020), domain adaptation and transfer learning (Mansour et al., 2009; Pan and Yang, 2009; Ben-David et al., 2010; Patel et al., 2015; Tan et al., 2018), and classical DRO (Ben-Tal et al., 2009; Shapiro, 2017).

Recent machine-learning work has applied DRO to language modeling (Oren et al., 2019), class-imbalance (Xu et al., 2020), group fairness (Hashimoto et al., 2018), and robust regression (Blanchet et al., 2021; Duchi and Namkoong, 2021; Chen and Paschalidis, 2018). However, these methods typically assume convex, Lipschitz losses and structured label noise, and they do not address nonconvex problems like learning a neuron with respect to the square loss. On the other hand, even the work that did consider nonconvex loss functions like Qi et al. (2021) and Sinha et al. (2018) only considered general, non-structured smooth loss functions, for which only convergence to stationary points can be guaranteed. In addition to not exploring the structured nonconvexity of common learning tasks, for concrete learning tasks such as those considered in our work, convergence to stationary points is known to be insufficient—even for the special case of learning a ReLU neuron without any distributional ambiguity, stationary points of the squared loss may not lead to any formal learning guarantees (Yehudai and Shamir, 2020).

Most closely related to our work in this context is Li et al. (2024), which explored the task of robustly learning a single neuron in a related, but distinct distributionally robust optimization setting, where there is a single reference distribution and each sample could be reweighted. The main body of the paper has already provided a comparison to Li et al. (2024) both in terms of the results and the techniques, thus we omit repeating it here.

Group DRO As a subclass of problems in Distributionally Robust Optimization, a line of work isolates distributional shift that manifests through a finite set of semantic or demographic groups and asks the learner to minimize the worst-case loss across those groups. Sagawa et al. (2020) first demonstrated its empirical effectiveness on over-parameterized neural networks, showing dramatic accuracy gains for minority groups. Subsequent theoretical work established sharp sample-complexity bounds under convex, unregularized losses (Haghtalab et al., 2022; Soma et al., 2022; Zhang et al., 2023). There are also related explorations into minimax regret optimization (Agarwal and Zhang, 2022) and no-regret dynamics (Wang et al., 2024). They provide algorithmic blueprints using iterative methods like online gradient descent and stochastic optimization techniques, with established convergence guarantees and complexity bounds.

In the context of large-scale language model pretraining, DoReMi (Xie et al., 2023) has empirically demonstrated the

effectiveness of dynamically adjusting mixture weights across domains. Building on this line of work, LLM-Shearing (Xia et al., 2024) further explored structured pruning as a means to accelerate pretraining while maintaining competitive performance of DoReMi.

Our setting addresses an even more aggressive setting, involving two robustness notions—agnostic label noise and adversarial covariate shifts across K groups—by penalizing deviations from uniform group weights via an f-divergence in a Group DRO formulation. This allows us to obtain the first provable, polynomial-time algorithm for learning a nonconvex neuron model under both sources of adversarial perturbation.

B Supplementary Preliminaries

In the appendix, we recall that $\phi(\widehat{\boldsymbol{\lambda}})$ generically to represent a divergence between $\widehat{\boldsymbol{\lambda}}$ and the uniform distribution, instantiated as either $\mathrm{KL}(\widehat{\boldsymbol{\lambda}}, \widehat{\boldsymbol{\lambda}}_0)$ or $\chi^2(\widehat{\boldsymbol{\lambda}}, \widehat{\boldsymbol{\lambda}}_0)$, where $\widehat{\boldsymbol{\lambda}}_0 = \frac{1}{K}\mathbf{1}$.

We state here several classical inequalities, and useful definitions and facts.

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 $t > 0$,*

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

Definition B.3 (Total Variation Distance)**.** Let P and Q be two probability distributions on a measurable space $(\Omega, \mathcal{F})$. The total variation (TV) distance between P and Q is defined as

$$TV(P, Q) := \sup_{A \in \mathcal{F}} |P(A) - Q(A)|,$$

which represents the largest difference in the probabilities that the two distributions assign to the same event. If P and Q admit probability mass functions $P(x)$ and $Q(x)$, respectively, then the total variation distance admits the equivalent form:

$$TV(P, Q) = \frac{1}{2}\sum_{x \in \Omega} |P(x) - Q(x)|.$$

Lemma B.4 (Pinsker's Inequality)**.** *Let P and Q be two probability distributions on the same sample space. Then,*

$$TV(P, Q) \leq \sqrt{\frac{1}{2}\mathrm{KL}(P, Q)},$$

where $\mathrm{KL}(P, Q)$ denotes the Kullback-Leibler divergence between P and Q.

A key consequence of our distributional assumption in Assumption 1.2 is a high-probability bound on the norm of the covariates, as formalized below.

Lemma B.5. *Let $\delta \in (0, 1)$. Under Assumption 1.2, with probability at least $1 - \delta$, we have that $\|\boldsymbol{x}\|_2 \leq S\sqrt{d}$ for all covariates $\boldsymbol{x}$ in a total of N samples from all groups, where $S = B\log(dN/\delta)$.*

Proof. Since each group has N/K independent samples, applying Assumption 1.2 gives for any one sample $\boldsymbol{x}_{[i]}^{(j)}$ and any unit $\boldsymbol{u}$,

$$\Pr\left[|\boldsymbol{u} \cdot \boldsymbol{x}_{[i]}^{(j)}| \geq S\right] \leq e^{-S/B}.$$

Applying this inequality with $\boldsymbol{u}$ chosen as each of the standard basis vectors and then applying a union bound over d coordinates leads to $\Pr[\|\boldsymbol{x}_{[i]}^{(j)}\|_2 \geq S\sqrt{d}] \leq d\, e^{-S/B}$. Another union bound over all N samples gives $\Pr[\exists\, \boldsymbol{x} : \|\boldsymbol{x}\|_2 \geq S\sqrt{d}] \leq dN\, e^{-S/B}$. Setting $S = B\log(dN/\delta)$ completes the argument. □

We will also need the following "concentration-via-clipping" argument, which we will apply in Appendix C.

Lemma B.6 (One-Dimensional Concentration via Clipping)**.** *Let Z be a real-valued random variable with mean $\mu = \mathbb{E}[Z]$. Assume Z satisfies the tail bound $\Pr(|Z| > t) \le 2\exp(-t^{1/\tau}/B)$ for all $t \ge t_0$, where $t_0 > 0$ and $\tau \ge 1$ is a bounded integer. Let $Z_1, \dots, Z_n$ be n i.i.d. copies of Z.*

For any error $\epsilon > 0$, failure probability $\delta \in (0,1)$, and sufficiently large n satisfying

$$n = \tilde{O}\left(\max\left(\frac{B^{2\tau}}{\epsilon^2}, \frac{t_0^2}{\epsilon^2}\right)\log\left(\frac{1}{\delta}\right)\right),$$

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

$$\left|\frac{1}{n}\sum_{i=1}^{n} Z_i - \mu\right| \le \epsilon.$$

Proof. Let the clipping threshold be $S \ge t_0$, to be determined later, and define $Z_c := \max(-S, \min(Z, S))$. We decompose the total error using the triangle inequality:

$$\left|\frac{1}{n}\sum_{i=1}^{n} Z_i - \mu\right| \le \underbrace{\left|\frac{1}{n}\sum_{i=1}^{n}(Z_i - Z_{i,c})\right|}_{\text{(I) Clipping Error}} + \underbrace{\left|\frac{1}{n}\sum_{i=1}^{n} Z_{i,c} - \mathbb{E}[Z_c]\right|}_{\text{(II) Concentration Error}} + \underbrace{|\mathbb{E}[Z_c] - \mu|}_{\text{(III) Bias Error}}.$$

We now bound each of these three terms.

Clipping term: The first term is non-zero only if at least one sample $|Z_i| > S$. Let $\mathscr{E}_{\text{clip}}$ be this event. By a union bound over the samples and the tail bound on Z:

$$\Pr(\mathscr{E}_{\text{clip}}) = \Pr(\exists i : |Z_i| > S) \le n \cdot \Pr(|Z| > S) \le 2n\exp(-S^{1/\tau}/B).$$

This means with probability $1 - 2n\exp(-S^{1/\tau}/B)$, we have $\left|\frac{1}{n}\sum_{i=1}^{n}(Z_i - Z_{i,c})\right| = 0$.

Bias term: The bias $|\mathbb{E}[Z_c] - \mu| = |\mathbb{E}[Z_c - Z]|$ is bounded by the integral of the tail probability:

$$|\mathbb{E}[Z_c - Z]| \le \mathbb{E}[|Z| \cdot \mathbb{I}(|Z| > S)] = \int_S^\infty \Pr(|Z| > t)dt \le \int_S^\infty 2e^{-t^{1/\tau}/B}dt,$$

where the first inequality follows from the natural coupling between Z_c and Z assigning every point in Z_c to the corresponding unclipped point in Z. If the point is unclipped, this contributes 0, if the point is clipped, then this means $|Z| > S$ and it contributes $|Z| - S < |Z|$. The equality is the standard way to express expectations of nonnegative random variables as tail integrals. The final inequality follows from the tail bound on $|Z|$.

We now further upper bound the right hand side above.

Let $u = t^{1/\tau}/B$, so $t = (Bu)^\tau$ and $dt = \tau B(Bu)^{\tau-1}du$. The lower limit of integration becomes $u_0 = S^{1/\tau}/B$.

$$\int_{u_0}^\infty 2e^{-u}\tau B(Bu)^{\tau-1}du = 2\tau B^\tau \int_{u_0}^\infty u^{\tau-1}e^{-u}du.$$

For $u_0 \ge \tau - 1$ (which will be satisfied by the choice of S later), we can bound $\int_{u_0}^\infty u^{\tau-1}e^{-u}du \le c_\tau u_0^{\tau-1}e^{-u_0}$ for some constant $c_\tau < 2$ depending on τ (by standard facts about exponential integrals). The bias is thus bounded by $2\tau B^\tau c_\tau u_0^{\tau-1}e^{-u_0} = 2\tau B^\tau c_\tau (S^{1/\tau}/B)^{\tau-1}e^{-(S^{1/\tau}/B)} = \Theta_\tau$.

Concentration term: Finally, we bound the concentration term. The clipped variable Z_c is bounded in $[-S, S]$. We can apply Hoeffding's inequality to its empirical mean. With probability at least $1 - \delta/2$:

$$\left|\frac{1}{n}\sum_{i=1}^{n} Z_{i,c} - \mathbb{E}[Z_c]\right| \le \sqrt{\frac{2S^2\log(4/\delta)}{n}} = S\sqrt{\frac{2\log(4/\delta)}{n}}.$$

Substituting $S = (B\log(4n/\delta))^\tau$, this deviation is bounded by $(B\log(4n/\delta))^\tau\sqrt{\frac{2\log(4/\delta)}{n}}$.

Accounting for everything, we see that, with high probability, the deviation is at most the sum of the concentration and bias errors. We need to choose the clipping threshold S and the sample size n to make this total error at most ϵ, while keeping the failure probability at most δ.

This requires satisfying the following conditions simultaneously:

$$S \geq (B \log(4n/\delta))^\tau \tag{16}$$
$$c_\tau B S^{(\tau-1)/\tau} e^{-S^{1/\tau}/B} \leq \epsilon/2 \tag{17}$$
$$n \geq \frac{8S^2}{\epsilon^2} \log(4/\delta) \tag{18}$$
$$S \geq t_0 \tag{19}$$
$$S \geq (\tau-1)^\tau B^\tau \tag{20}$$

where the first condition ensures the clipping event probability is at most $\delta/2$, the second bounds the bias (where the constants have been absorbed into c_τ), and the third bounds the concentration error of the clipped estimators. The second-last condition is just to ensure that the tail bound applies, and the final condition corresponds to $u_0 \geq \tau - 1$.

Choosing S: To satisfy all of these inequalities simultaneously, we define the clipping threshold S:

$$S := \max\left(\left(B \log\left(\frac{C_\tau B^\tau}{\epsilon}\right)\right)^\tau, \left((\tau-1)\, B \log(4n/\delta)\right)^\tau, t_0\right)$$

for a sufficiently large constant C_τ. The first term in the $\max$ is chosen to satisfy the bias condition (17), and the second term satisfies the clipping condition (16) as well as (20). We now explain how to derive the first term.

Let us define the variable $X = S^{1/\tau}/B$. The inequality can be rewritten in terms of X by substituting $S^{1/\tau} = BX$:

$$c_\tau B (BX)^{\tau-1} e^{-X} \leq \frac{\epsilon}{2}$$
$$c_\tau B^\tau X^{\tau-1} e^{-X} \leq \frac{\epsilon}{2}$$

To satisfy this, we need to find a sufficiently large X. For any $\tau \geq 1$, there exists a (constant) threshold $X_0(\tau)$ such that for all $X \geq X_0(\tau)$, the polynomial term $X^{\tau-1}$ is bounded by an exponential, i.e., $X^{\tau-1} \leq e^{X/2}$. Assuming X is large enough to meet this condition, our inequality becomes:

$$c_\tau B^\tau e^{X/2} e^{-X} = c_\tau B^\tau e^{-X/2} \leq \frac{\epsilon}{2}$$

Solving for X:

$$e^{-X/2} \leq \frac{\epsilon}{2c_\tau B^\tau}$$
$$-X/2 \leq \log\left(\frac{\epsilon}{2c_\tau B^\tau}\right) = -\log\left(\frac{2c_\tau B^\tau}{\epsilon}\right)$$
$$X \geq 2\log\left(\frac{2c_\tau B^\tau}{\epsilon}\right)$$

This choice of X is consistent with the initial assumption that X is large (for small ϵ), as the logarithm will be large. We can absorb the constant factors into a new constant C_τ and state that the condition is satisfied if $X \geq 2\log\left(C_\tau \frac{B^\tau}{\epsilon}\right)$. Substituting back $X = S^{1/\tau}/B$, we get:

$$\frac{S^{1/\tau}}{B} \geq 2\log\left(C_\tau \frac{B^\tau}{\epsilon}\right)$$

This gives the required lower bound on S:

$$S \geq \left(2B \log\left(C_\tau \frac{B^\tau}{\epsilon}\right)\right)^\tau$$

This demonstrates that a choice of S satisfying $S = \tilde{O}((B \log(B/\epsilon))^\tau)$ is sufficient to control the bias term.

Solving for n: Substituting this choice of S into the concentration condition (18) yields a lower bound on n. This gives three requirements on n, corresponding to which term in the $\max$ defining S is dominant.

1. If S is determined by the bias term (the first term), then n must satisfy: $n \geq \frac{8\log(4/\delta)}{\epsilon^2}\left(B\log\left(\frac{C_\tau B^\tau}{\epsilon}\right)\right)^{2\tau}$. This contributes to $n = \tilde{O}\left(\frac{B^{2\tau}}{\epsilon^2}\log\left(\frac{1}{\delta}\right)\right)$ in the stated sample complexity.

2. If S is determined by the clipping probability term (the second term), we get a recursive condition on n:

$$n \geq \frac{8\log(4/\delta)}{\epsilon^2}(B\log(4n/\delta))^{2\tau}.$$

   This is a recursive inequality of the form $n \geq K\log^{2\tau}(n)$ where $K = \frac{8B^{2\tau}\log(4/\delta)}{\epsilon^2}$; by the standard technique for solving such recursive inequalities this is satisfied by all $n \geq \max((2\tau e)^{2\tau}, 2K\log^{2\tau}(2K)) = \Theta(K\log^{2\tau}(K))$, and so it suffices that:

$$n = \tilde{O}\left(K\log^{2\tau}(K)\right) = \tilde{O}\left(\frac{B^{2\tau}}{\epsilon^2}\log^{2\tau}\left(\frac{B^{2\tau}}{\epsilon^2}\right)\right).$$

   Since the $\tilde{O}$ notation absorbs polylogarithmic factors, this simplifies to $n = \tilde{O}\left(\frac{B^{2\tau}}{\epsilon^2}\log(\frac{1}{\delta})\right)$.

3. If S is determined by the threshold term t_0 (the third term), then $S = t_0$ and we need: $n \geq \frac{8\log(4/\delta)}{\epsilon^2}t_0^2$. This contributes $n = O\left(\frac{t_0^2}{\epsilon^2}\log(1/\delta)\right)$ to the sample complexity.

The total sample complexity is the maximum of these three requirements, giving us the sample complexity stated in the lemma. □

C Uniform Convergence and Population Approximation

In Section 3 we carried out our main convergence analysis under the empirical distributions $\widehat{p}_{[i]}$. Here we show that, with high probability, all required empirical expectations closely match their population counterparts under each true distribution $p_{[i]}$. Our approach combines Lemma B.6 with a standard net argument after clipping the functions involved, to obtain convergence uniformly over the weight domain $\mathcal{B}(W)$, and then applies this result to the sharpness and moment bounds as well as to control the difference between the empirical and population optima.

Lemma C.1 (Uniform Concentration for Heavy-Tailed Variables)**.** *Let $h : \mathcal{B}(W) \times \mathbb{R}^d \to \mathbb{R}$ be a function. Assume $\boldsymbol{x}$ satisfies Assumption 1.2 with parameter B_x, and for any $\boldsymbol{w} \in \mathcal{B}(W)$, the random variable $Z_{\boldsymbol{w}} = h(\boldsymbol{w}; \boldsymbol{x})$ satisfies the tail bound $\Pr(|Z_{\boldsymbol{w}}| > t) \leq 2\exp(-t^{1/\tau}/B_h)$ for all $t \geq t_0$, where $t_0 = \tilde{O}(WB_x\beta)$ may depend on problem parameters, $\tau \geq 1$ is a bounded integer and $B_h \geq B_x$.*

Let $a(\boldsymbol{x})$ be the Lipschitz constant of $h(\boldsymbol{w}; \boldsymbol{x})$ with respect to $\boldsymbol{w} \in \mathcal{B}(W)$, and suppose that $a(\boldsymbol{x}) \leq a_{S_x}$ whenever $\|\boldsymbol{x}\|_2 \leq \sqrt{d}S_x$, for $S_x = \tilde{O}(B_x)$ chosen according to Lemma B.5 and $a_{S_x} = \tilde{O}(d^2B_x^4)$ as required in Lemma C.2 below.

Then, for sufficiently large N such that

$$N = \tilde{O}\left(K \cdot \frac{t_0^2 + B_h^{2\tau}}{\epsilon^2}\left(d\log\frac{dWB_x}{\epsilon} + \log\frac{K}{\delta}\right)\right),$$

with probability at least $1 - \delta$, for every group $i \in [K]$ and all $\boldsymbol{w} \in \mathcal{B}(W)$:

$$\left|\mathbb{E}_{\widehat{p}_{[i]}}[h(\boldsymbol{w}; \boldsymbol{x})] - \mathbb{E}_{p_{[i]}}[h(\boldsymbol{w}; \boldsymbol{x})]\right| \leq \epsilon.$$

Proof. We will follow a clipping argument similar to that in Lemma B.6. We define a clipping threshold for the data norm, $S_x = B_x\log(3dN^2/\delta)$. Let $\mathscr{E}_x$ be the event $\|\boldsymbol{x}\|_2 \leq \sqrt{d}S_x$. We decompose h into a "clipped" part and a "tail" part—$h_c(\boldsymbol{w}; \boldsymbol{x}) := h(\boldsymbol{w}; \boldsymbol{x}) \cdot \mathbb{I}[\mathscr{E}_x]$ and $h_t(\boldsymbol{w}; \boldsymbol{x}) := h(\boldsymbol{w}; \boldsymbol{x}) \cdot \mathbb{I}[\neg\mathscr{E}_x]$, so that $h = h_c + h_t$. The total deviation can then be bounded using the triangle inequality:

$$|\mathbb{E}_{\widehat{p}_{[i]}}[h] - \mathbb{E}_{p_{[i]}}[h]| \leq \underbrace{|\mathbb{E}_{\widehat{p}_{[i]}}[h_c] - \mathbb{E}_{p_{[i]}}[h_c]|}_{\text{(I)}} + \underbrace{|\mathbb{E}_{\widehat{p}_{[i]}}[h_t]|}_{\text{(II)}} + \underbrace{|\mathbb{E}_{p_{[i]}}[h_t]|}_{\text{(III)}}.$$

We will show that (I) is uniformly bounded by ϵ and that (II) and (III) are each uniformly bounded by $\epsilon/3$ with high probability.

Bounding the Tail Contribution (Terms II and III) By Lemma B.5, with probability at least $1-\delta/(3N)$, all covariates $\boldsymbol{x}$ in the total sample of size N satisfy $\|\boldsymbol{x}\|_2 \le \sqrt{d}S_x$, where $S_x = B_x \log(3dN^2/\delta)$. On this high-probability event, the indicator $\mathbb{I}[\|\boldsymbol{x}_j\|_2 > \sqrt{d}S_x]$ is zero for all samples $\boldsymbol{x}_j$. Consequently, for any group i, the empirical expectation of the tail part is zero:

$$\mathbb{E}_{\widehat{p}_{[i]}}[h_t] = \frac{K}{N}\sum_{j=1}^{N/K} h(\boldsymbol{w};\boldsymbol{x}_{ij})\mathbb{I}[\|\boldsymbol{x}_{ij}\|_2 > \sqrt{d}S_x] = 0.$$

Since this does not depend on $\boldsymbol{w}$ or i, Term (II) is zero for all $\boldsymbol{w}$ and all groups $i \in [K]$ simultaneously.

For Term (III), we bound the population mean of the tail part for all $\boldsymbol{w}$, $|\mathbb{E}_{p_{[i]}}[h_t]|$, using the Cauchy-Schwarz inequality:

$$\begin{aligned}
|\mathbb{E}_{p_{[i]}}[h_t]| &\le \mathbb{E}_{p_{[i]}}\left[|h(\boldsymbol{w};\boldsymbol{x})| \cdot \mathbb{I}[\|\boldsymbol{x}\|_2 > \sqrt{d}S_x]\right] \\
&\le \sqrt{\mathbb{E}_{p_{[i]}}[h(\boldsymbol{w};\boldsymbol{x})^2]} \cdot \sqrt{\Pr(\|\boldsymbol{x}\|_2 > \sqrt{d}S_x)} \\
&\le \sqrt{\mathbb{E}_{p_{[i]}}[h(\boldsymbol{w};\boldsymbol{x})^2]} \cdot \sqrt{\delta/(3N)}.
\end{aligned}$$

The second inequality is the bound on the data probability, which is bounded by $\delta/(3N)$ by our choice of S_x. The first term is the square root of the second moment of $h(\boldsymbol{w};\boldsymbol{x})$. We can bound this moment using its tail parameters. Using the tail integral formula, $\mathbb{E}[Z^2] = \int_0^\infty \Pr(|Z| > \sqrt{u})du$, we split the integral at $u = t_0^2$:

$$\mathbb{E}_{p_{[i]}}[h(\boldsymbol{w};\boldsymbol{x})^2] \le \int_0^{t_0^2} 1du + \int_{t_0^2}^\infty 2\exp(-u^{1/(2\tau)}/B_h)du = t_0^2 + 4\tau B_h^{2\tau}\Gamma(2\tau, t_0^{1/\tau}/B_h) \le \tilde{O}(t_0^2 + B_h^{2\tau}),$$

where the final inequality is a consequence standard upper bounds on the Γ function. Thus, we have $|\mathbb{E}_{p_{[i]}}[h_t]| \le \sqrt{\tilde{O}(t_0^2 + B_h^{2\tau})}\sqrt{\delta/(3N)}$. To ensure this is less than $\epsilon/3$, it suffices to choose a sufficiently large $N = \tilde{O}(t_0^2 + B_h^{2\tau})\ \delta/\epsilon^2$.

Uniform Concentration of the Clipped Part (Term I) Let $a_{S_x} > \sup_{\|\boldsymbol{x}\|_2 \le \sqrt{d}S_x} a(\boldsymbol{x})$ be a deterministic upper bound on $a(\boldsymbol{x})$. For any $\boldsymbol{x}$ where $\mathbb{I}[\mathscr{E}_x] = 1$, we have $\|\boldsymbol{x}\|_2 \le \sqrt{d}S_x$, and therefore $|h_c(\boldsymbol{w}_1;\boldsymbol{x}) - h_c(\boldsymbol{w}_2;\boldsymbol{x})| \le a(\boldsymbol{x})\|\boldsymbol{w}_1 - \boldsymbol{w}_2\|_2 \le a_{S_x}\|\boldsymbol{w}_1 - \boldsymbol{w}_2\|_2$.

Since their random variable $Z_{\boldsymbol{w},c} = h_c(\boldsymbol{w};\boldsymbol{x})$ is a clipped version of $Z_{\boldsymbol{w}} = h(\boldsymbol{w};\boldsymbol{x})$, its tails are at least as light. We follow the standard net argument for h_c. Let $\mathcal{N}$ be an r-net over $\mathscr{B}(W)$ with $|\mathcal{N}| \le (3W/r)^d$. We want to achieve $|\mathbb{E}_{\widehat{p}_{[i]}}[h_c] - \mathbb{E}_{p_{[i]}}[h_c]| \le \epsilon/3$ for all points in the net and all groups. Using Lemma B.6 with accuracy $\epsilon/3$ and failure probability $\delta' = \frac{\delta/3}{K|\mathcal{N}|}$, further, substituting $r = \epsilon/6a_{S_x}$, the required sample size $N = nK$ satisfies:

$$N = \tilde{O}\left(K \cdot \max\left(\frac{B_h^{2\tau}}{\epsilon^2}, \frac{t_0^2}{\epsilon^2}\right)\left(d\log\frac{Wa_{S_x}}{\epsilon} + \log\frac{K}{\delta}\right)\right).$$

Plugging in bound on a_{S_x} from the lemma statement, we recover the stated sample complexity. With this sample size, with probability at least $1-\delta/3$, we have $|\mathbb{E}_{\widehat{p}_{[i]}}[h_c(\boldsymbol{w}')] - \mathbb{E}_{p_{[i]}}[h_c(\boldsymbol{w}')]| \le \epsilon/3$ for all $\boldsymbol{w}' \in \mathcal{N}$.

Extending to the continuum via the triangle inequality and the Lipschitz constant a_{S_x}, for any $\boldsymbol{w} \in \mathscr{B}(W)$:

$$|\mathbb{E}_{\widehat{p}_{[i]}}[h_c] - \mathbb{E}_{p_{[i]}}[h_c]| \le |\mathbb{E}_{\widehat{p}_{[i]}}[h_c(\boldsymbol{w})] - \mathbb{E}_{\widehat{p}_{[i]}}[h_c(\boldsymbol{w}')]| + \epsilon/3 + |\mathbb{E}_{p_{[i]}}[h_c(\boldsymbol{w}')] - \mathbb{E}_{p_{[i]}}[h_c(\boldsymbol{w})]| \le 2a_{S_x}r + \epsilon/3.$$

We have three events we need to hold: the empirical tail is zero, the population tail is small, and the clipped part concentrates. By a union bound, all three hold with probability at least $1-\delta$. On this combined event, we have:

$$|\mathbb{E}_{\widehat{p}_{[i]}}[h] - \mathbb{E}_{p_{[i]}}[h]| \le (2a_{S_x}r + \epsilon/3) + 0 + \epsilon/3 \le 2a_{S_x}r + (2\epsilon/3) \le \epsilon.$$

While the stated inequalities for S_x and N lead to a recursive bound on N—the inequality for N has a $\log\log(N)$ term on the RHS due to S_x scaling logarithmically with N—using standard techniques like the one from the end of Lemma B.6, we can obtain an explicit bound on N, at a poly-log cost absorbed by the $\tilde{O}$ notation. □

Lemma C.2 (Empirical Sharpness and Moment Bounds with Heavy Tails)**.** *Under Assumptions 1.2 and 1.3, let the number of samples*

$$N = \tilde{O}_{\beta,B}\left(\frac{KW^4}{\epsilon^2}\left(d\log\frac{dW}{\epsilon} + \log\frac{K}{\delta}\right)\right)$$

be sufficiently large. Then with probability at least $1-\delta$*, for every group* $i \in [K]$*, every* $\boldsymbol{w} \in \mathcal{B}(3\|\boldsymbol{w}_*\|)$ *with* $\|\boldsymbol{w}-\boldsymbol{w}_*\| \geq \sqrt{\epsilon}$*, and every unit vector* $\boldsymbol{u}$*, the following hold:*

$$\mathbb{E}_{\boldsymbol{x}\sim\widehat{p}_{[i]}}\left[(\sigma(\boldsymbol{w}\cdot\boldsymbol{x}) - \sigma(\boldsymbol{w}_*\cdot\boldsymbol{x}))(\boldsymbol{w}\cdot\boldsymbol{x} - \boldsymbol{w}_*\cdot\boldsymbol{x})\right] \geq \tfrac{c_0}{2}\|\boldsymbol{w}-\boldsymbol{w}_*\|_2^2,$$
$$\mathbb{E}_{\boldsymbol{x}\sim\widehat{p}_{[i]}}\left[(\boldsymbol{u}\cdot\boldsymbol{x})^\tau\right] \leq 6B, \quad \textit{for } \tau = 2, 4.$$

Proof. The proof consists of applying the uniform concentration result for heavy-tailed variables (Lemma C.1) to three different functions. We will analyze the sample complexity for each and take the maximum to get the final result.

1. Sharpness Bound. Let $h_1(\boldsymbol{w};\boldsymbol{x}) = (\sigma(\boldsymbol{w}\cdot\boldsymbol{x}) - \sigma(\boldsymbol{w}_*\cdot\boldsymbol{x}))(\boldsymbol{w}\cdot\boldsymbol{x} - \boldsymbol{w}_*\cdot\boldsymbol{x})$. By Assumption 1.3, its true expectation is lower bounded: $\mathbb{E}_{p_{[i]}}[h_1(\boldsymbol{w};\boldsymbol{x})] \geq c_0\|\boldsymbol{w}-\boldsymbol{w}_*\|_2^2$. We want to show that with high probability, the empirical expectation is close to this value.

The function class $\{h_1(\boldsymbol{w};\cdot) : \boldsymbol{w} \in \mathcal{B}(W)\}$ is defined over the parameter ball $\boldsymbol{w} \in \mathcal{B}(W)$. Let $\Delta\boldsymbol{w} := \boldsymbol{w} - \boldsymbol{w}_*$, and assume $\|\Delta\boldsymbol{w}\|_2 \leq 2W$ for all $\boldsymbol{w}$ under consideration. Consider the random variable $Z_{\boldsymbol{w}} = h_1(\boldsymbol{w};\boldsymbol{x})$. Its tail behavior can be bounded as follows:

Since σ is β-Lipschitz, we have $|\sigma(\boldsymbol{w}\cdot\boldsymbol{x}) - \sigma(\boldsymbol{w}_*\cdot\boldsymbol{x})| \leq \beta|\Delta\boldsymbol{w}\cdot\boldsymbol{x}|$. This implies:

$$|Z_{\boldsymbol{w}}| = |h_1(\boldsymbol{w};\boldsymbol{x})| \leq \beta(\Delta\boldsymbol{w}\cdot\boldsymbol{x})^2 = \beta\|\Delta\boldsymbol{w}\|_2^2(\boldsymbol{u}\cdot\boldsymbol{x})^2 \leq 4\beta W^2(\boldsymbol{u}\cdot\boldsymbol{x})^2,$$

where $\boldsymbol{u} = \Delta\boldsymbol{w}/\|\Delta\boldsymbol{w}\|_2$.

By Assumption 1.2, the random variable $(\boldsymbol{u}\cdot\boldsymbol{x})^2$ has a tail bound $\Pr((\boldsymbol{u}\cdot\boldsymbol{x})^2 > t) \leq 2\exp(-t^{1/2}/B)$. The tail bound for $Z_{\boldsymbol{w}}$ is therefore:

$$\Pr(|Z_{\boldsymbol{w}}| > t) \leq \Pr\left((\boldsymbol{u}\cdot\boldsymbol{x})^2 > t/(4\beta W^2)\right) \leq 2\exp\left(-\frac{(t/(4\beta W^2))^{1/2}}{B}\right) = 2\exp\left(-\frac{t^{1/2}}{2B\sqrt{\beta W^2}}\right).$$

This shows that $Z_{\boldsymbol{w}}$ has a heavy tail with shape parameter $\tau_{h_1} = 2$ and an effective scale parameter $B_{h_1} = 2\sqrt{\beta}WB$.

To apply Lemma C.1, we also need the effective Lipschitz constant of the function class, which we denote by a_1. To find the Lipschitz constant of $h_1(\boldsymbol{w};\boldsymbol{x})$ with respect to $\boldsymbol{w}$, we bound its gradient's norm:

$$\begin{aligned}
\|\nabla_{\boldsymbol{w}} h_1(\boldsymbol{w};\boldsymbol{x})\|_2 &= \|\sigma'(\boldsymbol{w}\cdot\boldsymbol{x})\boldsymbol{x}(\boldsymbol{w}\cdot\boldsymbol{x} - \boldsymbol{w}_*\cdot\boldsymbol{x}) + (\sigma(\boldsymbol{w}\cdot\boldsymbol{x}) - \sigma(\boldsymbol{w}_*\cdot\boldsymbol{x}))\boldsymbol{x}\|_2 \\
&\leq |\sigma'(\boldsymbol{w}\cdot\boldsymbol{x})|\cdot\|\boldsymbol{x}\|_2\cdot|\Delta\boldsymbol{w}\cdot\boldsymbol{x}| + |\sigma(\boldsymbol{w}\cdot\boldsymbol{x}) - \sigma(\boldsymbol{w}_*\cdot\boldsymbol{x})|\cdot\|\boldsymbol{x}\|_2 \\
&\leq \beta\|\boldsymbol{x}\|_2\cdot(\|\Delta\boldsymbol{w}\|_2\|\boldsymbol{x}\|_2) + (\beta\|\Delta\boldsymbol{w}\|_2\|\boldsymbol{x}\|_2)\cdot\|\boldsymbol{x}\|_2 \\
&\leq 4\beta W\|\boldsymbol{x}\|_2^2.
\end{aligned}$$

This gives a random Lipschitz constant $a_1(\boldsymbol{x}) = 4\beta W\|\boldsymbol{x}\|_2^2$. Note that $a_1(\boldsymbol{x}) \leq 4\beta W S^2 d \leq \tilde{O}(\beta dWB^2)$ if $\|\boldsymbol{x}\|_2^2 \leq dS_x^2$ for $S_x = \tilde{O}(B)$.

Since we consider $\|\boldsymbol{w}-\boldsymbol{w}_*\|^2 \geq \epsilon$, we can now use Lemma C.1 with a target additive error $\frac{c_0}{4}\epsilon$ (i.e. replace ϵ in C.2 with $c_0\epsilon/4$), and a net radius $r = \frac{c_0\epsilon}{8a_1}$. The required number of samples N_1 is dominated by the heavy-tailed term in the sample complexity bound, which must use the parameters specific to h_1: Setting $t_0 = 0$ in Lemma C.1, we start with the given sample complexity:

$$N_1 = \tilde{O}\left(K\cdot\frac{(B_{h_1})^{2\tau_{h_1}}}{(c_0\epsilon/4)^2}\left(\log\frac{K}{\delta} + d\log\frac{dWB}{\epsilon}\right)\right)$$

Substituting the parameters $\tau_{h_1} = 2$, $B_{h_1} = 2\sqrt{\beta}WB$, we proceed with the calculation:

$$\begin{aligned}
N_1 &= \tilde{O}\left(K\cdot\frac{(2\sqrt{\beta}WB)^{2\cdot 2}}{(c_0\epsilon/4)^2}\left(\log\frac{K}{\delta} + d\log\frac{dWB}{\epsilon}\right)\right) \\
&= \tilde{O}\left(K\cdot\frac{64\beta^2W^4B^4}{c_0^2\epsilon^2}\left(d\log\frac{dWB}{\epsilon} + \log\frac{K}{\delta}\right)\right).
\end{aligned}$$

Absorbing the numerical constants and the logarithmic factors into the $\tilde{O}$ notation, we arrive at the simplified expression for the sample complexity:

$$N_1 = \tilde{O}\left(\frac{K\beta^2 W^4 B^4}{\epsilon^2}\left(d\log\frac{dWB}{\epsilon} + \log\frac{K}{\delta}\right)\right).$$

With this sample size, with probability at least $1-\delta/2$, for all $\boldsymbol{w} \in \mathscr{B}(W)$ with $\|\boldsymbol{w}-\boldsymbol{w}_*\|_2 \geq \sqrt{\epsilon}$:

$$\mathbb{E}_{\widehat{p}_{[i]}}[h_1(\boldsymbol{w};\boldsymbol{x})] \geq \mathbb{E}_{p_{[i]}}[h_1(\boldsymbol{w};\boldsymbol{x})] - (2a_1 r + t) \geq c_0\|\boldsymbol{w}-\boldsymbol{w}_*\|_2^2 - \left(\tfrac{c_0\epsilon}{4} + \tfrac{c_0\epsilon}{4}\right) \geq \frac{c_0}{2}\|\boldsymbol{w}-\boldsymbol{w}_*\|_2^2.$$

2. Moment Bounds. Let $h_2(\boldsymbol{u};\boldsymbol{x}) = (\boldsymbol{u}\cdot\boldsymbol{x})^4$ and $h_3(\boldsymbol{u};\boldsymbol{x}) = (\boldsymbol{u}\cdot\boldsymbol{x})^2$. These are defined over the space of unit vectors, $\boldsymbol{u} \in \mathbb{S}^{d-1}$. By Assumption 1.2, the population means are bounded, $\mathbb{E}_{p_{[i]}}[h_j] \leq 5B$, and their tails are given by:

$$\Pr(h_2(\boldsymbol{u};\boldsymbol{x}) > t) \leq 2\exp(-t^{1/4}/B) \quad (\tau_2 = 4, B_2 = B)$$
$$\Pr(h_3(\boldsymbol{u};\boldsymbol{x}) > t) \leq 2\exp(-t^{1/2}/B) \quad (\tau_3 = 2, B_3 = B)$$

To apply Lemma C.1, we first need upper bounds on the Lipschitz constants for h_2 and h_3 in the appropriate bounded region.

Lipschitz Constants: The gradients are $\nabla_{\boldsymbol{u}} h_2(\boldsymbol{u};\boldsymbol{x}) = 4(\boldsymbol{u}\cdot\boldsymbol{x})^3\boldsymbol{x}$ and $\nabla_{\boldsymbol{u}} h_3(\boldsymbol{u};\boldsymbol{x}) = 2(\boldsymbol{u}\cdot\boldsymbol{x})\boldsymbol{x}$. Their norms are bounded by random variables:

$$\|\nabla_{\boldsymbol{u}} h_2(\boldsymbol{u};\boldsymbol{x})\|_2 \leq 4\|\boldsymbol{x}\|_2^4, \quad \text{and} \quad \|\nabla_{\boldsymbol{u}} h_3(\boldsymbol{u};\boldsymbol{x})\|_2 \leq 2\|\boldsymbol{x}\|_2^2.$$

Similar to getting the upper bound for $a_1(\boldsymbol{w})$, we have $a_2(\boldsymbol{x}) = \tilde{O}(d^2B^4)$ and $a_3(\boldsymbol{x}) = \tilde{O}(dB^2)$ if $\|\boldsymbol{w}\|_2^2 \leq dS_x^2$.

Sample Complexity and Concentration: We apply Lemma C.1 with a target deviation of $\epsilon = B/2$ and net radii $r_j = \epsilon/(2a_j) = B/(4a_j)$ for $j = 2, 3$. For the functions h_2 and h_3, the tail bounds hold for all $t \geq 0$, so we can take $t_0 = 0$. The domain for $\boldsymbol{u}$ is the unit sphere, so we take $W = 1$.

For $h_2(\boldsymbol{u};\boldsymbol{x}) = (\boldsymbol{u}\cdot\boldsymbol{x})^4$, we have tail parameters $\tau_2 = 4$ and $B_2 = B$. Since the Lipschitz constant is upper bounded by $a_2 = \tilde{\Theta}(d^2B^4)$ whenever $\|x\|_2^2 \leq dS_x^2$, the sample complexity N_2 is given by the formula from Lemma C.1:

$$\begin{aligned} N_2 &= \tilde{O}\left(K\cdot\frac{B^{2\cdot 4}}{(B/2)^2}\left(d\log\frac{4a_2}{B} + \log\frac{K}{\delta}\right)\right) \\ &= \tilde{O}\left(KB^6\left(d + \log\frac{K}{\delta}\right)\right), \end{aligned}$$

where we absorbed the polylogarithmic factors in B and d into the $\tilde{O}$ notation. For $h_3(\boldsymbol{u};\boldsymbol{x}) = (\boldsymbol{u}\cdot\boldsymbol{x})^2$, we have tail parameters $\tau_3 = 2$ and $B_3 = B$. The Lipschitz constant is upper bounded by $a_3 = \tilde{\Theta}(dB^2)$ whenever $\|x\|_2^2 \leq dS_x^2$. The sample complexity N_3 is similarly calculated to be:

$$N_3 = \tilde{O}\left(KB^2\left(d + \log\frac{K}{\delta}\right)\right).$$

The sample complexity for the moment bounds is dominated by N_2. With this many samples, with probability at least $1-\delta/2$, we have for all unit vectors $\boldsymbol{u}$:

$$\mathbb{E}_{\widehat{p}_{[i]}}[h_2(\boldsymbol{u};\boldsymbol{x})] \leq \mathbb{E}_{p_{[i]}}[h_2(\boldsymbol{u};\boldsymbol{x})] + (2a_2 r_2 + \epsilon) \leq 5B + \left(2a_2\frac{\epsilon}{2a_2} + \epsilon\right) = 5B + 2\epsilon.$$

With our choice of $\epsilon = B/2$, the total bound is $5B + 2(B/2) = 6B$. A similar analysis for h_3 also gives us a bound of $6B$ for that case. By combining the three sample complexity bounds and omitting the dependence on β and B, we obtain the overall sample complexity in the lemma statement. □

Lemma C.3 (Empirical vs. Population Optima, Heavy-Tailed Version)**.** *Under Assumptions 1.2 and 1.3, and in the setting of Fact 2.3, let the number of samples be*

$$N = \tilde{O}_{\beta,B}\left(\frac{KW^4}{\epsilon^2}\log\left(\frac{1}{\delta}\right)\right)$$

Then with probability at least $1-\delta$*, it holds that* $\widehat{\mathrm{OPT}_{\mathrm{m}}} \leq \mathrm{OPT}_{\mathrm{m}} + \epsilon$.

Proof. The proof proceeds in three steps: first, we derive the tail properties of the loss function; second, we apply a concentration inequality with a union bound; and third, we use a property of the maximum function to conclude.

Recall that Fact 2.3 bounds $|y| \le M$ with $M = C_M WB\beta \log(\beta BW/\epsilon)$ for a large enough constant C_M.

Tail Bound of the Loss Function. Let the loss be the random variable $Z = l(\boldsymbol{w}_*; \boldsymbol{x}, y) = (\sigma(\boldsymbol{w}_* \cdot \boldsymbol{x}) - y)^2$. We know σ is β-Lipschitz continuous, $\sigma(0) = 0$, and y is bounded by M. To find the tail bound for Z, we first establish an upper bound on its value:

$$Z = (\sigma(\boldsymbol{w}_* \cdot \boldsymbol{x}) - y)^2 \le 2(\sigma(\boldsymbol{w}_* \cdot \boldsymbol{x})^2 + y^2) \le 2(\beta^2(\boldsymbol{w}_* \cdot \boldsymbol{x})^2 + M^2).$$

The term $(\boldsymbol{w}_* \cdot \boldsymbol{x})^2$ can be written as $\|\boldsymbol{w}_*\|_2^2(\boldsymbol{u} \cdot \boldsymbol{x})^2 \le W^2(\boldsymbol{u} \cdot \boldsymbol{x})^2$, where $\boldsymbol{u} = \frac{\boldsymbol{w}_*}{\|\boldsymbol{w}_*\|_2}$ is a unit vector. This gives $Z \le 2(W^2\beta^2(\boldsymbol{u} \cdot \boldsymbol{x})^2 + M^2)$. We can now bound the tail probability of Z for any $t > 2M^2$:

$$\Pr(Z > t) \le \Pr(2(W^2\beta^2(\boldsymbol{u} \cdot \boldsymbol{x})^2 + M^2) > t) = \Pr\left((\boldsymbol{u} \cdot \boldsymbol{x})^2 > \frac{t/2 - M^2}{(\beta W)^2}\right).$$

From Assumption 1.2, the variable $(\boldsymbol{u} \cdot \boldsymbol{x})^2$ has a sub-exponential tail with $\tau = 2$ and parameter B. Let $t' = (t/2 - M^2)/(\beta W)^2$. The tail bound is $\Pr((\boldsymbol{u} \cdot \boldsymbol{x})^2 > t') \le 2\exp(-(t')^{1/2}/B)$. Substituting t' back gives:

$$\Pr(Z > t) \le 2\exp\left(-\frac{\sqrt{t/2 - M^2}}{\beta WB}\right).$$

To establish a clean sub-exponential form of the type $\exp(-t^{1/2}/B_l)$, we consider the tail for $t \ge 4M^2$. For this range of t, we have $t/2 - M^2 \ge t/2 - t/4 = t/4$. Therefore, $\sqrt{t/2 - M^2} \ge \sqrt{t/4} = \sqrt{t}/2$. Substituting this into the exponent gives:

$$\Pr(Z > t) \le 2\exp\left(-\frac{\sqrt{t}/2}{\beta WB}\right) = 2\exp\left(-\frac{t^{1/2}}{2\beta WB}\right).$$

This bound holds for all $t \ge 4M^2 = \tilde{\Theta}(\beta^2 W^2 B^2)$. This is a valid sub-exponential tail with exponent $\tau_l = 2$ and an effective tail parameter $B_l = 2WB\beta$.

Concentration and Union Bound over Groups. Recall that $\widehat{\mathrm{OPT}_{\mathrm{m}}} = \max_{i\in[K]} \mathbb{E}_{\widehat{p}_{[i]}}[l(\boldsymbol{w}_*; \boldsymbol{x}, y)]$ and $\mathrm{OPT}_{\mathrm{m}} = \max_{i\in[K]} \mathbb{E}_{p_{[i]}}[l(\boldsymbol{w}_*; \boldsymbol{x}, y)]$. Also, let $\boldsymbol{\ell}^* = [l^*_{[1]}, \ldots, l^*_{[K]}]$, where $l^*_{[i]} = \mathbb{E}_{p_{[i]}}(l(\boldsymbol{w}_*; \boldsymbol{x}, y))$, for all $i \in [K]$, and we define $\hat{\boldsymbol{\ell}}^*$, $\hat{l}^*_{[i]}$ for the empirical case similarly. To ensure the empirical loss $\hat{l}^*_{[i]}$ concentrates around the true loss $l^*_{[i]}$ for all K groups simultaneously, we apply Lemma B.6 combined with a union bound. We require the deviation for each group to be at most ϵ. To achieve a total failure probability of at most δ, we set the per-group failure probability to $\delta' = \delta/K$. Lemma B.6 states that for $\tau_l = 2$ and $t_0 = \tilde{\Theta}(\beta^2 W^2 B^2)$, the number of samples per group, $n = N/K$, must satisfy:

$$n = \tilde{O}\left(\frac{B_l^{2\tau_l} + \beta^4 W^4 B^4}{\epsilon^2}\left(\log\frac{B_l}{\epsilon}\right)^{2\tau_l}\log\left(\frac{1}{\delta}\right)\right) = \tilde{O}\left(\frac{\beta^4 W^4 B^4}{\epsilon^2}\log\left(\frac{1}{\delta}\right)\right).$$

Consequently, the total number of samples $N = nK$ must satisfy the bound stated in the lemma. With the sample size N, the union bound guarantees that with probability at least $1 - \delta$, we have $|\hat{l}^*_{[i]} - l^*_{[i]}| \le \epsilon$ for all $i \in [K]$, i.e., $\|\hat{\boldsymbol{\ell}}^* - \boldsymbol{\ell}^*\|_\infty \le \epsilon$.

Finally, note that $\boldsymbol{a}, \boldsymbol{b} \in \mathbb{R}^K$, $|\max_i a_i - \max_j b_j| \le \|\boldsymbol{a} - \boldsymbol{b}\|_\infty$, we have:

$$|\widehat{\mathrm{OPT}_{\mathrm{m}}} - \mathrm{OPT}_{\mathrm{m}}| = |\max_i \hat{l}^*_{[i]} - \max_j l^*_{[j]}| \le \|\hat{\boldsymbol{\ell}}^* - \boldsymbol{\ell}^*\|_\infty \le \epsilon.$$

This implies $\widehat{\mathrm{OPT}_{\mathrm{m}}} \le \mathrm{OPT}_{\mathrm{m}} + \epsilon$, completing the proof. □

D Proof of Gap Lower Bound

In this section, we adapt techniques from Li et al. (2024) to establish Lemma 3.2, restated below for convenience.

Lemma 3.2 (Gap Lower Bound)**.** *Under Assumptions 1.2 and 1.3, if the per-group sample size N/K is sufficiently large (see Lemma C.2), for all $\boldsymbol{w} \in \mathscr{B}(3\|\boldsymbol{w}_*\|_2)$ and all $\widehat{\boldsymbol{\lambda}} \in \Delta_K$, we have* $\mathrm{Gap}(\boldsymbol{w}, \widehat{\boldsymbol{\lambda}}) \ge -\frac{12\beta^2 B}{c_1}\widehat{\mathrm{OPT}_{\mathrm{m}}} + \frac{c_1}{2}\|\boldsymbol{w} - \boldsymbol{w}_*\|_2^2 + \nu D_\phi(\widehat{\boldsymbol{\lambda}}^*, \widehat{\boldsymbol{\lambda}})$.

Proof. We split the primal-dual gap into the primal gap and the dual gap: $\mathrm{Gap}(\boldsymbol{w}, \widehat{\boldsymbol{\lambda}}) = \big[L(\boldsymbol{w}, \widehat{\boldsymbol{\lambda}}^*) - L(\boldsymbol{w}_*, \widehat{\boldsymbol{\lambda}}^*)\big] + \big[L(\boldsymbol{w}_*, \widehat{\boldsymbol{\lambda}}^*) - L(\boldsymbol{w}_*, \widehat{\boldsymbol{\lambda}})\big]$. We expand the primal gap as

$$\begin{aligned} &L(\boldsymbol{w}, \widehat{\boldsymbol{\lambda}}^*) - L(\boldsymbol{w}_*, \widehat{\boldsymbol{\lambda}}^*) = \sum_{i=1}^{K} \widehat{\lambda}^*_{[i]} \mathbb{E}_{(\boldsymbol{x},y)\sim\widehat{p}_{[i]}}[(\sigma(\boldsymbol{w}\cdot\boldsymbol{x}) - y)^2 - (\sigma(\boldsymbol{w}_*\cdot\boldsymbol{x}) - y)^2] \\ =& \sum_{i=1}^{K} \widehat{\lambda}^*_{[i]} \mathbb{E}_{\widehat{p}_{[i]}}[((\sigma(\boldsymbol{w}\cdot\boldsymbol{x}) - \sigma(\boldsymbol{w}_*\cdot\boldsymbol{x}))^2] - 2\sum_{i=1}^{K} \widehat{\lambda}^*_{[i]} \mathbb{E}_{\widehat{p}_{[i]}}[(\sigma(\boldsymbol{w}_*\cdot\boldsymbol{x}) - y)(\sigma(\boldsymbol{w}\cdot\boldsymbol{x}) - \sigma(\boldsymbol{w}_*\cdot\boldsymbol{x}))]. \end{aligned}$$

We now bound both of the above terms. For the first term, Equation (5) implies,

$$\sum_{i=1}^{K} \widehat{\lambda}^*_{[i]} \mathbb{E}_{\widehat{p}_{[i]}}[((\sigma(\boldsymbol{w}\cdot\boldsymbol{x}) - \sigma(\boldsymbol{w}_*\cdot\boldsymbol{x}))^2] \geq \sum_{i=1}^{K} \widehat{\lambda}^*_{[i]} c_1 \|\boldsymbol{w} - \boldsymbol{w}_*\|_2^2 = c_1\|\boldsymbol{w} - \boldsymbol{w}_*\|_2^2. \tag{21}$$

For the second term, by the Cauchy-Schwarz inequality, we have

$$\begin{aligned} &\sum_{i=1}^{K} \widehat{\lambda}^*_{[i]} \mathbb{E}_{\widehat{p}_{[i]}}[(\sigma(\boldsymbol{w}_*\cdot\boldsymbol{x}) - y)(\sigma(\boldsymbol{w}\cdot\boldsymbol{x}) - \sigma(\boldsymbol{w}_*\cdot\boldsymbol{x}))] \\ \leq& \sum_{i=1}^{K} \widehat{\lambda}^*_{[i]} \sqrt{\mathbb{E}_{\widehat{p}_{[i]}}[(\sigma(\boldsymbol{w}_*\cdot\boldsymbol{x}) - y)^2] \mathbb{E}_{\widehat{p}_{[i]}}[(\sigma(\boldsymbol{w}\cdot\boldsymbol{x}) - \sigma(\boldsymbol{w}_*\cdot\boldsymbol{x}))^2]} \\ \leq& \sum_{i=1}^{K} \widehat{\lambda}^*_{[i]} \sqrt{\mathbb{E}_{\widehat{p}_{[i]}}[(\sigma(\boldsymbol{w}_*\cdot\boldsymbol{x}) - y)^2]}\beta\sqrt{6B}\|\boldsymbol{w} - \boldsymbol{w}_*\|_2 \leq \sqrt{\widehat{\mathrm{OPT}_{\mathrm{m}}}}\beta\sqrt{6B}\|\boldsymbol{w} - \boldsymbol{w}_*\|_2, \end{aligned} \tag{22}$$

where the second inequality follows from Equation (5) and the last inequality applies Jensen's inequality to the concave function $\sqrt{\cdot}$ and then uses the definition of $\widehat{\mathrm{OPT}_{\mathrm{m}}}$.

Combining Equation (21) and Equation (22), the primal gap is bounded below by

$$-2\beta\sqrt{6B}\|\boldsymbol{w} - \boldsymbol{w}_*\|_2\sqrt{\widehat{\mathrm{OPT}_{\mathrm{m}}}} + c_1\|\boldsymbol{w} - \boldsymbol{w}_*\|_2^2 \geq -\frac{12\beta^2 B}{c_1}\widehat{\mathrm{OPT}_{\mathrm{m}}} + \frac{c_1}{2}\|\boldsymbol{w} - \boldsymbol{w}_*\|_2^2, \tag{23}$$

where we use Young's inequality (Fact B.1): $2\beta\sqrt{6B}\|\boldsymbol{w} - \boldsymbol{w}_*\|_2\sqrt{\widehat{\mathrm{OPT}_{\mathrm{m}}}} \leq \frac{4\beta^2 6B}{2c_1}\widehat{\mathrm{OPT}_{\mathrm{m}}} + \frac{c_1}{2}\|\boldsymbol{w} - \boldsymbol{w}_*\|_2^2$. For the dual gap, we have

$$\begin{aligned} &L(\boldsymbol{w}_*, \widehat{\boldsymbol{\lambda}}^*) - L(\boldsymbol{w}_*, \widehat{\boldsymbol{\lambda}}) = -L(\boldsymbol{w}_*, \widehat{\boldsymbol{\lambda}}) - (-L(\boldsymbol{w}_*, \widehat{\boldsymbol{\lambda}}^*)) \\ =& \nabla_{\widehat{\boldsymbol{\lambda}}}(-L(\boldsymbol{w}_*, \widehat{\boldsymbol{\lambda}}^*)) \cdot (\widehat{\boldsymbol{\lambda}} - \widehat{\boldsymbol{\lambda}}^*) + D_{-L(\boldsymbol{w}_*,\cdot)}(\widehat{\boldsymbol{\lambda}}, \widehat{\boldsymbol{\lambda}}^*) \geq \nu D_\phi(\widehat{\boldsymbol{\lambda}}^*, \widehat{\boldsymbol{\lambda}}), \end{aligned} \tag{24}$$

where the second equality and the last inequality follow from Facts 2.4 and 2.5. Combining Equation (23) and Equation (24) completes the proof. □

E Proof of Gap Upper Bound

In this section we prove Proposition 3.4, which shows that the cumulative primal-dual gap $\sum_{t=1}^{n} a_t \,\mathrm{Gap}(\boldsymbol{w}_t, \widehat{\boldsymbol{\lambda}}_t)$ can be controlled by telescoping the contributions from the primal updates in $\boldsymbol{w}$, the dual updates in $\widehat{\boldsymbol{\lambda}}$, and a small residual term that is absorbed by our choice of step sizes. The argument proceeds in three steps:

1. **Per-iteration decomposition** We combine Lemmas E.1 and E.2 (an upper bound on $a_t L(\boldsymbol{w}_t, \widehat{\boldsymbol{\lambda}}^*)$ and a lower bound on $a_t L(\boldsymbol{w}_*, \widehat{\boldsymbol{\lambda}}_t)$, respectively) to bound from above each $a_t \mathrm{Gap}(\boldsymbol{w}_t, \widehat{\boldsymbol{\lambda}}_t)$ by a sum of differences in squared distances, Bregman divergences, and a residual error E_t. Most of the technical work is devoted to proving Lemma E.2.
2. **Telescoping Sum** Summing the per-iteration bounds from $t = 1$ to n causes most distance and divergence terms to telescope, leaving only boundary terms at $t = 0$ and $t = n$, plus one remaining inner product $-a_n \sum_{i=1}^{K}(\widehat{\lambda}_{n[i]} - \widehat{\lambda}_{n-1[i]})\mathbb{E}_{\widehat{p}_{[i]}}[\langle \boldsymbol{v}(\boldsymbol{w}_{n-1}; \boldsymbol{x}, y), \boldsymbol{w}_* - \boldsymbol{w}_n\rangle]$.

3. **Residual Control** We bound that final inner product via Young's inequality and our step-size choice, absorbing it back into the telescoped Bregman and distance terms.

Before we carry out these steps in detail, we first state the two necessary bounds on $a_t L(\boldsymbol{w}_t, \widehat{\boldsymbol{\lambda}}^*)$ and $a_t L(\boldsymbol{w}_*, \widehat{\boldsymbol{\lambda}}_t)$ that are required per-iteration decomposition. We defer their proofs to Appendices E.1 and E.2.

Lemma E.1 (Upper Bound for $a_t L(\boldsymbol{w}_t, \widehat{\boldsymbol{\lambda}}^*)$)**.** *In each iteration t, let*

$$\widehat{\boldsymbol{\lambda}}_t \;=\; \arg\max_{\widehat{\boldsymbol{\lambda}} \in \Delta_K} \big\{\, a_t L(\boldsymbol{w}_t, \widehat{\boldsymbol{\lambda}}) \;-\; (\nu_0 + \nu A_{t-1})\, D_\phi(\widehat{\boldsymbol{\lambda}}, \widehat{\boldsymbol{\lambda}}_{t-1}) \big\}.$$

Then for all $t \geq 1$,

$$\begin{aligned} a_t L(\boldsymbol{w}_t, \widehat{\boldsymbol{\lambda}}^*) \;\leq\; & a_t L(\boldsymbol{w}_t, \widehat{\boldsymbol{\lambda}}_t) \;-\; (\nu_0 + \nu A_{t-1})\, D_\phi(\widehat{\boldsymbol{\lambda}}_t, \widehat{\boldsymbol{\lambda}}_{t-1}) \;-\; (\nu_0 + \nu A_t)\, D_\phi(\widehat{\boldsymbol{\lambda}}^*, \widehat{\boldsymbol{\lambda}}_t) \\ & +\; (\nu_0 + \nu A_{t-1})\, D_\phi(\widehat{\boldsymbol{\lambda}}^*, \widehat{\boldsymbol{\lambda}}_{t-1}). \end{aligned}$$

Lemma E.2 (Lower Bound for $a_t L(\boldsymbol{w}_*, \widehat{\boldsymbol{\lambda}}_t)$)**.** *Under Assumptions 1.2 and 1.3, if the per-group sample size N/K is sufficiently large (see Lemma C.2), where $\boldsymbol{w}_t$ is the sequence of iterates in Algorithm 1, we have for each $t \geq 1$:*

$$\begin{aligned} a_t L(\boldsymbol{w}_*, \widehat{\boldsymbol{\lambda}}_t) \geq\; & a_t L(\boldsymbol{w}_t, \widehat{\boldsymbol{\lambda}}_t) - \frac{1 + 0.5 c_1 A_{t-1}}{2} \|\boldsymbol{w}_* - \boldsymbol{w}_{t-1}\|_2^2 + \frac{1 + 0.5 c_1 A_t}{2} \|\boldsymbol{w}_* - \boldsymbol{w}_t\|_2^2 \\ & -\; a_{t-1} \sum_{i=1}^{K} (\widehat{\lambda}_{t-1[i]} - \widehat{\lambda}_{t-2[i]}) \mathbb{E}_{\widehat{p}_{[i]}} [\langle \boldsymbol{v}(\boldsymbol{w}_{t-2}; \boldsymbol{x}, y), \boldsymbol{w}_* - \boldsymbol{w}_{t-1} \rangle] \\ & +\; a_t \sum_{i=1}^{K} (\widehat{\lambda}_{t[i]} - \widehat{\lambda}_{t-1[i]}) \mathbb{E}_{\widehat{p}_{[i]}} [\langle \boldsymbol{v}(\boldsymbol{w}_{t-1}; \boldsymbol{x}, y), \boldsymbol{w}_* - \boldsymbol{w}_t \rangle] \\ & -\; \frac{1 + 0.5 c_1 A_{t-1}}{4} \|\boldsymbol{w}_{t-2} - \boldsymbol{w}_{t-1}\|_2^2 + \frac{1 + 0.5 c_1 A_t}{4} \|\boldsymbol{w}_{t-1} - \boldsymbol{w}_t\|_2^2 \\ & -\; (\nu_0 + \nu A_{t-2}) D_\phi(\widehat{\boldsymbol{\lambda}}_{t-1}, \widehat{\boldsymbol{\lambda}}_{t-2}) - \frac{28 \beta^2 B \widehat{\mathrm{OPT}_{\mathrm{m}}} a_t}{c_1}. \end{aligned}$$

For ease of recall, we restate here the quantities defined in Algorithm 1 and Equation (6):

$$\begin{aligned} C_3 &:= 31 \beta \sqrt{B} / c_1, \\ C_4 &:= 27 c_1 + 2163 \beta^4 B^2 / c_1, \\ C_W &:= \sqrt{6\beta^2 + c_M^2 B \log^2 \Big(\frac{\beta B W}{\epsilon} \Big)}, \\ C'_W &:= 2\sqrt{3} C_W \beta W B, \\ a_t &= \min \Big\{ \Big(1 + \frac{c_1}{8 C_4} \Big)^{t-1} \frac{1}{4 C_4}, \max \Big\{ \Big(1 + \frac{\sqrt{c_1 \nu}}{4\sqrt{2} C'_W} \Big)^{t-1} \frac{\sqrt{\nu_0}}{4 C'_W}, \frac{c_1 \nu_0}{(4\sqrt{2} C'_W)^2} t \Big\} \Big\}, \\ A_n &= \sum_{t=0}^{n} a_t. \end{aligned}$$

We now carry out these steps in detail, and prove our main upper bound on the gap.

Proposition 3.4 (Gap Upper Bound)**.** *Let the sequences $\{a_t\}$, $\{A_t\}$, $\{\boldsymbol{w}_t\}$, and $\{\widehat{\boldsymbol{\lambda}}_t\}$ be generated by Algorithm 1, where, by convention, $a_{-1} = a_0 = A_{-1} = A_0 = 0$, $\boldsymbol{w}_{-1} = \boldsymbol{w}_0 = \mathbf{0}$, and $\widehat{\boldsymbol{\lambda}}_{-1} = \widehat{\boldsymbol{\lambda}}_0 = \frac{1}{K}\mathbf{1}$. Under Assumptions 1.2 and 1.3, if*

the per-group sample size N/K is sufficiently large (see Lemma C.2), then for any $n \geq 1$:

$$\begin{aligned}
&\sum_{t=1}^{n} a_t \mathrm{Gap}(\boldsymbol{w}_t, \widehat{\boldsymbol{\lambda}}_t) \\
&\leq \frac{1}{2}\|\boldsymbol{w}_* - \boldsymbol{w}_0\|_2^2 - \frac{1+0.5c_1A_n}{4}\|\boldsymbol{w}_* - \boldsymbol{w}_n\|_2^2 \\
&\quad + \nu_0 D_\phi(\widehat{\boldsymbol{\lambda}}^*, \widehat{\boldsymbol{\lambda}}_0) - \frac{1+0.5c_1A_n}{4}\|\boldsymbol{w}_n - \boldsymbol{w}_{n-1}\|_2^2 \\
&\quad - (\nu_0 + \nu A_n) D_\phi(\widehat{\boldsymbol{\lambda}}^*, \widehat{\boldsymbol{\lambda}}_n) + \frac{28\beta^2 B}{c_1} A_n \widehat{\mathrm{OPT}_{\mathrm{m}}}.
\end{aligned}$$

Proof. **1. Per-iteration bound.** Recall

$$a_t \,\mathrm{Gap}(\boldsymbol{w}_t, \widehat{\boldsymbol{\lambda}}_t) \;=\; a_t\, L(\boldsymbol{w}_t, \widehat{\boldsymbol{\lambda}}^*) \;-\; a_t\, L(\boldsymbol{w}_*, \widehat{\boldsymbol{\lambda}}_t).$$

Combining Lemmas E.1 and E.2 gives, for each t,

$$\begin{aligned}
a_t \,\mathrm{Gap}(\boldsymbol{w}_t, \widehat{\boldsymbol{\lambda}}_t) \;\leq\; & \frac{1+0.5c_1A_{t-1}}{2}\|\boldsymbol{w}_* - \boldsymbol{w}_{t-1}\|_2^2 - \frac{1+0.5c_1A_t}{2}\|\boldsymbol{w}_* - \boldsymbol{w}_t\|_2^2 \\
&+\; a_{t-1}\sum_{i=1}^{K}(\widehat{\lambda}_{t-1[i]} - \widehat{\lambda}_{t-2[i]})\mathbb{E}_{\widehat{p}_{[i]}}[\langle \boldsymbol{v}(\boldsymbol{w}_{t-2}; \boldsymbol{x}, y), \boldsymbol{w}_* - \boldsymbol{w}_{t-1}\rangle] \\
&-\; a_t\sum_{i=1}^{K}(\widehat{\lambda}_{t[i]} - \widehat{\lambda}_{t-1[i]})\mathbb{E}_{\widehat{p}_{[i]}}[\langle \boldsymbol{v}(\boldsymbol{w}_{t-1}; \boldsymbol{x}, y), \boldsymbol{w}_* - \boldsymbol{w}_t\rangle] \\
&+\; (\nu_0 + \nu A_{t-2})\, D_\phi(\widehat{\boldsymbol{\lambda}}_{t-1}, \widehat{\boldsymbol{\lambda}}_{t-2}) - (\nu_0 + \nu A_{t-1})\, D_\phi(\widehat{\boldsymbol{\lambda}}_t, \widehat{\boldsymbol{\lambda}}_{t-1}) \\
&+\; \frac{1+0.5c_1A_{t-1}}{4}\|\boldsymbol{w}_{t-1} - \boldsymbol{w}_{t-2}\|_2^2 - \frac{1+0.5c_1A_t}{4}\|\boldsymbol{w}_t - \boldsymbol{w}_{t-1}\|_2^2 \\
&+\; (\nu_0 + \nu A_{t-1})\, D_\phi(\widehat{\boldsymbol{\lambda}}^*, \widehat{\boldsymbol{\lambda}}_{t-1}) - (\nu_0 + \nu A_t)\, D_\phi(\widehat{\boldsymbol{\lambda}}^*, \widehat{\boldsymbol{\lambda}}_t) \\
&+\; \frac{28\beta^2 B\widehat{\mathrm{OPT}_{\mathrm{m}}} a_t}{c_1}.
\end{aligned}$$

2. Telescoping over $t = 1, \ldots, n$. Summing the above inequalities from $t = 1$ to n causes all intermediate distance and divergence terms to cancel, leaving only the boundary terms at $t = 0$ and $t = n$, plus the term $-a_n \sum_{i=1}^{K}(\widehat{\lambda}_{n[i]} - \widehat{\lambda}_{n-1[i]})\mathbb{E}_{\widehat{p}_{[i]}}[\langle \boldsymbol{v}(\boldsymbol{w}_{n-1}; \boldsymbol{x}, y), \boldsymbol{w}_* - \boldsymbol{w}_n\rangle]$. Concretely, one obtains

$$\begin{aligned}
\sum_{t=1}^{n} a_t \,\mathrm{Gap}(\boldsymbol{w}_t, \widehat{\boldsymbol{\lambda}}_t) \;\leq\; & \frac{1}{2}\|\boldsymbol{w}_* - \boldsymbol{w}_0\|_2^2 - \frac{1+0.5c_1A_n}{2}\|\boldsymbol{w}_* - \boldsymbol{w}_n\|_2^2 \\
&-\; a_n\sum_{i=1}^{K}(\widehat{\lambda}_{n[i]} - \widehat{\lambda}_{n-1[i]})\mathbb{E}_{\widehat{p}_{[i]}}[\langle \boldsymbol{v}(\boldsymbol{w}_{n-1}; \boldsymbol{x}, y), \boldsymbol{w}_* - \boldsymbol{w}_n\rangle] \\
&-\; (\nu_0 + \nu A_{n-1})\, D_\phi(\widehat{\boldsymbol{\lambda}}_n, \widehat{\boldsymbol{\lambda}}_{n-1}) + \nu_0\, D_\phi(\widehat{\boldsymbol{\lambda}}^*, \widehat{\boldsymbol{\lambda}}_0) - (\nu_0 + \nu A_n)\, D_\phi(\widehat{\boldsymbol{\lambda}}^*, \widehat{\boldsymbol{\lambda}}_n) \\
&-\; \frac{1+0.5c_1A_n}{4}\|\boldsymbol{w}_n - \boldsymbol{w}_{n-1}\|_2^2 + \frac{28\beta^2 B\,\widehat{\mathrm{OPT}}\,A_n}{c_1}.
\end{aligned}$$

3. Bounding the final residual. It remains only to absorb the term $-a_n \sum_{i=1}^{K}(\widehat{\lambda}_{n[i]} - \widehat{\lambda}_{n-1[i]})\mathbb{E}_{\widehat{p}_{[i]}}[\langle \boldsymbol{v}(\boldsymbol{w}_{n-1}; \boldsymbol{x}, y), \boldsymbol{w}_* - \boldsymbol{w}_n\rangle]$. By Young's inequality and our choice of α_1 and step size a_t in Algorithm 1, we can show that (see the derivation of

Equation (33) in the proof of Lemma E.2 for the complete argument)

$$\begin{aligned}
&- a_n \sum_{i=1}^{K} (\widehat{\lambda}_{n[i]} - \widehat{\lambda}_{n-1[i]}) \mathbb{E}_{\widehat{p}_{[i]}}[\langle \boldsymbol{v}(\boldsymbol{w}_{n-1}; \boldsymbol{x}, y), \boldsymbol{w}_* - \boldsymbol{w}_n \rangle] \\
&\leq 4\sqrt{3} C_W \beta W B a_n (\frac{\alpha_1}{2} 2 D_\phi(\widehat{\boldsymbol{\lambda}}_n, \widehat{\boldsymbol{\lambda}}_{n-1}) + \frac{1}{2\alpha_1} \|\boldsymbol{w}_* - \boldsymbol{w}_n\|_2^2) \\
&\leq (\nu_0 + \nu A_{n-1}) D_\phi(\widehat{\boldsymbol{\lambda}}_n, \widehat{\boldsymbol{\lambda}}_{n-1}) + \frac{1 + 0.5 c_1 A_n}{4} \|\boldsymbol{w}_* - \boldsymbol{w}_n\|_2^2, \qquad (25)
\end{aligned}$$

substituting this back into the telescoped sum completes the proof. □

E.1 Proof of Lemma E.1

We now prove the upper bound on $a_t L(\boldsymbol{w}_t, \widehat{\boldsymbol{\lambda}}^*)$.

Lemma E.1 (Upper Bound for $a_t L(\boldsymbol{w}_t, \widehat{\boldsymbol{\lambda}}^*)$)**.** *In each iteration t, let*

$$\widehat{\boldsymbol{\lambda}}_t \;=\; \arg\max_{\widehat{\boldsymbol{\lambda}} \in \Delta_K} \big\{\, a_t L(\boldsymbol{w}_t, \widehat{\boldsymbol{\lambda}}) \;-\; (\nu_0 + \nu A_{t-1})\, D_\phi(\widehat{\boldsymbol{\lambda}}, \widehat{\boldsymbol{\lambda}}_{t-1}) \big\}.$$

Then for all $t \geq 1$,

$$\begin{aligned}
a_t L(\boldsymbol{w}_t, \widehat{\boldsymbol{\lambda}}^*) \;\leq\; & a_t L(\boldsymbol{w}_t, \widehat{\boldsymbol{\lambda}}_t) \;-\; (\nu_0 + \nu A_{t-1})\, D_\phi(\widehat{\boldsymbol{\lambda}}_t, \widehat{\boldsymbol{\lambda}}_{t-1}) \;-\; (\nu_0 + \nu A_t)\, D_\phi(\widehat{\boldsymbol{\lambda}}^*, \widehat{\boldsymbol{\lambda}}_t) \\
& +\; (\nu_0 + \nu A_{t-1})\, D_\phi(\widehat{\boldsymbol{\lambda}}^*, \widehat{\boldsymbol{\lambda}}_{t-1}).
\end{aligned}$$

Proof. We begin by introducing the quantity

$$h(\widehat{\boldsymbol{\lambda}}) \;:=\; a_t\, L(\boldsymbol{w}_t, \widehat{\boldsymbol{\lambda}}) \;-\; (\nu_0 + \nu A_{t-1})\, D_\phi(\widehat{\boldsymbol{\lambda}}, \widehat{\boldsymbol{\lambda}}_{t-1}).$$

Then observe that

$$a_t L(\boldsymbol{w}_t, \widehat{\boldsymbol{\lambda}}^*) = h(\widehat{\boldsymbol{\lambda}}^*) + (\nu_0 + \nu A_{t-1})\, D_\phi(\widehat{\boldsymbol{\lambda}}^*, \widehat{\boldsymbol{\lambda}}_{t-1}),$$

since we have simply added and subtracted the same Bregman term.

By construction, $h(\widehat{\boldsymbol{\lambda}})$ is the sum of a linear function in $\widehat{\boldsymbol{\lambda}}$ and a $-(\nu_0 + \nu A_{t-1})$-strongly concave term, so h itself is strongly concave. Hence its maximizer is exactly

$$\widehat{\boldsymbol{\lambda}}_t \;=\; \arg\max_{\widehat{\boldsymbol{\lambda}} \in \Delta_K} h(\widehat{\boldsymbol{\lambda}}).$$

We now apply the standard Bregman-divergence expansion around this maximizer:

$$h(\widehat{\boldsymbol{\lambda}}^*) = h(\widehat{\boldsymbol{\lambda}}_t) + \big\langle \nabla h(\widehat{\boldsymbol{\lambda}}_t),\, \widehat{\boldsymbol{\lambda}}^* - \widehat{\boldsymbol{\lambda}}_t \big\rangle + D_h(\widehat{\boldsymbol{\lambda}}^*, \widehat{\boldsymbol{\lambda}}_t).$$

Since $\widehat{\boldsymbol{\lambda}}_t$ maximizes h, by Fact 2.4, the linear term is nonpositive, and we have

$$\begin{aligned}
h(\widehat{\boldsymbol{\lambda}}^*) \;&\leq h(\widehat{\boldsymbol{\lambda}}_t) + D_h(\widehat{\boldsymbol{\lambda}}^*, \widehat{\boldsymbol{\lambda}}_t) \\
&= a_t L(\boldsymbol{w}_t, \widehat{\boldsymbol{\lambda}}_t) - (\nu_0 + \nu A_{t-1})\, D_\phi(\widehat{\boldsymbol{\lambda}}_t, \widehat{\boldsymbol{\lambda}}_{t-1}) - (\nu_0 + \nu A_t)\, D_\phi(\widehat{\boldsymbol{\lambda}}^*, \widehat{\boldsymbol{\lambda}}_t),
\end{aligned}$$

where in the last step we used Fact 2.5 to replace $D_h(\widehat{\boldsymbol{\lambda}}^*, \widehat{\boldsymbol{\lambda}}_t)$ by the appropriate shift in the Bregman divergence D_ϕ.

Putting these pieces together gives

$$\begin{aligned}
a_t L(\boldsymbol{w}_t, \widehat{\boldsymbol{\lambda}}^*) \;\leq\; & a_t L(\boldsymbol{w}_t, \widehat{\boldsymbol{\lambda}}_t) \;-\; (\nu_0 + \nu A_{t-1})\, D_\phi(\widehat{\boldsymbol{\lambda}}_t, \widehat{\boldsymbol{\lambda}}_{t-1}) \;-\; (\nu_0 + \nu A_t)\, D_\phi(\widehat{\boldsymbol{\lambda}}^*, \widehat{\boldsymbol{\lambda}}_t) \\
& +\; (\nu_0 + \nu A_{t-1})\, D_\phi(\widehat{\boldsymbol{\lambda}}^*, \widehat{\boldsymbol{\lambda}}_{t-1}).
\end{aligned}$$

as claimed. Notice that the last two terms telescope when summed over t, and the negative term involving $D_\phi(\widehat{\boldsymbol{\lambda}}_t, \widehat{\boldsymbol{\lambda}}_{t-1})$ can be used to absorb error contributions in the overall gap analysis. □

E.2 Proof of Lemma E.2

In this section, we prove the lower bound on $a_t L(\boldsymbol{w}_*, \widehat{\boldsymbol{\lambda}}_t)$. This is the most challenging part of the convergence analysis that motivates and analyzes the extrapolation of the dual variable (Line 4), determines the convergence rate (Claim E.4), and proves the key linearization lemma (Lemma 3.5) that underlies the definition of the surrogate gradient in Line 3. Note that in the analysis below we assume the labels are bounded by M when we expand the definition of $\boldsymbol{v}(\boldsymbol{w}; \boldsymbol{x}, y)$, which is w.l.o.g. due to Fact 2.3 and can be ensured by the appropriate pre-processing of samples, as discussed in Section 2.

Lemma E.2 (Lower Bound for $a_t L(\boldsymbol{w}_*, \widehat{\boldsymbol{\lambda}}_t)$)**.** *Under Assumptions 1.2 and 1.3, if the per-group sample size N/K is sufficiently large (see Lemma C.2), where $\boldsymbol{w}_t$ is the sequence of iterates in Algorithm 1, we have for each $t \geq 1$:*

$$\begin{aligned}
a_t L(\boldsymbol{w}_*, \widehat{\boldsymbol{\lambda}}_t) \geq\ & a_t L(\boldsymbol{w}_t, \widehat{\boldsymbol{\lambda}}_t) - \frac{1 + 0.5c_1 A_{t-1}}{2}\|\boldsymbol{w}_* - \boldsymbol{w}_{t-1}\|_2^2 + \frac{1 + 0.5c_1 A_t}{2}\|\boldsymbol{w}_* - \boldsymbol{w}_t\|_2^2 \\
& - a_{t-1}\sum_{i=1}^{K}(\widehat{\lambda}_{t-1[i]} - \widehat{\lambda}_{t-2[i]})\mathbb{E}_{\widehat{p}_{[i]}}[\langle \boldsymbol{v}(\boldsymbol{w}_{t-2}; \boldsymbol{x}, y), \boldsymbol{w}_* - \boldsymbol{w}_{t-1}\rangle] \\
& + a_t \sum_{i=1}^{K}(\widehat{\lambda}_{t[i]} - \widehat{\lambda}_{t-1[i]})\mathbb{E}_{\widehat{p}_{[i]}}[\langle \boldsymbol{v}(\boldsymbol{w}_{t-1}; \boldsymbol{x}, y), \boldsymbol{w}_* - \boldsymbol{w}_t\rangle] \\
& - \frac{1 + 0.5c_1 A_{t-1}}{4}\|\boldsymbol{w}_{t-2} - \boldsymbol{w}_{t-1}\|_2^2 + \frac{1 + 0.5c_1 A_t}{4}\|\boldsymbol{w}_{t-1} - \boldsymbol{w}_t\|_2^2 \\
& - (\nu_0 + \nu A_{t-2}) D_\phi(\widehat{\boldsymbol{\lambda}}_{t-1}, \widehat{\boldsymbol{\lambda}}_{t-2}) - \frac{28\beta^2 B\widehat{\mathrm{OPT}_{\mathrm{m}}} a_t}{c_1}.
\end{aligned}$$

Proof. Since $L(\boldsymbol{w}_*, \widehat{\boldsymbol{\lambda}}_t) = \sum_{i=1}^K \widehat{\lambda}_{t[i]}\mathbb{E}_{\widehat{p}_{[i]}}[(\sigma(\boldsymbol{w}_* \cdot \boldsymbol{x}) - y)^2] - \nu d_f(\widehat{\boldsymbol{\lambda}}_t, \widehat{\boldsymbol{\lambda}}_0)$, we first expand the square

$$\begin{aligned}
(\sigma(\boldsymbol{w}_* \cdot \boldsymbol{x}) - y)^2 &= ((\sigma(\boldsymbol{w}_{t-1} \cdot \boldsymbol{x}) - y) + (\sigma(\boldsymbol{w}_* \cdot \boldsymbol{x}) - \sigma(\boldsymbol{w}_{t-1} \cdot \boldsymbol{x})))^2 \\
&= (\sigma(\boldsymbol{w}_{t-1} \cdot \boldsymbol{x}) - y)^2 + (\sigma(\boldsymbol{w}_* \cdot \boldsymbol{x}) - \sigma(\boldsymbol{w}_{t-1} \cdot \boldsymbol{x}))^2 \\
&\quad + 2(\sigma(\boldsymbol{w}_{t-1} \cdot \boldsymbol{x}) - y)(\sigma(\boldsymbol{w}_* \cdot \boldsymbol{x}) - \sigma(\boldsymbol{w}_{t-1} \cdot \boldsymbol{x}))
\end{aligned} \tag{26}$$

and obtain

$$\begin{aligned}
L(\boldsymbol{w}_*, \widehat{\boldsymbol{\lambda}}_t) = & \sum_{i=1}^{K}\widehat{\lambda}_{t[i]}\mathbb{E}_{\widehat{p}_{[i]}}[2(\sigma(\boldsymbol{w}_{t-1} \cdot \boldsymbol{x}) - y)(\sigma(\boldsymbol{w}_* \cdot \boldsymbol{x}) - \sigma(\boldsymbol{w}_{t-1} \cdot \boldsymbol{x}))] - \nu d_f(\widehat{\boldsymbol{\lambda}}_t, \widehat{\boldsymbol{\lambda}}_0) \\
& + \sum_{i=1}^{K}\widehat{\lambda}_{t[i]}\mathbb{E}_{\widehat{p}_{[i]}}[(\sigma(\boldsymbol{w}_{t-1} \cdot \boldsymbol{x}) - y)^2] + \sum_{i=1}^{K}\widehat{\lambda}_{t[i]}\mathbb{E}_{\widehat{p}_{[i]}}[(\sigma(\boldsymbol{w}_* \cdot \boldsymbol{x}) - \sigma(\boldsymbol{w}_{t-1} \cdot \boldsymbol{x}))^2].
\end{aligned}$$

We use Lemma 2.2 to bound below the last term, use Lemma 3.5 to bound below the first term, and get

$$\begin{aligned}
L(\boldsymbol{w}_*, \widehat{\boldsymbol{\lambda}}_t) \geq & \sum_{i=1}^{K}\widehat{\lambda}_{t[i]}\mathbb{E}_{\widehat{p}_{[i]}}[\langle \boldsymbol{v}(\boldsymbol{w}_{t-1}; \boldsymbol{x}, y), \boldsymbol{w}_* - \boldsymbol{w}_{t-1}\rangle] - \nu d_f(\widehat{\boldsymbol{\lambda}}_t, \widehat{\boldsymbol{\lambda}}_0) - \frac{24\beta^2 B\widehat{\mathrm{OPT}_{\mathrm{m}}}}{c_1} \\
& + \sum_{i=1}^{K}\widehat{\lambda}_{t[i]}\mathbb{E}_{\widehat{p}_{[i]}}[(\sigma(\boldsymbol{w}_{t-1} \cdot \boldsymbol{x}) - y)^2] + c_1\|\boldsymbol{w}_* - \boldsymbol{w}_{t-1}\|_2^2 - \frac{c_1}{4}\|\boldsymbol{w}_* - \boldsymbol{w}_{t-1}\|_2^2.
\end{aligned}$$

Recall that the extrapolated group weights at $(t-1)$-th iteration are $\bar{\boldsymbol{\lambda}}_{t-1} := \widehat{\boldsymbol{\lambda}}_{t-1} + \frac{a_{t-1}}{a_t}(\widehat{\boldsymbol{\lambda}}_{t-1} - \widehat{\boldsymbol{\lambda}}_{t-2})$, thus

$$\begin{aligned}
L(\boldsymbol{w}_*, \widehat{\boldsymbol{\lambda}}_t) \geq & \sum_{i=1}^{K}(\widehat{\lambda}_{t[i]} - \bar{\lambda}_{t-1[i]})\mathbb{E}_{\widehat{p}_{[i]}}[\langle \boldsymbol{v}(\boldsymbol{w}_{t-1}; \boldsymbol{x}, y), \boldsymbol{w}_* - \boldsymbol{w}_{t-1}\rangle] - \nu d_f(\widehat{\boldsymbol{\lambda}}_t, \widehat{\boldsymbol{\lambda}}_0) \\
& + \sum_{i=1}^{K}\widehat{\lambda}_{t[i]}\mathbb{E}_{\widehat{p}_{[i]}}[(\sigma(\boldsymbol{w}_{t-1} \cdot \boldsymbol{x}) - y)^2] + \frac{3}{4}c_1\|\boldsymbol{w}_* - \boldsymbol{w}_{t-1}\|_2^2 - \frac{24\beta^2 B\widehat{\mathrm{OPT}_{\mathrm{m}}}}{c_1} \\
& + \sum_{i=1}^{K}\bar{\lambda}_{t-1[i]}\mathbb{E}_{\widehat{p}_{[i]}}[\langle \boldsymbol{v}(\boldsymbol{w}_{t-1}; \boldsymbol{x}, y), \boldsymbol{w}_* - \boldsymbol{w}_{t-1}\rangle].
\end{aligned} \tag{27}$$

Now we observe that the last term in (27) is a linearization term defining the primal update. This allows us to carry out a similar proof as for the primal gap upper bound in Lemma E.1, and avoid an implicit dependency issue by choosing $\boldsymbol{v}(\boldsymbol{w}_{t-1}; \boldsymbol{x}, y)$ instead of $\boldsymbol{v}(\boldsymbol{w}_t; \boldsymbol{x}, y)$ to define $\psi(\boldsymbol{w})$, since $\boldsymbol{v}(\boldsymbol{w}_t; \boldsymbol{x}, y)$ depends on $\boldsymbol{w}_t$. Let $\psi(\boldsymbol{w}) = a_t \sum_{i=1}^K \bar{\lambda}_{t-1[i]} \mathbb{E}_{\widehat{p}_{[i]}}[\langle \boldsymbol{v}(\boldsymbol{w}_{t-1}; \boldsymbol{x}, y), \boldsymbol{w}\rangle] + \frac{1+0.5c_1A_t}{2}\|\boldsymbol{w} - \boldsymbol{w}_{t-1}\|_2^2$, then $\boldsymbol{w}_t = \arg\min_{\boldsymbol{w}\in\mathcal{B}(W)} \psi(\boldsymbol{w})$ is the update rule for $\boldsymbol{w}_t$. Since $\psi(\boldsymbol{w})$ is $(1 + 0.5c_1A_{t-1})$-strongly convex in $\boldsymbol{w}$, we have

$$
\begin{aligned}
&a_t \sum_{i=1}^K \bar{\lambda}_{t-1[i]} \mathbb{E}_{\widehat{p}_{[i]}}[\langle \boldsymbol{v}(\boldsymbol{w}_{t-1}; \boldsymbol{x}, y), \boldsymbol{w}_* - \boldsymbol{w}_{t-1}\rangle] \\
\geq\ & a_t \sum_{i=1}^K \bar{\lambda}_{t-1[i]} \mathbb{E}_{\widehat{p}_{[i]}}[\langle \boldsymbol{v}(\boldsymbol{w}_{t-1}; \boldsymbol{x}, y), \boldsymbol{w}_t - \boldsymbol{w}_{t-1}\rangle] \\
&+ \frac{1+0.5c_1A_t}{2}\|\boldsymbol{w}_t - \boldsymbol{w}_{t-1}\|_2^2 - \frac{1+0.5c_1A_{t-1}}{2}\|\boldsymbol{w}_* - \boldsymbol{w}_{t-1}\|_2^2 \\
&+ \frac{1+0.5c_1A_t}{2}\|\boldsymbol{w}_* - \boldsymbol{w}_t\|_2^2 - \frac{c_1a_t}{4}\|\boldsymbol{w}_* - \boldsymbol{w}_{t-1}\|_2^2. \qquad (28)
\end{aligned}
$$

Multiplying both sides by a_t and plugging Equation (28) into the last term in RHS in Equation (27), we get

$$
\begin{aligned}
a_tL(\boldsymbol{w}_*, \widehat{\boldsymbol{\lambda}}_t) \geq\ & a_t \sum_{i=1}^K (\widehat{\lambda}_{t[i]} - \bar{\lambda}_{t-1[i]}) \mathbb{E}_{\widehat{p}_{[i]}}[\langle \boldsymbol{v}(\boldsymbol{w}_{t-1}; \boldsymbol{x}, y), \boldsymbol{w}_* - \boldsymbol{w}_{t-1}\rangle] - \nu a_t d_f(\widehat{\boldsymbol{\lambda}}_t, \widehat{\boldsymbol{\lambda}}_0) - a_t \frac{24\beta^2 B\widehat{\mathrm{OPT}_{\mathrm{m}}}}{c_1} \\
&+ a_t \sum_{i=1}^K \widehat{\lambda}_{t[i]} \mathbb{E}_{\widehat{p}_{[i]}}[(\sigma(\boldsymbol{w}_{t-1} \cdot \boldsymbol{x}) - y)^2] + \frac{3}{4}c_1a_t\|\boldsymbol{w}_* - \boldsymbol{w}_{t-1}\|_2^2 \\
&+ a_t \sum_{i=1}^K \bar{\lambda}_{t-1[i]} \mathbb{E}_{\widehat{p}_{[i]}}[\langle \boldsymbol{v}(\boldsymbol{w}_{t-1}; \boldsymbol{x}, y), \boldsymbol{w}_t - \boldsymbol{w}_{t-1}\rangle] \\
&+ \frac{1+0.5c_1A_t}{2}\|\boldsymbol{w}_t - \boldsymbol{w}_{t-1}\|_2^2 - \frac{1+0.5c_1A_{t-1}}{2}\|\boldsymbol{w}_* - \boldsymbol{w}_{t-1}\|_2^2 \\
&+ \frac{1+0.5c_1A_t}{2}\|\boldsymbol{w}_* - \boldsymbol{w}_t\|_2^2 - \frac{c_1a_t}{4}\|\boldsymbol{w}_* - \boldsymbol{w}_{t-1}\|_2^2.
\end{aligned}
$$

Recalling that, by definition, $L(\boldsymbol{w}_t, \widehat{\boldsymbol{\lambda}}_t) = \sum_{i=1}^K \widehat{\lambda}_{t[i]} \mathbb{E}_{\widehat{p}_{[i]}}[(\sigma(\boldsymbol{w}_t \cdot \boldsymbol{x}) - y)^2]$, we further have:

$$
\begin{aligned}
a_tL(\boldsymbol{w}_*, \widehat{\boldsymbol{\lambda}}_t) \geq\ & a_tL(\boldsymbol{w}_t, \widehat{\boldsymbol{\lambda}}_t) + a_t \sum_{i=1}^K (\widehat{\lambda}_{t[i]} - \bar{\lambda}_{t-1[i]}) \mathbb{E}_{\widehat{p}_{[i]}}[\langle \boldsymbol{v}(\boldsymbol{w}_{t-1}; \boldsymbol{x}, y), \boldsymbol{w}_* - \boldsymbol{w}_{t-1}\rangle] \\
&+ a_t \sum_{i=1}^K \widehat{\lambda}_{t[i]} \mathbb{E}_{\widehat{p}_{[i]}}[(\sigma(\boldsymbol{w}_{t-1} \cdot \boldsymbol{x}) - y)^2 - (\sigma(\boldsymbol{w}_t \cdot \boldsymbol{x}) - y)^2] \\
&+ a_t \sum_{i=1}^K \bar{\lambda}_{t-1[i]} \mathbb{E}_{\widehat{p}_{[i]}}[\langle \boldsymbol{v}(\boldsymbol{w}_{t-1}; \boldsymbol{x}, y), \boldsymbol{w}_t - \boldsymbol{w}_{t-1}\rangle] \\
&+ \frac{1+0.5c_1A_t}{2}\|\boldsymbol{w}_t - \boldsymbol{w}_{t-1}\|_2^2 - \frac{1+0.5c_1A_{t-1}}{2}\|\boldsymbol{w}_* - \boldsymbol{w}_{t-1}\|_2^2 \\
&+ \frac{1+0.5c_1A_t}{2}\|\boldsymbol{w}_* - \boldsymbol{w}_t\|_2^2 + \frac{c_1a_t}{2}\|\boldsymbol{w}_* - \boldsymbol{w}_{t-1}\|_2^2 - a_t \frac{24\beta^2 B\widehat{\mathrm{OPT}_{\mathrm{m}}}}{c_1}. \qquad (29)
\end{aligned}
$$

Now we need to deal with the second, the third, and the fourth term in RHS of Equation (29). For the second term, we

carry out the following decomposition

$$
\begin{aligned}
& a_t \sum_{i=1}^{K} (\widehat{\lambda}_{t[i]} - \bar{\lambda}_{t-1[i]}) \mathbb{E}_{\widehat{p}_{[i]}}[\langle \boldsymbol{v}(\boldsymbol{w}_{t-1}; \boldsymbol{x}, y), \boldsymbol{w}_* - \boldsymbol{w}_{t-1} \rangle] \\
= & \, a_t \sum_{i=1}^{K} (\widehat{\lambda}_{t[i]} - \widehat{\lambda}_{t-1[i]}) \mathbb{E}_{\widehat{p}_{[i]}}[\langle \boldsymbol{v}(\boldsymbol{w}_{t-1}; \boldsymbol{x}, y), \boldsymbol{w}_* - \boldsymbol{w}_{t-1} \rangle] \\
& - a_{t-1} \sum_{i=1}^{K} (\widehat{\lambda}_{t-1[i]} - \widehat{\lambda}_{t-2[i]}) \mathbb{E}_{\widehat{p}_{[i]}}[\langle \boldsymbol{v}(\boldsymbol{w}_{t-2}; \boldsymbol{x}, y), \boldsymbol{w}_* - \boldsymbol{w}_{t-2} \rangle] \\
& + a_{t-1} \sum_{i=1}^{K} (\widehat{\lambda}_{t-1[i]} - \widehat{\lambda}_{t-2[i]}) \mathbb{E}_{\widehat{p}_{[i]}}[\langle \boldsymbol{v}(\boldsymbol{w}_{t-2}; \boldsymbol{x}, y) - \boldsymbol{v}(\boldsymbol{w}_{t-1}; \boldsymbol{x}, y), \boldsymbol{w}_* - \boldsymbol{w}_{t-1} \rangle] \\
& + a_{t-1} \sum_{i=1}^{K} (\widehat{\lambda}_{t-1[i]} - \widehat{\lambda}_{t-2[i]}) \mathbb{E}_{\widehat{p}_{[i]}}[\langle \boldsymbol{v}(\boldsymbol{w}_{t-2}; \boldsymbol{x}, y), \boldsymbol{w}_{t-1} - \boldsymbol{w}_{t-2} \rangle], \qquad (30)
\end{aligned}
$$

where the last equality follows the definition of $\bar{\boldsymbol{\lambda}}_{t-1}$. Since the first two terms telescope, we only need to focus on bounding the last two terms in Equation (30). For the second to last term in Equation (30), we have

$$
\begin{aligned}
& - a_{t-1} \sum_{i=1}^{K} (\widehat{\lambda}_{t-1[i]} - \widehat{\lambda}_{t-2[i]}) \mathbb{E}_{\widehat{p}_{[i]}}[\langle \boldsymbol{v}(\boldsymbol{w}_{t-2}; \boldsymbol{x}, y) - \boldsymbol{v}(\boldsymbol{w}_{t-1}; \boldsymbol{x}, y), \boldsymbol{w}_* - \boldsymbol{w}_{t-1} \rangle] \\
\leq & \, 2a_{t-1} \max_{i \in [K]} |\mathbb{E}_{\widehat{p}_{[i]}}[2\beta(\sigma(\boldsymbol{w}_{t-2} \cdot \boldsymbol{x}) - \sigma(\boldsymbol{w}_{t-1} \cdot \boldsymbol{x}))(\boldsymbol{w}_* \cdot \boldsymbol{x} - \boldsymbol{w}_{t-1} \cdot \boldsymbol{x})]| \\
\leq & \, 4a_{t-1}\beta \sqrt{6\beta^2 B \|\boldsymbol{w}_{t-2} - \boldsymbol{w}_{t-1}\|_2^2 \cdot 6B \|\boldsymbol{w}_* - \boldsymbol{w}_{t-1}\|_2^2} \\
\leq & \, \frac{24 \cdot 36\beta^4 B^2 a_{t-1}}{c_1} \|\boldsymbol{w}_{t-2} - \boldsymbol{w}_{t-1}\|_2^2 + \frac{c_1 a_t}{6} \|\boldsymbol{w}_* - \boldsymbol{w}_{t-1}\|_2^2, \qquad (31)
\end{aligned}
$$

where in the first inequality, we applied the definition of $\boldsymbol{v}(\boldsymbol{w}; \boldsymbol{x}, y)$ and Hölder's inequality with the fact that $\sum_{i=1}^{K} |\widehat{\lambda}_{t-1[i]} - \widehat{\lambda}_{t-2[i]}| \leq 2$. The second inequality follows from Lemma 2.2 and Cauchy-Schwarz inequality, the third follows from Young's inequality and the fact that $a_{t-1} \leq a_t$ for all $t \geq 1$.
For the last term in Equation (30), by Cauchy-Schwarz inequality and Lemma 2.2, we have

$$
\begin{aligned}
& - a_{t-1} \sum_{i=1}^{K} (\widehat{\lambda}_{t-1[i]} - \widehat{\lambda}_{t-2[i]}) \mathbb{E}_{\widehat{p}_{[i]}}[\langle \boldsymbol{v}(\boldsymbol{w}_{t-2}; \boldsymbol{x}, y), \boldsymbol{w}_{t-1} - \boldsymbol{w}_{t-2} \rangle] \\
\leq & \, 2\beta a_{t-1} \max_{i \in [K]} |\mathbb{E}_{\widehat{p}_{[i]}}[(\sigma(\boldsymbol{w}_{t-2} \cdot \boldsymbol{x}) - y)(\boldsymbol{w}_{t-1} \cdot \boldsymbol{x} - \boldsymbol{w}_{t-2} \cdot \boldsymbol{x})]| \sum_{i=1}^{K} |\widehat{\lambda}_{t-1[i]} - \widehat{\lambda}_{t-2[i]}| \\
\leq & \, 2\beta a_{t-1} \max_{i \in [K]} \sqrt{\mathbb{E}_{\widehat{p}_{[i]}}[(\sigma(\boldsymbol{w}_{t-2} \cdot \boldsymbol{x}) - y)^2] \cdot 6B \|\boldsymbol{w}_{t-1} - \boldsymbol{w}_{t-2}\|_2^2} \sum_{i=1}^{K} |\widehat{\lambda}_{t-1[i]} - \widehat{\lambda}_{t-2[i]}|. \qquad (32)
\end{aligned}
$$

Moreover, following again from Lemma 2.2 and Fact 2.3, for all groups $i \in [K]$ and any $\boldsymbol{w} \in \mathscr{B}(3\|\boldsymbol{w}_*\|_2)$,

$$
\begin{aligned}
\mathbb{E}_{\widehat{p}_{[i]}}[(\sigma(\boldsymbol{w} \cdot \boldsymbol{x}) - y)^2] &\leq 2(\mathbb{E}_{\widehat{p}_{[i]}}[(\sigma(\boldsymbol{w} \cdot \boldsymbol{x}))^2] + \mathbb{E}_{\widehat{p}_{[i]}}[y^2]) \\
&= 2\big(6\beta^2 W^2 B + C_M^2 W^2 B^2 \log^2(\frac{\beta B W}{\epsilon})\big).
\end{aligned}
$$

Plugging this bound into (32) above, and recalling that $C_W = \sqrt{6\beta^2 + C_M^2 B \log^2(\frac{\beta BW}{\epsilon})}$, we get

$$
\begin{aligned}
&- a_{t-1}\sum_{i=1}^{K}(\widehat{\lambda}_{t-1[i]} - \widehat{\lambda}_{t-2[i]})\mathbb{E}_{\widehat{p}_{[i]}}[\langle \boldsymbol{v}(\boldsymbol{w}_{t-2};\boldsymbol{x},y), \boldsymbol{w}_{t-1}-\boldsymbol{w}_{t-2}\rangle] \\
&\le 2a_{t-1}\beta\sqrt{2W^2B\big(6\beta^2 + C_M^2 B\log^2(\frac{\beta BW}{\epsilon})\big)}\sqrt{6B}\|\boldsymbol{w}_{t-1}-\boldsymbol{w}_{t-2}\|_2\sum_{i=1}^{K}|\widehat{\lambda}_{t-1[i]} - \widehat{\lambda}_{t-2[i]}| \\
&\le 4\sqrt{3}a_{t-1}C_W\beta WB\|\boldsymbol{w}_{t-1}-\boldsymbol{w}_{t-2}\|_2\sum_{i=1}^{K}|\widehat{\lambda}_{t-1[i]} - \widehat{\lambda}_{t-2[i]}| \\
&\le 4\sqrt{3}C_W\beta WBa_{t-1}(\frac{\alpha_1}{2}2D_\phi(\widehat{\boldsymbol{\lambda}}_{t-1},\widehat{\boldsymbol{\lambda}}_{t-2}) + \frac{1}{2\alpha_1}\|\boldsymbol{w}_{t-1}-\boldsymbol{w}_{t-2}\|_2^2),
\end{aligned} \tag{33}
$$

where we apply Young's inequality (Fact B.1) and Claim E.3 (see below) for the last inequality.

To bound the third term on the RHS of Equation (29), we first use a similar method as in Equation (26) to expand the difference of the square terms in the expectation to get:

$$
\begin{aligned}
&a_t\sum_{i=1}^{K}\widehat{\lambda}_{t[i]}\mathbb{E}_{\widehat{p}_{[i]}}[(\sigma(\boldsymbol{w}_t\cdot\boldsymbol{x})-y)^2 - (\sigma(\boldsymbol{w}_{t-1}\cdot\boldsymbol{x})-y)^2] \\
&\le a_t\max_{i\in[K]}\mathbb{E}_{\widehat{p}_{[i]}}[(\sigma(\boldsymbol{w}_t\cdot\boldsymbol{x})-\sigma(\boldsymbol{w}_{t-1}\cdot\boldsymbol{x}))^2] \\
&\quad + \ a_t\max_{i\in[K]}|\mathbb{E}_{\widehat{p}_{[i]}}[2(\sigma(\boldsymbol{w}_{t-1}\cdot\boldsymbol{x})-y)(\sigma(\boldsymbol{w}_t\cdot\boldsymbol{x})-\sigma(\boldsymbol{w}_{t-1}\cdot\boldsymbol{x}))]| \\
&\le 6a_t\beta^2B\|\boldsymbol{w}_t-\boldsymbol{w}_{t-1}\|_2^2 + 2a_t\beta\max_{i\in[K]}|\mathbb{E}_{\widehat{p}_{[i]}}[(\sigma(\boldsymbol{w}_{t-1}\cdot\boldsymbol{x})-y)(\boldsymbol{w}_t\cdot\boldsymbol{x}-\boldsymbol{w}_{t-1}\cdot\boldsymbol{x}))]|,
\end{aligned} \tag{34}
$$

where the second inequality follows from Lemma 2.2 and Lipschitzness of $\sigma(\cdot)$. We split the second term in RHS above and apply the definition of $\widehat{\mathrm{OPT}_{\mathrm{m}}}$ together with the Cauchy-Schwarz inequality, respectively, for all groups $i \in [K]$,

$$
\begin{aligned}
&|\mathbb{E}_{\widehat{p}_{[i]}}[(\sigma(\boldsymbol{w}_{t-1}\cdot\boldsymbol{x})-y)(\boldsymbol{w}_t\cdot\boldsymbol{x}-\boldsymbol{w}_{t-1}\cdot\boldsymbol{x}))]| \\
&\le |\mathbb{E}_{\widehat{p}_{[i]}}[(\sigma(\boldsymbol{w}_*\cdot\boldsymbol{x})-y)(\boldsymbol{w}_t\cdot\boldsymbol{x}-\boldsymbol{w}_{t-1}\cdot\boldsymbol{x})]| + |\mathbb{E}_{\widehat{p}_{[i]}}[(\sigma(\boldsymbol{w}_{t-1}\cdot\boldsymbol{x})-\sigma(\boldsymbol{w}_*\cdot\boldsymbol{x}))(\boldsymbol{w}_t\cdot\boldsymbol{x}-\boldsymbol{w}_{t-1}\cdot\boldsymbol{x}))]| \\
&\le \sqrt{\widehat{\mathrm{OPT}_{\mathrm{m}}}\cdot 6B\|\boldsymbol{w}_t-\boldsymbol{w}_{t-1}\|_2^2} + \sqrt{6\beta^2B\|\boldsymbol{w}_{t-1}-\boldsymbol{w}_*\|_2^2\cdot 6B\|\boldsymbol{w}_t-\boldsymbol{w}_{t-1}\|_2^2}.
\end{aligned}
$$

Combining with Equation (34) and applying Young's inequality (Fact B.1), we have

$$
\begin{aligned}
&a_t\sum_{i=1}^{K}\widehat{\lambda}_{t[i]}\mathbb{E}_{\widehat{p}_{[i]}}[(\sigma(\boldsymbol{w}_t\cdot\boldsymbol{x})-y)^2 - (\sigma(\boldsymbol{w}_{t-1}\cdot\boldsymbol{x})-y)^2] \\
&\le 6\beta^2Ba_t\|\boldsymbol{w}_t-\boldsymbol{w}_{t-1}\|_2^2 + 12a_t\beta^2B\Big(\frac{36\beta^2B}{2c_1}\|\boldsymbol{w}_t-\boldsymbol{w}_{t-1}\|_2^2 + \frac{c_1}{2\cdot 36\beta^2B}\|\boldsymbol{w}_{t-1}-\boldsymbol{w}_*\|_2^2\Big) \\
&\quad + \ 2a_t\Big(\frac{c_1}{2\beta^2B}6\beta^2B\|\boldsymbol{w}_t-\boldsymbol{w}_{t-1}\|_2^2 + \frac{\beta^2B}{2c_1}\widehat{\mathrm{OPT}_{\mathrm{m}}}\Big).
\end{aligned} \tag{35}
$$

Now we bound the fourth term in Equation (29). Noting that $\sum_{i=1}^K |\bar{\lambda}_{t-1[i]}| \le 3$ since $a_{t-1} \le a_t$ for any $t \ge 0$, we have

$$
\begin{aligned}
&- a_t \sum_{i=1}^K \bar{\lambda}_{t-1[i]} \mathbb{E}_{\widehat{p}_{[i]}}[\langle \boldsymbol{v}(\boldsymbol{w}_{t-1}; \boldsymbol{x}, y), \boldsymbol{w}_t - \boldsymbol{w}_{t-1}\rangle] \\
&\le 6\beta a_t \max_{i\in[K]} |\mathbb{E}_{\widehat{p}_{[i]}}[(\sigma(\boldsymbol{w}_{t-1}\cdot\boldsymbol{x}) - \sigma(\boldsymbol{w}_*\cdot\boldsymbol{x}))(\boldsymbol{w}_t\cdot\boldsymbol{x} - \boldsymbol{w}_{t-1}\cdot\boldsymbol{x})| \\
&\quad + 6\beta a_t \max_{i\in[K]} |\mathbb{E}_{\widehat{p}_{[i]}}[(\sigma(\boldsymbol{w}_*\cdot\boldsymbol{x}) - y)(\boldsymbol{w}_t\cdot\boldsymbol{x} - \boldsymbol{w}_{t-1}\cdot\boldsymbol{x})| \\
&\le 6\beta a_t \sqrt{6\beta^2 B\|\boldsymbol{w}_{t-1} - \boldsymbol{w}_*\|_2^2 \cdot 6B\|\boldsymbol{w}_t - \boldsymbol{w}_{t-1}\|_2^2} + 6\beta a_t \sqrt{\widehat{\mathrm{OPT}_{\mathrm{m}}} \cdot 6B\|\boldsymbol{w}_t - \boldsymbol{w}_{t-1}\|_2^2} \\
&\le 36\beta^2 B a_t \Big(\frac{c_1}{2\cdot 108\beta^2 B}\|\boldsymbol{w}_{t-1} - \boldsymbol{w}_*\|_2^2 + \frac{108\beta^2 B}{2c_1}\|\boldsymbol{w}_t - \boldsymbol{w}_{t-1}\|_2^2\Big) \\
&\quad + \ 6\beta a_t \Big(\frac{\beta B}{2c_1}\widehat{\mathrm{OPT}_{\mathrm{m}}} + \frac{6Bc_1}{2\beta B}\|\boldsymbol{w}_t - \boldsymbol{w}_{t-1}\|_2^2\Big). \qquad (36)
\end{aligned}
$$

We again here used the definition of $\boldsymbol{v}(\boldsymbol{w}; \boldsymbol{x}, y)$, a similar method as in Equation (26) to split terms, and used Hölder's inequality to get the first inequality, then applied Cauchy-Schwarz inequality in the second inequality, and used Young's inequality (Fact B.1) in the third.

It remains to choose the appropriate α_1 and argue that the step sizes a_t satisfy $4\sqrt{3}C_W\beta BW a_{t-1} \le \nu_0 + \nu A_{t-2}$, $\frac{2\sqrt{3}C_W\beta W B a_{t-1}}{\alpha_1} + \frac{36\cdot 24\beta^4 B^2 a_{t-1}}{c_1} \le \frac{1+0.5c_1A_{t-1}}{4}$, and $(24c_1 + 6\beta^2 B + \frac{36\cdot 60\beta^4 B^2}{c_1})a_t \le \frac{1+0.5c_1A_t}{4}$, where the first inequality is to construct a term to telescope with $-(\nu_0+\nu A_{t-1})D_\phi(\widehat{\boldsymbol{\lambda}}_t, \widehat{\boldsymbol{\lambda}}_{t-1})$ in Lemma E.1, the second and the third inequality are to construct telescoping terms $\|\boldsymbol{w}_t - \boldsymbol{w}_{t-1}\|_2^2$ and $\|\boldsymbol{w}_{t-1} - \boldsymbol{w}_{t-2}\|_2^2$. We use Claim E.4 below to find and justify an appropriate choice of α_1 and a_t. Combining Equation (30), Equation (31), Equation (33), Equation (35) and Equation (36) together and then substituting them back into Equation (29), we get

$$
\begin{aligned}
a_t L(\boldsymbol{w}_*, \widehat{\boldsymbol{\lambda}}_t) \ge a_t L(\boldsymbol{w}_t, \widehat{\boldsymbol{\lambda}}_t) &- \frac{1+0.5c_1A_{t-1}}{2}\|\boldsymbol{w}_* - \boldsymbol{w}_{t-1}\|_2^2 + \frac{1+0.5c_1A_t}{2}\|\boldsymbol{w}_* - \boldsymbol{w}_t\|_2^2 \\
&- \ a_{t-1}\sum_{i=1}^K (\widehat{\lambda}_{t-1[i]} - \widehat{\lambda}_{t-2[i]})\mathbb{E}_{\widehat{p}_{[i]}}[\langle \boldsymbol{v}(\boldsymbol{w}_{t-2}; \boldsymbol{x}, y), \boldsymbol{w}_* - \boldsymbol{w}_{t-1}\rangle] \\
&+ \ a_t \sum_{i=1}^K (\widehat{\lambda}_{t[i]} - \widehat{\lambda}_{t-1[i]})\mathbb{E}_{\widehat{p}_{[i]}}[\langle \boldsymbol{v}(\boldsymbol{w}_{t-1}; \boldsymbol{x}, y), \boldsymbol{w}_* - \boldsymbol{w}_t\rangle] \\
&- \ \frac{1+0.5c_1A_{t-1}}{4}\|\boldsymbol{w}_{t-2} - \boldsymbol{w}_{t-1}\|_2^2 + \frac{1+0.5c_1A_t}{4}\|\boldsymbol{w}_{t-1} - \boldsymbol{w}_t\|_2^2 \\
&- \ (\nu_0 + \nu A_{t-2})D_\phi(\widehat{\boldsymbol{\lambda}}_{t-1}, \widehat{\boldsymbol{\lambda}}_{t-2}) - \frac{28\beta^2 B\widehat{\mathrm{OPT}_{\mathrm{m}}} a_t}{c_1},
\end{aligned}
$$

which completes our proof. □

Claim E.3 (TV via χ^2/KL on the simplex)**.** *For $\widehat{\boldsymbol{\lambda}}_0 = \frac{\mathbf{1}}{K}$ and any $t_1, t_2 \ge 0$, s.t. $\widehat{\boldsymbol{\lambda}}_{t_1}, \widehat{\boldsymbol{\lambda}}_{t_2} \in \Delta_K$, $\big(\sum_{i=1}^K |\widehat{\lambda}_{t_1[i]} - \widehat{\lambda}_{t_2[i]}|\big)^2 \le 2D_\phi(\widehat{\boldsymbol{\lambda}}_{t_1}, \widehat{\boldsymbol{\lambda}}_{t_2})$, where $\phi = KL(\cdot, \widehat{\boldsymbol{\lambda}}_0)$ or $\phi = \chi^2(\cdot, \widehat{\boldsymbol{\lambda}}_0)$.*

Proof. **Case 1:** $\phi = KL(\cdot, \widehat{\boldsymbol{\lambda}}_0)$. By Pinsker's inequality (Lemma B.4),

$$
(\sum_{i=1}^K |\widehat{\lambda}_{t_1[i]} - \widehat{\lambda}_{t_2[i]}|)^2 = \|\widehat{\boldsymbol{\lambda}}_{t_1} - \widehat{\boldsymbol{\lambda}}_{t_2}\|_1^2 \ \le \ \big(2\,TV(\widehat{\boldsymbol{\lambda}}_{t_1}, \widehat{\boldsymbol{\lambda}}_{t_2})\big)^2 \ \le \ 2\,KL(\widehat{\boldsymbol{\lambda}}_{t_1}, \widehat{\boldsymbol{\lambda}}_{t_2}) \ = \ 2\,D_{\mathrm{KL}(\cdot, \widehat{\boldsymbol{\lambda}}_0)}(\widehat{\boldsymbol{\lambda}}_{t_1}, \widehat{\boldsymbol{\lambda}}_{t_2}),
$$

so altogether

$$
(\sum_{i=1}^K |\widehat{\lambda}_{t_1[i]} - \widehat{\lambda}_{t_2[i]}|)^2 \le 2D_\phi(\widehat{\boldsymbol{\lambda}}_{t_1}, \widehat{\boldsymbol{\lambda}}_{t_2}).
$$

Case 2: $\phi = \chi^2(\cdot, \widehat{\boldsymbol{\lambda}}_0)$. Alternatively, divide and multiply each coordinate by $\widehat{\lambda}_{0[i]}$ and use Jensen's inequality,

$$
(\sum_{i=1}^K |(\widehat{\lambda}_{t_1[i]} - \widehat{\lambda}_{t_2[i]})|)^2 = (\sum_{i=1}^K |\frac{\widehat{\lambda}_{t_1[i]} - \widehat{\lambda}_{t_2[i]}}{\widehat{\lambda}_{0[i]}}|\ \widehat{\lambda}_{0[i]})^2 \le \sum_{i=1}^K \Big(\frac{\widehat{\lambda}_{t_1[i]} - \widehat{\lambda}_{t_2[i]}}{\widehat{\lambda}_{0[i]}}\Big)^2 \widehat{\lambda}_{0[i]} = D_{\chi^2(\cdot, \widehat{\boldsymbol{\lambda}}_0)}(\widehat{\boldsymbol{\lambda}}_{t_1}, \widehat{\boldsymbol{\lambda}}_{t_2}).
$$

Since $D_{\chi^2} \le 2D_{\chi^2}$ trivially, this completes the bound $(\sum_{i=1}^K |\widehat{\lambda}_{t_{[i]}} - \widehat{\lambda}_{t_2[i]}|)^2 \le 2D_\phi(\widehat{\boldsymbol{\lambda}}_{t_1}, \widehat{\boldsymbol{\lambda}}_{t_2})$. □

Claim E.4 (Convergence Rate)**.** *For all $t \ge 0$, let a_t be defined as in Line 2, then it holds that $4\sqrt{3}C_W\beta BWa_{t-1} \le \nu_0 + \nu A_{t-2}$, $\frac{2\sqrt{3}C_W\beta WBa_{t-1}}{\alpha_1} + \frac{36\cdot 24\beta^4B^2a_{t-1}}{c_1} \le \frac{1+0.5c_1A_{t-1}}{4}$, and $(24c_1 + 6\beta^2B + \frac{36\cdot 60\beta^4B^2}{c_1})a_t \le \frac{1+0.5c_1A_t}{4}$. Note that the choice of a_t also makes the last inequality of Equation* (25) *hold.*

Proof. By Young's inequality (Fact B.1), $6\beta^2B \le 3c_1 + 3\beta^2B^4/c_1$, which means that to make the third inequality hold, it suffices that

$$(27c_1 + \frac{2163\beta^4B^2}{c_1})a_t \le \frac{1+0.5c_1A_t}{4}.$$

Let $C_4 = 27c_1 + 2163\beta^4B^2/c_1$, it suffices to enforce

$$a_t \le (1 + \frac{c_1}{8C_4})^{t-1}\frac{1}{4C_4}. \tag{37}$$

For the second inequality to hold, it suffices to enforce the two inequalities below:

$$\frac{2\sqrt{3}C_W\beta WBa_t}{\alpha_1} \le \frac{1+0.5c_1A_t}{8}, \quad \frac{864\beta^4B^2}{c_1}a_t \le \frac{1+0.5c_1A_t}{8},$$

where the second one holds if the first inequality listed in the proof holds. Therefore, to satisfy both the first and the second inequalities in the statement of the lemma, let $C'_W = 2\sqrt{3}C_W\beta WB$, it suffices:

$$2C'_W\alpha_1a_t \le \nu_0 + \nu A_{t-1}, \quad \frac{C'_Wa_t}{\alpha_1} \le \frac{1+0.5c_1A_t}{8}.$$

For each iteration t, we choose $\alpha_1 = 2\sqrt{(\nu_0 + \nu A_{t-1})/(2 + c_1A_t)}$ and it suffices:

$$a_t \le \frac{1}{4\sqrt{2}C'_W}\sqrt{(\nu_0 + \nu A_{t-1})(2 + c_1A_t)}. \tag{38}$$

Using $A_{t-1} \le A_t$ again, it suffices that $a_1 \le \sqrt{\nu_0}/(4C'_W)$ and $a_t \le \sqrt{c_1\nu}A_{t-1}/(4\sqrt{2}C'_W)$ for $t \ge 2$, which means

$$a_t \le (1 + \frac{\sqrt{c_1\nu}}{4\sqrt{2}C'_W})^{t-1}\frac{\sqrt{\nu_0}}{4C'_W},$$

Note that when $\nu = 0$, to satisfy Equation (38), it also suffices $a_t^2 \le \nu_0c_1A_t/(4\sqrt{2}C'_W)^2$, which means it suffices to have $a_t \le \nu_0c_1t/(4\sqrt{2}C'_W)^2$. Therefore, together with Equation (37), it suffices

$$a_t = \min\left\{\left(1 + \frac{c_1}{8C_4}\right)^{t-1}\frac{1}{4C_4}, \max\left\{\left(1 + \frac{\sqrt{c_1\nu}}{4\sqrt{2}C'_W}\right)^{t-1}\frac{\sqrt{\nu_0}}{4C'_W}, \frac{c_1\nu_0}{(4\sqrt{2}C'_W)^2}t\right\}\right\}, \quad A_n = \sum_{t=0}^n a_t.$$

The analysis above provides a valid choice for a_t that satisfies the conditions of the claim, which concludes the proof. □

E.3 Proof of Lemma 3.3

Finally, we inductively establish the boundedness of the iterates that is necessary for the sharpness results (Fact 2.1 and Lemma 2.2).

Lemma 3.3. *For all iterations $t \ge 0$ of Algorithm 1, the iterates satisfy $\|\boldsymbol{w}_t\|_2 \le 3\|\boldsymbol{w}_*\|_2$.*

Proof. We prove this lemma by induction on t. When $t = 0$, it is trivial that $\mathbf{0} = \boldsymbol{w}_0 \in \mathscr{B}(3\|\boldsymbol{w}_*\|)$. Now assume $\|\boldsymbol{w}_t\| \le 3\|\boldsymbol{w}_*\|$ holds for all $0 \le t \le n$, we would like to prove $\|\boldsymbol{w}_{n+1}\| \le 3\|\boldsymbol{w}_*\|$ also holds. We apply Lemma 3.2 and Proposition 3.4 to sandwich the quantity $\sum_{t=1}^{n+1} a_t\mathrm{Gap}(\boldsymbol{w}_t, \widehat{\boldsymbol{\lambda}}_t)$. Note that we only have $\|\boldsymbol{w}_n\| \le 3\|\boldsymbol{w}_*\|$, which means

that we can only add the term $a_{n+1}\mathrm{Gap}(\boldsymbol{w}_{n+1}, \widehat{\boldsymbol{\lambda}}_{n+1})$ to the lower bound in Lemma 3.2. However, we can directly apply Proposition 3.4 for the $(n+1)$-th iteration since it only requires $\|\boldsymbol{w}_n\| \leq 3\|\boldsymbol{w}_*\|$. Therefore, we get the inequality below:

$$
\begin{aligned}
&-\frac{12\beta^2 B}{c_1}\widehat{\mathrm{OPT}_{\mathrm{m}}}A_n + \sum_{t=1}^{n} a_t \frac{c_1}{2}\|\boldsymbol{w}_t - \boldsymbol{w}_*\|_2^2 + \sum_{t=1}^{n} \nu a_t D_\phi(\widehat{\boldsymbol{\lambda}}^*, \widehat{\boldsymbol{\lambda}}_t) + a_{n+1}\mathrm{Gap}(\boldsymbol{w}_{n+1}, \widehat{\boldsymbol{\lambda}}_{n+1}) \\
\leq &\sum_{t=1}^{n+1} a_t \mathrm{Gap}(\boldsymbol{w}_t, \widehat{\boldsymbol{\lambda}}_t) \\
\leq &\frac{1}{2}\|\boldsymbol{w}_* - \boldsymbol{w}_0\|_2^2 + \nu_0 D_\phi(\widehat{\boldsymbol{\lambda}}^*, \widehat{\boldsymbol{\lambda}}_0) - \frac{1 + 0.5c_1 A_{n+1}}{4}\|\boldsymbol{w}_* - \boldsymbol{w}_{k+1}\|_2^2 - (\nu_0 + \nu A_{n+1}) D_\phi(\widehat{\boldsymbol{\lambda}}^*, \widehat{\boldsymbol{\lambda}}_{n+1}) \\
&+ \frac{28\beta^2 B\widehat{\mathrm{OPT}_{\mathrm{m}}}A_{n+1}}{c_1}.
\end{aligned} \tag{39}
$$

Also, similar to Lemma 3.2 (see the proof of Lemma 3.2 in Appendix D), we split $a_{n+1}\mathrm{Gap}(\boldsymbol{w}_{n+1}, \widehat{\boldsymbol{\lambda}}_{n+1})$ into two terms and get

$$
\begin{aligned}
a_{n+1}\mathrm{Gap}(\boldsymbol{w}_{n+1}, \widehat{\boldsymbol{\lambda}}_{n+1}) = &\ [L(\boldsymbol{w}_{n+1}, \widehat{\boldsymbol{\lambda}}^*) - L(\boldsymbol{w}_*, \widehat{\boldsymbol{\lambda}}^*)] + [L(\boldsymbol{w}_*, \widehat{\boldsymbol{\lambda}}^*) - L(\boldsymbol{w}_*, \widehat{\boldsymbol{\lambda}}_{n+1})] \\
= &\ \sum_{i=1}^{K} \widehat{\lambda}_{[i]}^* \mathbb{E}_{\widehat{p}_{[i]}}[((\sigma(\boldsymbol{w}_{n+1}\cdot\boldsymbol{x}) - y)^2 - (\sigma(\boldsymbol{w}_*\cdot\boldsymbol{x}) - y)^2)] + \nu D_\phi(\widehat{\boldsymbol{\lambda}}^*, \widehat{\boldsymbol{\lambda}}_{n+1}) \\
\geq &\ -\widehat{\mathrm{OPT}_{\mathrm{m}}},
\end{aligned}
$$

where we lower bound the first term as $-\widehat{\mathrm{OPT}_{\mathrm{m}}}$ and simply ignore the second term due to the nonnegativity of Bregman divergence. Since $\sum_{i=1}^{n} a_t \frac{c_1}{2}\|\boldsymbol{w}_t - \boldsymbol{w}_*\|_2^2$ and $\sum_{t=1}^{n} \nu a_t D_\phi(\widehat{\boldsymbol{\lambda}}^*, \widehat{\boldsymbol{\lambda}}_t)$in LHS of Equation (39) are nonnegative, $-(\nu_0 + \nu A_{n+1}) D_\phi(\widehat{\boldsymbol{\lambda}}^*, \widehat{\boldsymbol{\lambda}}_{n+1})$ and $(1 + 0.5c_1 A_{n+1})\|\boldsymbol{w}_{n+1} - \boldsymbol{w}_n\|_2^2/4$ in RHS are nonpositive, we can ignore them. Plugging the inequality above into Equation (39) and rearranging the terms, we get

$$
\begin{aligned}
&\frac{2 + c_1 A_{n+1}}{8}\|\boldsymbol{w}_* - \boldsymbol{w}_{n+1}\|_2^2 \\
\leq &\ \frac{1}{2}\|\boldsymbol{w}_* - \boldsymbol{w}_0\|_2^2 + \nu_0 D_\phi(\widehat{\boldsymbol{\lambda}}^*, \widehat{\boldsymbol{\lambda}}_0) + \big(\frac{40\beta^2 B}{c_1}A_n + (1 + \frac{28\beta^2 B}{c_1})a_{n+1}\big)\widehat{\mathrm{OPT}_{\mathrm{m}}}.
\end{aligned}
$$

Let both sides be divided by $(2 + c_1 A_{n+1})/8$, for the first two terms in RHS, we use $2 + c_1 A_{n+1} \geq 2$, for the third term, we use $2 + c_1 A_{n+1} \geq c_1 A_n$, and for the last term, following the choice of a_t, we have $\frac{a_{n+1}}{2 + c_1 A_{n+1}} \leq \frac{a_{n+1}}{c_1 A_n} \leq \max\{\frac{1}{n}, \frac{1}{8C_4}\} \leq 1$. Therefore, the inequality above becomes

$$
\|\boldsymbol{w}_* - \boldsymbol{w}_{n+1}\|_2^2 \leq 2\|\boldsymbol{w}_* - \boldsymbol{w}_0\|_2^2 + 4\nu_0 D_\phi(\widehat{\boldsymbol{\lambda}}^*, \widehat{\boldsymbol{\lambda}}_0) + \big(\frac{544\beta^2 B}{c_1^2} + \frac{1}{\min\{\frac{n}{8}, C_4\}}\big)\widehat{\mathrm{OPT}_{\mathrm{m}}}.
$$

Since $\frac{1}{\min\{\frac{n}{8}, C_4\}}$ is at most of constant order, the coefficient of $\widehat{\mathrm{OPT}_{\mathrm{m}}}$ is also a constant. Choosing $\nu_0 = \frac{\epsilon}{4K}$, following the similar logic of claim E.2 in Li et al. (2024), we can assume without loss of generality that $\big(\frac{544\beta^2 B}{c_1^2} + \frac{1}{\min\{\frac{n}{8}, C_4\}}\big)\widehat{\mathrm{OPT}_{\mathrm{m}}} + \epsilon \leq \|\boldsymbol{w}_*\|_2^2$, otherwise we can compare 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 obtain an $O(\mathrm{OPT}) + \epsilon$ solution.

Now we complete the induction step showing $\|\boldsymbol{w}_* - \boldsymbol{w}_{n+1}\|_2^2 \leq 3\|\boldsymbol{w}_*\|_2^2$, which means $\|\boldsymbol{w}_{n+1}\|_2 \leq (1 + \sqrt{3})\|\boldsymbol{w}_*\|_2 \leq 3\|\boldsymbol{w}_*\|_2$, and we finish the proof. □

F Supplementary Details of Experiments

F.1 Existing Assets Used

We used the publicly available RedPajama dataset (Together Computer, 2023), which is released under a combination of open licenses consistent with the licenses of the original data sources (e.g., CC-BY for Wikipedia). We followed the official RedPajama license statement. Since the data from the `book` domain is no longer publicly available, we instead

downloaded approximately 160GB of raw data from the remaining six domains, using normalized weights based on the original dataset's initial proportions.

For the code, we built on the code base of Xia et al. (2024) (see Xia et al. (*Sheared LLaMA: Accelerating Language Model Pre-training*)), adding our implementation of the primal–dual methods below and modifying the function `update_proportion` in the `dynamic_loading_callback.py` file accordingly.

PD-KL updates

```
elif self.update_type == "pd-kl":
    new_lambdas = torch.log(new_lambdas + 1e-6) + eta * diff
    new_lambdas = torch.nn.functional.softmax(new_lambdas, dim=0)
    updated_domain_weights = \
        new_lambdas + extrapolation_factor * (new_lambdas - torch.tensor(current_lambdas))
            # extrapolation
    updated_domain_weights = (1-c) * updated_domain_weights + c / self.n_domains
```

F.2 Additional Experiment Details

We largely followed the pipelines and instructions provided in the aforementioned code base. For data preparation and model setup, we used virtual machines on Google Cloud Platform (GCP): a VM with 8 vCPUs and 64GB memory for tokenization and data sampling, and a VM with a single NVIDIA A100 80GB GPU for converting checkpoints into HuggingFace format and running model evaluations. For training, we employed 4 NVIDIA A100 80GB GPUs on the high-performance computing clusters of the Center for High Throughput Computing (CHTC) at UW-Madison, equipped with 32 CPUs and 256GB memory. We followed all training parameters from Xia et al. (2024), except that we set the evaluation interval to 8.4M tokens instead of 16.8M tokens, i.e., twice as frequent as in their setup.