Achieving Group Distributional Robustness and Minimax Group Fairness with Interpolating Classifiers

Natalia Martinez Gil[†]
IBM Research

Martin Bertran[†*]
Amazon Science

Guillermo Sapiro
Duke University & Apple

Abstract

Group distributional robustness optimization methods (GDRO) learn models that guarantee performance across a broad set of demographics. GDRO is often framed as a minimax game where an adversary proposes data distributions under which the model performs poorly; importance weights are used to mimic the adversarial distribution on finite samples. Prior work has show that applying GDRO with interpolating classifiers requires strong regularization to generalize to unseen data. Moreover, these classifiers are not responsive to importance weights in the asymptotic training regime. In this work we propose Bi-level GDRO, a provably convergent formulation that decouples the adversary's and model learner's objective and improves generalization guarantees. To address non-responsiveness of importance weights, we combine Bi-level GDRO with a learner that optimizes a temperature-scaled loss that can provably trade off performance between demographics, even on interpolating classifiers. We experimentally demonstrate the effectiveness of our proposed method on learning minimax classifiers on a variety of datasets. Code is available at github.com/MartinBertran/BiLevelGDRO.

1 Introduction

Deep neural networks have proven to be effective tools for classification in practice, achieving high average accuracy. Recent works [Hashimoto et al., 2018, Sagawa et al., 2019] have focused on producing models that are robust to distribution shifts, providing classifiers that perform well on a variety of demographics (sub-groups), regardless of how likely they are to appear on the training dataset [Diana et al., 2020, Martinez et al., 2020, Yang et al., 2023]. This is commonly done via importance weighting which modifies the relative importance of the training samples as a function of their demographic of origin. Importance weighting produces unbiased (albeit potentially high-variance) estimators of the loss over a candidate test distribution from samples drawn from a different training distribution; it allows for simple and robust minimax optimization algorithms, since the weights can be computed based on the train and test demographic priors, and can be adjusted with a variety of training algorithms, e.g., projected gradient ascent (PGA), multiplicative weight updates (MWU) [Chen et al., 2017, Diana et al., 2020].

One common objective to address distribution shifts is to formulate the problem as a minimax optimization game between the model learner and an adversary [Chen et al., 2017]. In this setting, the model learner attempts to minimize the error over a data distribution proposed by the adversary. The adversary in turn proposes data distributions within a set distance of the original training distribution to maximize the model's error; these distributions are realized with importance weights placed on each sample (or groups of samples in settings where demographics are known at training time). While this formulation is theoretically well grounded, it can suffer from two problems. One being the significant generalization errors for both the model learner and the adversary when training in the offline setting [Sagawa et al., 2019], and the other being a lack of responsiveness of the learned model to the importance weights themselves [Byrd and Lipton, 2019, Wang et al., 2021]. Some works have tackled the generalization issue by using regularization tuning on the model learner to produce best results [Sagawa et al., 2019].

[†]Equal contribution
[*]Work unrelated to position at Amazon

Proceedings of the 27[th] International Conference on Artificial Intelligence and Statistics (AISTATS) 2024, Valencia, Spain. PMLR: Volume TBD.

Recent works [Zhou et al., 2022] make use of bi-level optimization to treat the model parameters as dependent variables of the importance weights, and differentiate through the learning process, leading to better generalization errors. We extend their results to minimax group fairness[1] as a Bi-level GDRO objective, with an adversary that makes use of out of sample data to reduce the bias of the group errors estimators. Our formulation does not require differentiation through the learning process, while possessing theoretical generalization guarantees that depend on the adversary's, rather than the learner's, model complexity; the improvement of this approach is validated empirically.

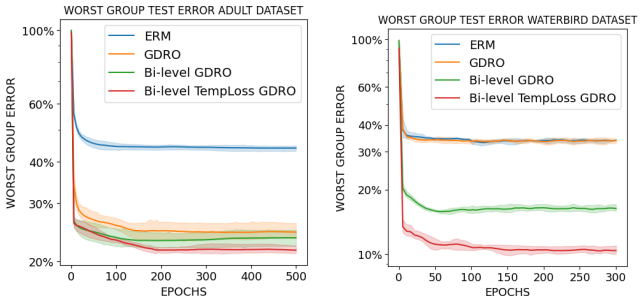
Recent evidence suggests that overparametrized neural networks, able to achieve zero-error over the training set, are not responsive to importance weights asymptotically [2][Byrd and Lipton, 2019, Xu et al., 2021]. The works in [Xu et al., 2021, Soudry et al., 2018, Nacson et al., 2019b], shows that (non-zero) importance weight coefficients are asymptotically ignored, with classifiers converging to max-margin solutions. These observations hold true for exponential tailed functions, which include both crossentropy and logistic regression losses. We modify our proposed Bi-level GDRO formulation to instead use a tempered loss on the learner, where the margin of each group is affected by the adversarial distribution. The resulting method improves worst group performance in the asymptotic training regime.

Main contributions. This work tackles generalization and lack of responsiveness to importance weights in the context of minimax group fairness/group distributional robustness. Representative results are shown in Figure 1. Our main contributions are the following:

- We improve generalization, and provide generalization bounds, by leveraging the bi-level formulation of group DRO (Bi-level GDRO), partially relaxing the error objective, and decoupling the adversary and model learner's datasets. We additionally propose a provably-convergent algorithm to solve our proposed bi-level optimization problem.

- We leverage the flexibility of the Bi-level GDRO formulation and replace the learner's base loss function with a tempered loss (TempLoss). This modification for exponentially-tailed losses (including cross-entropy) is asymptotically responsive to importance weights. TempLoss has an alternative interpretation as a weighted max-margin classifier in the Support Vector Machine (SVM) setting for separable data. We show that the Bi-level TempLoss GDRO objective can be optimized via MWU while keeping its convergence guarantees.

- We experimentally compare our proposed approach against standard worst-case group optimization using importance weighting with cross-entropy losses, as well as the recently proposed polynomial-tailed loss [Wang et al., 2021] and vector scaled loss [Kini et al., 2021].



(a) Adult (b) Waterbirds

Figure 1: Evolution of worst group test error across training epochs on the Adult and Waterbird datasets [Becker and Kohavi, 1996, Sagawa et al., 2019]; deviations computed over 5 splits. We compare ERM; GDRO (adversary uses training group losses to update group priors); our proposed Bi-level GDRO, were the adversary uses the error on a held-out dataset to update group priors; and Bi-level TempLoss GDRO, where the learner additionally minimizes our proposed TempLoss. Bi-level GDRO outperforms regular GDRO, especially over the last training epochs. The TempLoss modification on the learner's objective further improves performance. Results are consistent across several datasets as shown in Section 5.

2 Problem Setting

Consider the supervised classification scenario where we have input features $X \in \mathcal{X}$ and target variable $Y \in \mathcal{Y}$. In the GDRO setting we also assume the existence of a set of groups (i.e., demographics), represented by variable $G \in \mathcal{G}$, that define a set of conditional distributions on the input and target variables $X, Y|G \sim P_{X,Y|G}$, $\forall G \in \mathcal{G}$. The objective of GDRO [Hashimoto et al., 2018] is to learn a score function $f : \mathcal{X} \to \mathbb{R}^{|\mathcal{Y}|}$,[3] from some function family $\mathcal{F}$ that

[1] Here we refer to group distributional robustness and minimax group fairness interchangeably.

[2] Here, asymptotically refers to the limiting behaviour of stochastic gradient descent (SGD) given an infinite number of training iterations over a finite dataset.

[3] To simplify future notation, we take the unnormalized logits or scores to be the output of our model.

minimizes the error of predicting label Y from input X over the worst possible group distribution Q. The distribution Q lies within some predefined set of group distributions $Q \in \mathcal{Q} \subseteq \mathbb{P}_{\mathcal{G}}$, where $\mathbb{P}_{\mathcal{G}}$ represents the set of all possible distributions over $\mathcal{G}$,

$$\min_{f \in \mathcal{F}} \max_{Q \in \mathcal{Q}} \mathbb{E}_{P_{X,Y|G} Q_G}[\epsilon(f(X), Y)]. \tag{1}$$

Here, $\epsilon(f(X), Y) = \mathbf{1}[\arg\max_{y \in \mathcal{Y}}\{f_y(X)\} \neq Y]$ is the error function. We note $Q_G = Pr_{G \sim Q}(G)$,[4] as the probability of the group random variable G under Q.

One challenge of this formulation is optimizing over possible group distributions Q. We assume we are given access to data with groups G distributed according to a prior $P_G \in \mathbb{P}_{\mathcal{G}}$, and make use of importance weights $\frac{Q_G}{P_G}$ to solve Eq.(1). The resulting equation becomes

$$\min_{f \in \mathcal{F}} \max_{Q \in \mathcal{Q}} \mathbb{E}_{P_{X,Y,G}}[\frac{Q_G}{P_G} \epsilon(f(X), Y)]. \tag{2}$$

We focus on the setting where the set of groups $\mathcal{G}$ is discrete and known at training time. In this context, one family of adversarial distributions of interest is $\mathcal{Q}_{\gamma} = \{Q \in \mathbb{P}_{\mathcal{G}} : Q_g \geq \gamma_g, \forall g \in \mathcal{G}\}$, where $\boldsymbol{\gamma} = \{\gamma_g\}_{g \in \mathcal{G}}$ represents the minimum likelihood of each group. This is equivalent to minimizing a linear trade-off between the worst (unconstrained) group distribution and the $\boldsymbol{\gamma}$-weighted group error (i.e., relaxed minimax group objective [Diana et al., 2020]).

In practice, we are given access to a training dataset of i.i.d. samples $\mathcal{D}^{tr} = \{(x_i, y_i, g_i)\}_{i=1}^n \sim P_{X,Y,G}^{\otimes n}$, we denote the set of samples from group g as $\mathcal{D}_g^{tr} = \{(x_i, y_i) \in \mathcal{D}^{tr} : g_i = g\}$. The expectation operator in Eq.(2) can be replaced by its empirical estimate

$$\mathbb{E}_{\mathcal{D}^{tr}}[\frac{Q_G}{P_G} \epsilon(f(X), Y)] \coloneqq \sum_{i \in [\mathcal{D}^{tr}]} \frac{Q_{g_i}}{|\mathcal{D}_{g_i}^{tr}|} \epsilon(f(x_i), y_i). \tag{3}$$

Where necessary, the non-differentiability of the error function w.r.t. the classifier function f is addressed by using a surrogate loss function $\ell : \mathbb{R}^{|\mathcal{Y}|} \times \mathcal{Y} \to \mathbb{R}$ to supplant the error loss (e.g., cross-entropy loss with a softmax non-linearity).[5]

In the standard offline setting (Figure 2) the empirical estimate of the minimax objective in Eq.(2) is obtained by repeatedly solving its maximin lower bound, as well as replacing the error objective with the surrogate loss function ℓ. Here, we present its bi-level formulation,

$$\begin{aligned} \max_{Q \in \mathcal{Q}_{\gamma}} \quad & \textstyle\sum_{g \in \mathcal{G}} Q_g \mathbb{E}_{\mathcal{D}_g^{tr}}[\ell(f^*(X), Y)], \\ & s.t.\ f^* = \arg\min_{f \in \mathcal{F}} \mathbb{E}_{\mathcal{D}^{tr}}[\tfrac{Q_G}{P_G} \ell(f(X), Y)]. \end{aligned} \tag{4}$$

[4] We use Q_g to denote $Pr_{G \sim Q}(G = g)$.

[5] $\ell(f(x), y) = -\log(e^{f_y(x)} / \sum_{i \in \mathcal{Y}} e^{f_i(x)})$.

The maximin objective in Eq.(4) can be solved using multiplicative weight updates on the adversary (or other no-regret algorithms), combined with approximate best response on the model learner [Chen et al., 2017, Sagawa et al., 2019, Diana et al., 2020]. In this scenario, at each optimization round t an adversary updates its worst group distribution $\mathcal{Q}^t \in \Delta^{|\mathcal{G}-1|}$ based on the sequence of empirical group loss estimates incurred by the preceding models $f^0, \ldots, f^{t-1}$ using the training dataset. Then, the learner proposes a new model f^t by (approximately) optimizing the empirical weighted loss on the training samples as in Eq.(3). The game is played for a total of T rounds. Regardless of the model learner's strategy, the adversary is guaranteed to asymptotically discover the best response (in hindsight) to the model learner's approach [Roughgarden, 2016].

There are two key issues that arise from solving the maximin problem in Eq.(4) on finite samples. Both of these problems are especially prevalent when the function family $\mathcal{F}$ is able of to achieve zero classification error in the training dataset. We summarize the issues and proposed solutions next.

P1. Poor generalization of the adversarial prior Q due to over-fitting. Ideally, the methods used to learn the adversarial prior Q should rely on unbiased estimates of the group-conditional expected error of the model. For offline learning, where samples are seen by both the model learner and the adversary several times, these estimates tend to be biased, potentially underestimating group risk, and may not accurately reflect which groups are in need of remedial attention during training. We re-frame the minimax optimization problem using a bi-level formulation to decouple the adversary's and model learner's objectives; this enables us to improve on the empirical generalization of our models, and to directly learn importance weights based on unbiased error estimates instead of training losses. Section 3 shows that our proposed approach converges under mild conditions and give theoretical bounds on generalization error for this framework.

P2. Importance weights are asymptotically ignored in the 0-training-error training regime. SGD on exponential-tailed losses such as cross-entropy or Brier score can produce models with 0 training error [Sagawa et al., 2019, Byrd and Lipton, 2019, Xu et al., 2021, Nacson et al., 2019b]. Since this limit is independent of the importance weights, it means that standard linearly-weighted losses are asymptotically ineffective on perfectly separable training datasets. This is especially relevant in settings like ours where linear weights are an integral part to minimax optimization. Here, we propose that the learner optimizes

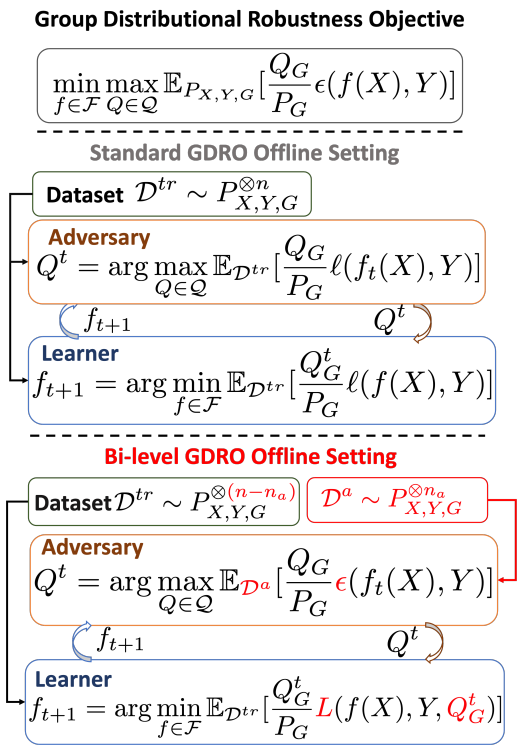


Figure 2: Diagram showing the standard offline approach and the proposed Bi-level offline approach for the GDRO objective. Differences marked in red. In the standard offline approach both the adversary and learner update their corresponding objectives by optimizing the weighted group conditional surrogate loss ℓ over the same empirical dataset $\mathcal{D}^{tr}$. In the proposed Bi-level GDRO approach the adversary directly optimizes the weighted group conditional error estimated using a dataset $\mathcal{D}^a$ that is independent of the one used by the learner. The learner optimizes a differentiable loss L that can take the adversarial group distribution Q as an additional input.

its model using a tempered loss (TempLoss), which modifies exponential tailed losses (e.g., cross entropy) by the incorporation of a multiplicative term on the sample margins that is active on correctly classified samples and responds to the importance weights or group adversarial priors. We present and analyze the properties of the proposed TempLoss in Section 3.4.

In Section 5 we show that the proposed Bi-level GDRO method, and its integration with a learner that minimizes the proposed tempered loss (Bi-level TempLoss GDRO), significantly improves worst group performance in the asymptotic training regime. These results do not rely on additional ad-hoc regularization techniques such as early stopping or weights regularization.

3 Methodology

3.1 Addressing minimax learning

To address the minimax objective in Eq 1 we propose the Bi-level GDRO objective which consists of the following changes: First, the adversary and the learner have access to independent datasets drawn from the same distribution. We use $\mathcal{D}^a \sim P^{\otimes n_a}_{X,Y,G}$ to denote the adversary dataset, and keep $\mathcal{D}^{tr}$ for the learner. Second, the adversary's objective is not replaced by a surrogate loss; it can rely on the group-conditional error estimates obtained with its independent dataset $\mathcal{D}^a$. Finally, we consider a generalized loss objective for the learner $L : \mathbb{R}^{|\mathcal{Y}|} \times \mathcal{Y} \times \mathbb{R}^+ \to \mathbb{R}$ that incorporates an additional optional input. This enables us to optionally incorporate the adversarial prior into the loss function which may prove to be a better surrogate to the original objective of minimizing the weighted group error loss. For example, our proposed TempLoss, described in Section 3.4, tempers the sample margins of the output based on the group prior to better control the error in the overparametrized regime. The proposed Bi-level GDRO objective is

$$\begin{aligned} \max_{Q\in\mathcal{Q}_\gamma} \quad & \sum_{g\in\mathcal{G}} Q_g \mathbb{E}_{\mathcal{D}^a_g}[\epsilon(f(X),Y)], \\ & s.t. f = \arg\min_{f\in\mathcal{F}} \mathbb{E}_{\mathcal{D}^{tr}}[\tfrac{Q_G}{P_G} L(f(X),Y,Q_G)]. \end{aligned} \tag{5}$$

To optimize our Bi-level GDRO objective, we propose Algorithm 1 where the adversary uses a MWU solver and the model learner uses $\alpha-$approximate best response, defined in the following section. Next, we show the convergence guarantees of the proposed approach and the generalization benefits of the independent dataset for the adversary.

3.2 Convergence

To prove the convergence of our approach, shown in Algorithm 1, we first consider the following definition.

Definition 3.1. $M(Q)$ is an $\alpha-$approximate oracle solver for the objective function $\sum_{g\in\mathcal{G}} Q_g \mathbb{E}_{\mathcal{D}^a_g}[\epsilon(f(X),Y)]$ if it produces a candidate function $\hat{f} = M(Q)$ such that

$$\begin{aligned} \sum_{g\in\mathcal{G}} Q_g \mathbb{E}_{\mathcal{D}^a_g}[\epsilon(\hat{f}(X),Y)] \leq \\ \alpha \min_{f\in\mathcal{F}} \sum_{g\in\mathcal{G}} Q_g \mathbb{E}_{\mathcal{D}^a_g}[\epsilon(f(X),Y)]. \end{aligned} \tag{6}$$

We note that $M(Q)$ is an approximate solver of the adversary's objective function. To instantiate this solver, we make the following assumption.

Assumption 3.2. Loss L is such that the solution to $\arg\min_{f\in\mathcal{F}} \mathbb{E}_{\mathcal{D}^{tr}}[\frac{Q_G}{P_G} L(f(X), Y, Q_G)]$ is an α-approximate solution to $\sum_{g\in\mathcal{G}} Q_g \mathbb{E}_{\mathcal{D}_g^a}[\epsilon(f(X), Y)]$.

That is, we assume the surrogate loss function in the inner loop of Eq.(5), computed over a separate dataset D^{tr}, produces high quality solutions to the empirical error objective for a fixed adversarial distribution Q. In these conditions, we can extend the results in [Chen et al., 2017] to show that Algorithm 1 can be used to compute an $\alpha-$approximate solution to our target optimization problem, as shown in Theorem 3.3.

Algorithm 1 Bi-level GDRO MWU Solver

Require: adversary set $\mathcal{D}^a$, constraint $\mathcal{Q}_\gamma$, parameters T, η,
α-approximate stochastic oracle $M(Q) : \mathbb{P}_G \to \mathcal{F}$ for the objective $\sum_{g\in\mathcal{G}} Q_g \epsilon_g(f)$, where

$$\epsilon_g(f) \coloneqq \mathbb{E}_{\mathcal{D}_g^a}[\epsilon(f(X), Y)].$$

Also, let $M(Q)$ be independent of D^a.
for $t = 1, \dots, T$ **do**
$\hat{Q}^t \propto \exp\{\eta \sum_{t'\in[t]} \epsilon_g(f^t)\}_{g\in\mathcal{G}}$,
$Q^t \leftarrow \hat{Q}^t(1 - ||\boldsymbol{\gamma}||_1^1) + \boldsymbol{\gamma}$
$f_t \leftarrow M(Q^t)$,
end for

output $\{f^1, \dots, f^T\}$

Theorem 3.3. *Algorithm 1 with* $\eta = \frac{\sqrt{2\frac{\log|\mathcal{G}|}{T}}}{(1-||\boldsymbol{\gamma}||_1^1)}$ *computes a uniform distribution over solutions* $P_F = U_{[\{f^1,\dots,f^T\}]}$ *such that*

$$\max_{Q\in\mathcal{Q}_\gamma} \mathbb{E}_{f\sim P_F} \sum_{g\in\mathcal{G}} Q_g \epsilon_g(f) \leq \alpha \min_{f\in\mathcal{F}} \max_{Q\in\mathcal{Q}_\gamma} \sum_{g\in\mathcal{G}} Q_g \epsilon_g(f) + \sqrt{\frac{2\log|\mathcal{G}|}{T}}. \quad (7)$$

Further, for any $\eta > 0$*, the solution to Algorithm 1 satisfies*

$$\max_{Q\in\mathcal{Q}_\gamma} \mathbb{E}_{f\sim P_F} \sum_{g\in\mathcal{G}} Q_g \epsilon_g(f) \leq \alpha(1+\eta) \min_{f\in\mathcal{F}} \max_{Q\in\mathcal{Q}_\gamma} \sum_{g\in\mathcal{G}} Q_g \epsilon_g(f) + \frac{\log|\mathcal{G}|}{\eta T}. \quad (8)$$

The MWU rule is a no-regret algorithm and guarantees near-optimal performance of the adversary regardless of the model learner's response strategy. The theorem states that an approximate solution to the current adversary's strategy is sufficient to approximately solve the empirical instantiation of Eq.(2). This gives us a convergent, finite-sample solution to our objective of interest. All proofs are provided in Appendix A. To avoid needing to maintain T classifiers at inference time, we follow [Chen et al., 2017] and [Sagawa et al., 2019] where the deployed model is the final classifier. In our experiments, this approach does not significantly impact performance. In the next section, we explore the generalization properties of this approach to out-of-sample data.

3.3 Generalization

Algorithm 1 requires our α-approximate solver $M(Q)$ to be independent of D^a. This is not a necessary condition for the convergence results shown in Theorem 3.3, but is of high importance in practice since we wish to ensure the performance of our model f not just on a finite set D^a but on out-of-sample data $(x, y, g) \sim P_{X,Y|G}Q$.

To motivate why we chose to decouple the model learner's and the adversary's datasets, we derive a finite sample generalization result for Algorithm 1 using a similar reasoning as in [Zhou et al., 2022]. We observe that the model learner's proposed classifier f is independent of D^a given a prior Q. We can obtain the following generalization bound with standard uniform convergence analysis on D^a.

Theorem 3.4. *Let* $\bar{n} = \min_{g\in\mathcal{G}} |D_g^a|$*. Further assume the set of priors* Q_γ *contains* $|Q_\gamma|$ *discrete choices. Let* $\{f^1, \dots, f^T\}$ *be the output of Algorithm 1 and let* $P_F = U_{[\{f^1,\dots,f^T\}]}$ *denote the uniform probability over the classifiers. With probability at least* $1 - \delta$

$$\max_{Q\in\mathcal{Q}_\gamma} \mathop{\mathbb{E}}_{f\sim P_F} \mathbb{E}_{P_{X,Y|GQ_G}}[\epsilon(f(X), Y)] \leq \max_{Q\in\mathcal{Q}_\gamma} \mathop{\mathbb{E}}_{f\sim P_F} \sum_g Q_g \mathbb{E}_{D_g^a}[\epsilon(f(X), Y)] + \sqrt{\frac{T\log|Q_\gamma| + \log\frac{|\mathcal{G}|}{\delta}}{2\bar{n}}}. \quad (9)$$

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

$$\max_{Q\in\mathcal{Q}_\gamma} \mathbb{E}_{P_{X,Y|G}Q_G}[\epsilon(f^T(X), Y)] \leq \max_{Q\in\mathcal{Q}_\gamma} \sum_g Q_g \mathbb{E}_{D_g^a}[\epsilon(f^T(X), Y)] + \sqrt{\frac{\log\frac{|Q_\gamma||\mathcal{G}|}{\delta}}{2\bar{n}}}. \quad (10)$$

This shows the generalization properties of our classifier on the worst-case distribution on unseen data. These results depend on the complexity of the adversarial prior set $|\mathcal{Q}_\gamma|$ instead of the (potentially much larger) model complexity $|\mathcal{F}|$, as well as the number of samples in the least represented group in D^a.

The above theorem does not exploit the fact that the classifiers f^t generated by Algorithm 1 are only weakly dependent on D^a through Q^t, and Q^t is only dependent on D^a through a handful of statistics $\epsilon_g(f^{t'}), t' \in [t-1], g \in \mathcal{G}$. This is not the case if the adversary reuses D^{tr}. If we assume D^a to be fully independent of the

classifier sequence (e.g., assume a fresh D^a is drawn after each round), then we can state a much stronger generalization result.

Theorem 3.5. *Consider the online version of Algorithm 1 where D^a is re-sampled at each round and $\bar{n} = \min_{g\in\mathcal{G}} |D^a_g|$. Let $P_F = U_{[\{f^1,\ldots,f^T\}]}$ denote the uniform probability over the output classifiers. With probability at least $1-\delta$*

$$\max_{Q\in\mathcal{Q}_\gamma} \mathop{\mathbb{E}}_{f\sim P_F} \mathbb{E}_{P_{X,Y|GQ_G}}[\epsilon(f(X),Y)] \leq \max_{Q\in\mathcal{Q}_\gamma} \mathop{\mathbb{E}}_{f\sim P_F} \sum_g Q_g \mathbb{E}_{D^a_g}[\epsilon(f(X),Y)] + \sqrt{\frac{\log\frac{|\mathcal{G}|}{\delta}}{2\bar{n}}}. \tag{11}$$

In this scenario, the generalization error as seen from the adversary is independent of both the adversary's and model's complexities. Appendix B.2 shows empirically that the expected error computed from D^a tracks true out-of-sample error much better than training error, supporting this idea that D^a is 'nearly independent' of the learned classifier.

3.4 Addressing importance weights and losses in the 0 training error scenario.

Consider the standard surrogate loss relaxation of the classifier $f = \arg\min_{f\in\mathcal{F}} \mathbb{E}_{\mathcal{D}^{tr}}[\frac{Q_G}{P_G}\ell(f(X),Y)]$. As described in **P2**, the minimizer f may be insensitive to the adversarial prior for high capacity networks able to achieve 0 error on the training set $\mathcal{D}^{tr}$ [Sagawa et al., 2019, Byrd and Lipton, 2019, Xu et al., 2021, Nacson et al., 2019b].

Here we derive a modification of the surrogate loss function $\ell(f(x),y)$ that explicitly incorporates the group prior information, $L(f(x),y,q)$ as denoted in Figure 2. We then give conditions under which this loss function addresses the problems identified in **P2**. For theoretical reasons, we derive our results in the classification scenario for univariate loss functions of the form $\ell : \mathbb{R} \to \mathbb{R}^+$ based on the margin score u, and later discuss necessary conditions on this loss function ℓ. One motivating example of such a univariate loss function that will satisfy all our desiderata is crossentropy loss over a softmax nonlinearity, in which case we can write

$$\begin{aligned} CE(f(x),y) &= \log(1+e^{-u}), \\ u(f(x),y) &= f_y(x) - \log\textstyle\sum_{y'\neq y} e^{f_{y'}(x)}. \end{aligned} \tag{12}$$

We propose the following weight-dependent transformation to the sample margin

$$v(u,q) := \begin{cases} u & \text{if } u < 0, \\ u[1-s(u)(1-q^{-1})] & \text{if } u \geq 0, \end{cases} \tag{13}$$

where q is the weight placed by the adversarial prior on the sample ($q = Q(g)$), and $s(u) = 1-e^{-\tau u}$ for any value $\tau > 0$. Essentially, this modification smoothly transitions between the standard margin function $u(f(x),y)$ and a margin function $\frac{u(f(x),y)}{q}$. We denote the resulting loss function from the mapping $\ell(v(u(f(x),y),q))$ as $L(f(x),y,q)$ or $L(u,q)$ for short.

Proposition 3.8 shows that this tempered loss L has interesting properties when the base loss ℓ is exponential-tailed and smooth, and L is also monotonically increasing w.r.t. q for $u \geq 0$. We extend the work in [Soudry et al., 2018] and show that we can control the resulting classifier even in the 0-error regime, converging to weighted max-margin solutions that depend on the weighting parameter q. The supporting definitions and assumptions are presented next.

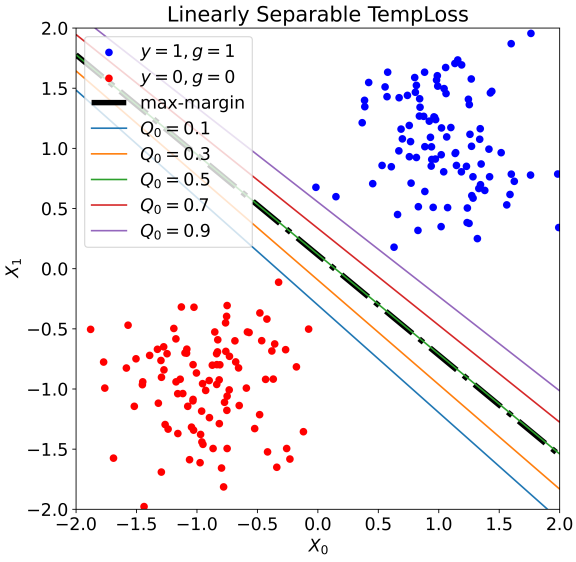


Figure 3: Decision boundary of an SGD-trained linear classifier using TempLoss on separable data, for different values of Q_0. Negative (blue, $y=0$) samples are given temperature scaling Q_0, while positive (red, $y=1$) samples use scaling $1-Q_0$. The decision boundary responds to the temperature parameter Q_0, unlike importance weighting on the base exponential loss.

Definition 3.6. [Soudry et al., 2018] A function $-\ell'(u)$ has a *tight exponential tail* if there exists positive constants c, a, μ_+, μ_-, u_+, and u_- such that

$$\begin{aligned} \forall u > u_+ : \quad & -\ell'(u) \leq c(1+e^{-\mu_+ u})e^{-au}, \\ \forall u > u_- : \quad & -\ell'(u) \geq c(1-e^{-\mu_- u})e^{-au}. \end{aligned} \tag{14}$$

Assumption 3.7. $l(u)$ is positive, differentiable, β-smooth, monotonically decreasing to zero, with $\lim_{u\to-\infty} -l'(u) \neq 0$ and $-l'(u)$ has a tight exponential tail.

Proposition 3.8. *For any loss function $\ell(u)$ satisfying Assumption 3.7 with smoothness parameter β, the tempered loss $L(u,q)$ also satisfies 3.7 with smoothness parameter β for any parameter $q \in (0,1]$. It additionally satisfies $\frac{dqL(u;q)}{dq} > 0 \, \forall q \in (0,1]$.*

The condition $\frac{dqL(u;q)}{dq} > 0$ ensures that an increase in the weight placed on a given sample or group increases the relative importance of the sample w.r.t. f. We show that this is sufficient to modify the solution to the inner loop objective in Eq.(5) for a simple linear classifier over a separable (potentially non-linear) representation function in Proposition 3.9. Figure 3 shows the results of using TempLoss instead of logistic regression on a linear SGD-trained model on synthetic data.

Proposition 3.9. *Given a binary classification dataset $\mathcal{D} = \{(x_i, y_i, g_i)\}_{i=1}^n \in \mathcal{X} \times \{-1,1\} \times \mathcal{G}$, a fixed representation $\phi : \mathcal{X} \to \mathbb{R}^d$ such that $\exists \theta \in \mathbb{R}^d : y_i \theta^T \phi(x_i) \geq 0$ $\forall i = 1, ..., n$ (data is linearly separable under ϕ), and a group distribution Q. Then, the limit of the gradient descent iterates of the linear classifier $f_\theta(x) = \theta^T \phi(x)$ for the inner loop objective in Eq.(5) with base loss $\ell(u)$ satisfying Assumption 3.7 is*

$$\begin{aligned} \lim_{t\to\infty} \frac{\theta^t}{||\theta^t||_2} &\to \frac{\hat{\theta}}{||\hat{\theta}||_2}, \\ \hat{\theta} &:= \arg\min_{\theta \in \mathbb{R}^d} ||\theta||_2^2 \\ &\quad s.t. \ \{y_i \theta^T \phi(x_i) \geq Q_{g_i}\}_{i=1}^n. \end{aligned} \tag{15}$$

This extends the results in [Soudry et al., 2018, Nacson et al., 2019b] to our tempered weighted loss. Similar results hold for multi-class classification. Here we show that SGD on the proposed loss leads to a weighted max-margin solution with sample margins proportional to the weighting coefficients Q_G.

Since our temperature-scaled loss modifies minimum classification margins per group on linearly separable data, we can establish perturbation robustness guarantees that depend on Q, as shown in Proposition 3.9.

Proposition 3.10. *In the setting of Proposition 3.9, where the representation function ϕ is $M-$Lipschitz, the solution is guaranteed to asymptotically satisfy*

$$\begin{aligned} &\min_{\bar{x}} y_i \theta^T \phi(\bar{x}) \geq 0 \qquad \forall (x_i, y_i, g_i) \in \mathcal{D}, \\ &s.t. \, ||\bar{x} - x_i||_2^2 \leq \frac{Q(g_i)^2}{||\hat{\theta}(Q)||_2^2} \frac{1}{M}. \end{aligned} \tag{16}$$

That is, the model is robust to ℓ_2 input perturbations at least up to $\frac{Q(g_i)^2}{||\hat{\theta}||_2^2} \frac{1}{M}$ on its training samples.

4 Related Work

GDRO [Sagawa et al., 2019] and similar approaches such as Minimax Group Fairness [Martinez et al., 2020, Diana et al., 2020] minimize worst-group loss. However, when applied to neural networks, these methods rely on strong ad-hoc regularization techniques to be responsive to sample re-weighting and learn a model that generalizes to unseen data [Sagawa et al., 2019, Byrd and Lipton, 2019]. This is in part due to the implicit bias of SGD for exponential tailed losses, where an interpolating family of models converges to the max margin classifier as discussed in [Nacson et al., 2019a, Xu et al., 2021].

Prior work has attempted to change the asymptotic behaviour of the learned classifier by modifying the cross-entropy loss. Several works have proposed to introduce correction terms to the logits of the classifier [Cao et al., 2019, Menon et al., 2020, Ye et al., 2020, Kini et al., 2021, Narasimhan and Menon, 2021, Lu et al., 2022]. In particular [Kini et al., 2021] proposed VS-loss, which integrates the mentioned approaches into a single loss function that applies additive and multiplicative correction terms to the logits. [Wang et al., 2021] proposed to apply importance weights to polynomial-tailed losses, which are asymptotically responsive to importance weights, but lack the interpretability of max-margin solutions. VS-loss [Kini et al., 2021] and the loss proposed by [Lu et al., 2022] are the most closely related to our TempLoss with the difference that we explicitly formalize the connection with the adversarial weights in the GDRO setting, and thus do not require to make any prior assumptions on which group will be most disadvantaged. Moreover, in comparison with VS-loss, we do not require any burn-in phase or separate (additive) corrections.

The work by [Zhou et al., 2022] proposed a Bi-level formulation for importance weight learning in the context of out of distribution learning. Their proposed framework directly maps the space of importance weights into the model parameter space and decouples the importance weight updates from the model optimization objective. Their work minimizes a risk metric on a validation set by learning the sample-level importance weights in the training sets for weighted ERM (and treating the resulting classifier's parameters as a dependent variable). As such, their method requires second order differentiation of the risk metric w.r.t. the model parameters, and each importance weight in the training set. Our Bi-level GDRO formulation fully decouples this dependency by instead learning group weights on the adversary's dataset. This, coupled with the minimax objective, enables the use of a convergent no-regret approach to learn the adversarial weights.

The work in [Cotter et al., 2019] analyzed a constrained optimization problem as a two-player game where each player has its own dataset. They showed

that this approach significantly improved the constrained violation on out-of-sample data. The work by [Sagawa et al., 2020] analyzes how over-parametrized models can hurt generalization on minority groups when spurious correlations and noisy features are present in the data, even when the average test error is improved. They provide a theoretical analysis based on a synthetic linear logistic regression that we adapt in the context of our approach in Appendix D. For an extended discussion on related work, see Appendix C, where we additionally motivate our choice of baselines for experimental comparisons.

5 Experiments and Results

We empirically evaluate the capabilities of our proposed Bi-Level approach with TempLoss, solved via Algorithm 1, at addressing distributional robustness on the over-parameterized classification setting. We compare our approach against empirical risk minimization (ERM), standard GDRO (with cross-entropy loss), VS-loss, and poly-tailed loss, as well as their Bi-level counterparts. We run our experiments on three image classification datasets, and 2 tabular classification datasets which we describe next.

5.1 Datasets and methods

We ran image classification experiments on a pretrained ResNet architecture [He et al., 2016], and tabular data classification over a fully connected network. In all scenarios, we used a small weight decay penalty (10^{-4}). We used Waterbirds [Sagawa et al., 2019] (2 targets, 4 groups), CIFAR-10 [Krizhevsky et al., 2009] (10 targets, 10 groups) and HAM-10K [Tschandl et al., 2018] (7 targets, 7 groups) as our image classification datasets; and UCI ADULT [Dua et al., 2017] (2 targets, 8 groups) and German Credit [Hofmann, 1994] (2 targets, 8 groups) as our tabular datasets. In all scenarios, the group variable G also includes the target variable Y. Further details on datasets, architecture, train-validation-test splitting, and training hyper-parameters for all methods are found on Appendix B.1. For all methods except ERM, we run a single epoch of training of the importance weighted function (i.e., $M(Q)$ in Algorithm 1) between adversarial prior updates, as it is standard in the literature [Diana et al., 2020, Sagawa et al., 2019].

5.2 Worst group error generalization

Figure 1 and Figure 4 in Appendix B.2 show the evolution of the worst group error on the test split across training epochs for ERM, GDRO, Bi-level GDRO (Algorithm 1 with learner minimizing standard cross entropy), and Bi-level Temploss GDRO. We observe that the Bi-level strategy considerably improves out-of-sample worst group error across all iterations. Further, using TempLoss instead of cross-entropy often yields meaningful improvements, and it is never worse than the standard cross-entropy objective.

5.3 Relaxed minimax fairness

Weighted errors as a function of γ[6] are shown in Table 1. We observe that the Bi-level variation of any base algorithm and loss combination (GDRO, VS-loss, Poly-tailed) significantly improves on the standard objective where the adversary's objective uses the same loss function and dataset as the learner. We compare results for various values of relaxed group fairness, where the user wants to achieve some γ-trade-off between worst case and average performance (with $\gamma = 0$ being the worst group loss/error objective and $\gamma = 1$ being ERM). This is achieved by setting the adversary's constraint to $\mathcal{Q}_\gamma = \{Q \in \mathbb{P}_\mathcal{G} : Q_g \geq \gamma P_g, \forall g \in \mathcal{G}\}$ in Algorithm 1. Moreover, the proposed Bi-level TempLoss GDRO method compares favorably with the Bi-level version of other approaches such as VS-loss and poly-loss; especially for unbalanced datasets.

6 Conclusions and Limitations

Here we discussed the use of a bi-level formulation of minimax fairness to improve generalization in the offline setting. We proposed a simple, provably convergent algorithm to solve this bi-level formulation and empirically showed it is capable of improving generalization for a variety of loss functions and datasets. We additionally discussed the lack of responsiveness to importance weighting exhibited by interpolating classifiers with exponential-tailed losses in the context of minimax robustness. We proposed a temperature-scaled exponential loss that provably converges to weighted-max-margin classifiers on separable training data, and discussed connections with robustness. The combination of our proposed TempLoss with our Bi-level formulation is able to effectively improve results for the relaxed minimax objective without requiring a-prior assumptions of the disadvantaged groups or additional ad-hoc regularization.

Acknowledgments

Work partially supported by NSF, ONR, NGA, and the Simons Foundation.

[6] $\gamma\times$ average error $+(1-\gamma)\times$ worst group error

Table 1: Weighted error results for each dataset and method as a function of γ. $\gamma = 0$ corresponds to worst group error objective, while $\gamma = 1$ corresponds to ERM. Deviations computed over 5 splits. The best results are consistently achieved for methods updating the importance weight with the held-out validation set as here proposed.

		Weighted error			
Dataset	Method	$\gamma = 0.0$	$\gamma = 0.1$	$\gamma = 0.2$	$\gamma = 0.5$
Waterbirds	ERM	33.8±0.3	30.7±0.3	27.6±0.2	18.2±0.2
	GDRO	33.2±0.5	30.1±0.7	27.2±0.7	18.7±0.2
	GDRO Polytail	36.3±0.3	33.5±0.4	30.6±0.1	21.2±0.1
	GDRO VS-loss	39.9±0.2	36.1±0.2	32.6±0.2	21.9±0.2
	Bi-level GDRO	16.5±0.8	24.9±0.3	23.6±0.4	17.1±0.2
	Bi-level Polytail	15.2±0.6	22.0±0.1	20.1±0.1	15.3±0.1
	Bi-level VS-loss	10.3±0.1	23.1±1.4	21.4±0.7	**13.8±0.2**
	Bi-level TempLoss	**10.1±0.3**	**18.8±1.3**	**17.5±0.4**	16.5±0.3
CIFAR10	ERM	29.0±0.8	27.6±0.7	26.2±0.7	21.9±0.5
	GDRO	25.5±0.2	24.2±0.3	23.1±0.9	20.1±0.3
	GDRO Polytail	26.2±0.2	25.3±0.5	24.4±0.4	20.9±0.3
	GDRO VS-loss	25.2±0.9	23.9±0.6	22.9±0.8	20.1±0.4
	Bi-level GDRO	**21.8±0.7**	21.1±0.9	20.8±1.0	19.3±0.9
	Bi-level Polytail	22.7±0.8	21.7±0.6	21.4±0.3	20.1±0.1
	Bi-level VS-loss	21.8±1.3	21.2±0.9	**21.0±1.0**	19.8±0.7
	Bi-level TempLoss	21.9±1.1	**21.±0.9**	21.2±1.2	**19.3±0.8**
UCI Adult	ERM	44.1±0.9	41.1±0.8	38.1±0.7	29.2±0.4
	GDRO	24.6±1.6	24.6±2.1	24.4±2.4	22.0±1.4
	GDRO Polytail	23.8±0.4	23.7±1.2	23.3±1.3	22.2±1.5
	GDRO VS-loss	39.4±2.2	41.1±2.6	43.5±0.7	38.3±1.3
	Bi-level GDRO	23.6±1.4	23.2±1.6	**22.6±1.0**	23.8±3.5
	Bi-level Polytail	25.6±0.4	25.4±0.2	24.1±0.6	23.9±3.2
	Bi-level VS-loss	24.8±1.7	24.6±2.1	25.0±2.4	30.9±4.7
	Bi-level TempLoss	**22.8±0.6**	**23.2±1.0**	23.4±0.8	**23.1±2.1**
HAM10K	ERM	16.1±7.4	14.7±6.7	13.3±6.0	9.2±3.8
	GDRO	9.0±0.6	8.7±0.5	8.7±1.0	7.3±0.3
	GDRO Polytail	8.7±0.4	7.4±0.7	7.3±0.8	6.9±0.7
	GDRO VS-loss	23.0±1.6	22.8±0.3	21.2±0.1	13.1±0.1
	Bi-level GDRO	7.3±0.2	**6.8±0.1**	6.5±0.1	5.4±0.2
	Bi-level Polytail	8.0±0.1	6.8±0.1	6.5±0.1	**5.3±0.1**
	Bi-level VS-loss	17.0±5.2	11.8±0.4	15.5±2.7	7.2±0.9
	Bi-level TempLoss	**7.1±0.5**	6.8±0.1	**5.6±0.8**	5.4±0.2
German	ERM	100.0±0.0	92.8±0.1	85.5±0.2	63.8±0.5
	GDRO	83.3±23.6	78.2±21.2	66.5±19.9	54.3±12.6
	GDRO Polytail	83.3±23.6	78.3±21.0	73.2±18.3	57.6±10.6
	GDRO VS-loss	100.0±0.0	92.8±0.2	85.7±0.4	64.3±0.3
	Bi-level GDRO	68.3±22.5	64.6±20.4	61.0±18.3	49.0±13.3
	Bi-level Polytail	68.3±22.5	64.9±20.3	60.0±19.2	49.2±13.0
	Bi-level VS-loss	**61.0±20.0**	**59.6±15.9**	59.1±15.0	57.4±8.8
	Bi-level TempLoss	63.8±21.0	60.8±18.9	**57.8±16.9**	**47.9±11.40**

References

[Becker and Kohavi, 1996] Becker, B. and Kohavi, R. (1996). Adult. UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C5XW20.

[Blondel et al., 2022] Blondel, M., Berthet, Q., Cuturi, M., Frostig, R., Hoyer, S., Llinares-López, F., Pedregosa, F., and Vert, J.-P. (2022). Efficient and modular implicit differentiation. *Advances in neural information processing systems*, 35:5230–5242.

[Buet-Golfouse, 2021] Buet-Golfouse, F. (2021). Narrow margins: Classification, margins and fat tails. In *International Conference on Machine Learning*, pages 1127–1135. PMLR.

[Byrd and Lipton, 2019] Byrd, J. and Lipton, Z. (2019). What is the effect of importance weighting in deep learning? In *International Conference on Machine Learning*, pages 872–881. PMLR.

[Cao et al., 2019] Cao, K., Wei, C., Gaidon, A., Arechiga, N., and Ma, T. (2019). Learning imbalanced datasets with label-distribution-aware margin loss. *Advances in Neural Information Processing Systems*, 32.

[Chen et al., 2017] Chen, R., Lucier, B., Singer, Y., and Syrgkanis, V. (2017). Robust optimization for non-convex objectives. *arXiv preprint arXiv:1707.01047*.

[Cotter et al., 2019] Cotter, A., Gupta, M., Jiang, H., Srebro, N., Sridharan, K., Wang, S., Woodworth, B., and You, S. (2019). Training well-generalizing classifiers for fairness metrics and other data-dependent constraints. In *International Conference on Machine Learning*, pages 1397–1405. PMLR.

[Diana et al., 2020] Diana, E., Gill, W., Kearns, M., Kenthapadi, K., and Roth, A. (2020). Convergent algorithms for (relaxed) minimax fairness. *arXiv e-prints*, pages arXiv–2011.

[Dua et al., 2017] Dua, D., Graff, C., et al. (2017). Uci machine learning repository.

[Hashimoto et al., 2018] Hashimoto, T., Srivastava, M., Namkoong, H., and Liang, P. (2018). Fairness without demographics in repeated loss minimization. In *International Conference on Machine Learning*, pages 1929–1938. PMLR.

[He et al., 2016] He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778.

[Hofmann, 1994] Hofmann, H. (1994). Statlog (German Credit Data). UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C5NC77.

[Kini et al., 2021] Kini, G. R., Paraskevas, O., Oymak, S., and Thrampoulidis, C. (2021). Label-imbalanced and group-sensitive classification under overparameterization. *Advances in Neural Information Processing Systems*, 34.

[Krizhevsky et al., 2009] Krizhevsky, A. et al. (2009). Learning multiple layers of features from tiny images.

[Liu et al., 2021] Liu, E. Z., Haghgoo, B., Chen, A. S., Raghunathan, A., Koh, P. W., Sagawa, S., Liang, P., and Finn, C. (2021). Just train twice: Improving group robustness without training group information. In *International Conference on Machine Learning*, pages 6781–6792. PMLR.

[Lu et al., 2022] Lu, Y., Ji, W., Izzo, Z., and Ying, L. (2022). Importance tempering: Group robustness for overparameterized models. *arXiv preprint arXiv:2209.08745*.

[Martinez et al., 2020] Martinez, N., Bertran, M., and Sapiro, G. (2020). Minimax pareto fairness: A multi objective perspective. In *International Conference on Machine Learning*, pages 6755–6764. PMLR.

[Menon et al., 2020] Menon, A. K., Jayasumana, S., Rawat, A. S., Jain, H., Veit, A., and Kumar, S. (2020). Long-tail learning via logit adjustment. *arXiv preprint arXiv:2007.07314*.

[Nacson et al., 2019a] Nacson, M. S., Lee, J., Gunasekar, S., Savarese, P. H. P., Srebro, N., and Soudry, D. (2019a). Convergence of gradient descent on separable data. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 3420–3428. PMLR.

[Nacson et al., 2019b] Nacson, M. S., Srebro, N., and Soudry, D. (2019b). Stochastic gradient descent on separable data: Exact convergence with a fixed learning rate. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 3051–3059. PMLR.

[Narasimhan and Menon, 2021] Narasimhan, H. and Menon, A. K. (2021). Training over-parameterized models with non-decomposable objectives. *Advances in Neural Information Processing Systems*, 34:18165–18181.

[Roughgarden, 2016] Roughgarden, T. (2016). *Twenty Lectures on Algorithmic Game Theory*. Cambridge University Press.

[Sagawa et al., 2019] Sagawa, S., Koh, P. W., Hashimoto, T. B., and Liang, P. (2019). Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. *arXiv preprint arXiv:1911.08731*.

[Sagawa et al., 2020] Sagawa, S., Raghunathan, A., Koh, P. W., and Liang, P. (2020). An investigation of why overparameterization exacerbates spurious correlations. In *International Conference on Machine Learning*, pages 8346–8356. PMLR.

[Soudry et al., 2018] Soudry, D., Hoffer, E., Nacson, M. S., Gunasekar, S., and Srebro, N. (2018). The implicit bias of gradient descent on separable data. *The Journal of Machine Learning Research*, 19(1):2822–2878.

[Tschandl et al., 2018] Tschandl, P., Rosendahl, C., and Kittler, H. (2018). The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific Data*, 5(1):1–9.

[Wah et al., 2011] Wah, C., Branson, S., Welinder, P., Perona, P., and Belongie, S. (2011). The caltech-ucsd birds-200-2011 dataset.

[Wainwright, 2019] Wainwright, M. J. (2019). *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge university press.

[Wang et al., 2021] Wang, K. A., Chatterji, N. S., Haque, S., and Hashimoto, T. (2021). Is importance weighting incompatible with interpolating classifiers? *arXiv preprint arXiv:2112.12986*.

[Xu et al., 2021] Xu, D., Ye, Y., and Ruan, C. (2021). Understanding the role of importance weighting for deep learning. *arXiv preprint arXiv:2103.15209*.

[Yang et al., 2023] Yang, Z., Ko, Y. L., Varshney, K. R., and Ying, Y. (2023). Minimax auc fairness: Efficient algorithm with provable convergence. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 11909–11917.

[Ye et al., 2020] Ye, H.-J., Chen, H.-Y., Zhan, D.-C., and Chao, W.-L. (2020). Identifying and compensating for feature deviation in imbalanced deep learning. *arXiv preprint arXiv:2001.01385*.

[Zhou et al., 2022] Zhou, X., Lin, Y., Pi, R., Zhang, W., Xu, R., Cui, P., and Zhang, T. (2022). Model agnostic sample reweighting for out-of-distribution learning. In *International Conference on Machine Learning*, pages 27203–27221. PMLR.

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. No.

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. Not Applicable
   (d) Information about consent from data providers/curators. Not Applicable
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. Not Applicable

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

A Proofs

A.1 Proof of Theorem 3.3

Theorem 3.3 Algorithm 1 with $\eta = \frac{\sqrt{2\frac{\log|\mathcal{G}|}{T}}}{(1-||\varepsilon||_1^1)}$ computes a uniform distribution over solutions $P_F = U_{[\{f_1,\ldots,f_T\}]}$ such that

$$\max_{Q\in\mathcal{Q}_\gamma} \mathbb{E}_{f\sim P_F} \sum_{g\in\mathcal{G}} Q_g \epsilon_g(f) \qquad \leq \alpha \min_{f\in\mathcal{F}} \max_{Q\in\mathcal{Q}_\gamma} \sum_{g\in\mathcal{G}} Q_g\epsilon_g(f) + \sqrt{\frac{2\log|\mathcal{G}|}{T}} \tag{17}$$

Further, for any $\eta > 0$, the solution to Algorithm 1 satisfies

$$\max_{Q\in\mathcal{Q}_\gamma} \mathbb{E}_{f\sim P_F} \sum_{g\in\mathcal{G}} Q_g \epsilon_g(f) \qquad \leq \alpha(1+\eta) \min_{f\in\mathcal{F}} \max_{Q\in\mathcal{Q}_\gamma} \sum_{g\in\mathcal{G}} Q_g\epsilon_g(f) + \frac{\log|\mathcal{G}|}{\eta T} \tag{18}$$

Proof. We first observe we can rewrite the outer level of the bi-level objective in Eq.(5) as

$$\max_{Q\in\mathcal{Q}_\gamma} \sum_{g\in\mathcal{G}} Q_g\epsilon_g(f) \quad = \max_{\hat{Q}\in\Delta^{G-1}} \sum_{g\in\mathcal{G}} \hat{Q}_g(1-|\gamma|_1^1)\epsilon_g(f) + \sum_{g'\in\mathcal{G}} \gamma_{g'}\epsilon_{g'}(f). \tag{19}$$

And that we can map any distribution $\hat{Q}\in\Delta^{G-1}$ to its corresponding distribution $Q = \hat{Q}(1-|\gamma|_1^1)+\gamma, Q\in\mathcal{Q}_\gamma$. We similarly note that the $\alpha-$approximate solver $M(Q)$ for the objective $\sum_{g\in\mathcal{G}} Q_g\epsilon_g(f)$ is also an $\alpha-$approximate solver $\hat{M}(\hat{Q})$ for the objective $\sum_{g\in\mathcal{G}} \hat{Q}_g(1-|\gamma|_1^1)\epsilon_g(f) + \sum_{g'\in\mathcal{G}} \gamma_{g'}\epsilon_{g'}(f)$

With this, it suffices to use Theorem 1 in [Chen et al., 2017] to see that the following update is convergent

$$\begin{aligned}
\hat{Q}^t &\propto \exp\{\eta'(1-|\gamma|_1^1)\textstyle\sum_{t'\in[t]}\epsilon_g(f^t) + \eta'\sum_{t'\in[t]}\sum_{g'\in\mathcal{G}}\gamma_{g'}\epsilon_{g'}(f^t)\}_{g\in\mathcal{G}},\\
&\propto \exp\{\eta'(1-|\gamma|_1^1)\textstyle\sum_{t'\in[t]}\epsilon_g(f^t)\}_{g\in\mathcal{G}},\\
f_t &\leftarrow \hat{M}(\hat{Q}^t),\\
&\leftarrow M((1-|\gamma|_1^1)\hat{Q}^t+\gamma).
\end{aligned} \tag{20}$$

Where we observe that the bias term $\eta'\sum_{g'\in\mathcal{G}}\gamma_{g'}\epsilon_{g'}(f^t)$ is independent of g and can thus be factored out of the update, and we identify the term $\eta = \eta'(1-|\gamma|_1^1)$ for our stated result.

Theorem: 3.4 Let $\bar{n} = \min_{g\in\mathcal{G}}|D_g^a|$. Further assume the set of priors Q_γ contains $|Q_\gamma|$ discrete choices. Let $\{f^1,\ldots,f^T\}$ be the output of Algorithm 1 and let $P_F = U_{[\{f^1,\ldots,f^T\}]}$ denote the uniform probability over the classifiers. With probability at least $1-\delta$

$$\max_{Q\in\mathcal{Q}_\gamma} \mathop{\mathbb{E}}_{f\sim P_F} \mathbb{E}_{P_{X,Y|GQ_G}}[\epsilon(f(X),Y)] \leq \quad \max_{Q\in\mathcal{Q}_\gamma} \mathop{\mathbb{E}}_{f\sim P_F} \textstyle\sum_g Q_g \mathbb{E}_{D_g^a}[\epsilon(f(X),Y)] + \sqrt{\frac{T\log|Q_\gamma| + \log\frac{|\mathcal{G}|}{\delta})}{2\bar{n}}}. \tag{21}$$

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

$$\max_{Q\in\mathcal{Q}_\gamma} \mathbb{E}_{P_{X,Y|G}Q_G}[\epsilon(f^T(X),Y)] \leq \quad \max_{Q\in\mathcal{Q}_\gamma} \textstyle\sum_g Q_g \mathbb{E}_{D_g^a}[\epsilon(f^T(X),Y)] + \sqrt{\frac{\log\frac{|Q_\gamma||\mathcal{G}|}{\delta}}{2\bar{n}}}. \tag{22}$$

Proof. We observe that all intermediary outputs of Algorithm 1 $f^t = M(Q^t)$ are implicit functions in $Q\in\mathcal{Q}\gamma$, as such, there are $|\mathcal{Q}_\gamma|$ such functions, and $|\mathcal{Q}_\gamma|$ choose T with repetition ways to get the distribution P_F, that is, there are $\bar{C}_{|\mathcal{Q}_\gamma|,T} \leq |\mathcal{Q}_\gamma|^T$ such distributions in P_F.

From this observation, we can use the Hoeffding and union bound (Corollary 2.2 in [Wainwright, 2019]) to state that, with probability at least $1-\delta/G$

$$\begin{aligned}
\mathbb{E}_{P_{X,Y|G=g}} \mathop{\mathbb{E}}_{f\sim P_F}[\epsilon(f(X),Y)] &\leq \mathbb{E}_{D_g^a} \mathop{\mathbb{E}}_{f\sim P_F}[\epsilon(f(X),Y)] + \sqrt{\frac{T\log|Q_\gamma| + \log\frac{G}{\delta}}{2|D_g^a|}},\\
&\leq \mathbb{E}_{D_g^a} \mathop{\mathbb{E}}_{f\sim P_F}[\epsilon(f(X),Y)] + \sqrt{\frac{T\log|Q_\gamma| + \log\frac{G}{\delta}}{2\bar{n}}}.
\end{aligned} \tag{23}$$

Where the second inequality follows from the definition of $\bar{n} = \min_{g \in G} |D_g^a|$. Taking the union bound across all G groups, we observe that the following bound holds with probability at least $1 - \delta$

$$\max_{Q \in \mathcal{Q}_\gamma} \mathop{\mathbb{E}}_{f \sim P_F} \mathbb{E}_{P_{X,Y|G} Q_G}[\epsilon(f(X), Y)] \quad \leq \max_{Q \in \mathcal{Q}_\gamma} \mathop{\mathbb{E}}_{f \sim P_F} \textstyle\sum_g \hat{Q}_g \mathbb{E}_{D_g^a}[\epsilon(f(X), Y)] + \sqrt{\frac{T \log |Q_\gamma| + \log \frac{G}{\delta}}{2\bar{n}}}. \tag{24}$$

The bound for just the final classifier f^T is derived identically by supplanting $T \log |Q_\gamma|$ by $\log |Q_\gamma|$. For it, we can state that with probability at least $1 - \delta$

$$\max_{Q \in \mathcal{Q}_\gamma} \mathbb{E}_{P_{X,Y|G} Q_G}[\epsilon(f^T(X), Y)] \quad \leq \max_{Q \in \mathcal{Q}_\gamma} \textstyle\sum_g \hat{Q}_g \mathbb{E}_{D_g^a}[\epsilon(f^T(X), Y)] + \sqrt{\frac{\log |Q_\gamma| + \log \frac{G}{\delta}}{2\bar{n}}}. \tag{25}$$

Theorem 3.5 Consider the online version of Algorithm 1 where D^a is re-sampled at each round and $\bar{n} = \min_{g \in \mathcal{G}} |D_g^a|$. Let $P_F = U_{[\{f^1, \ldots, f^T\}]}$ denote the uniform probability over the output classifiers. With probability at least $1 - \delta$

$$\begin{aligned} &\max_{Q \in \mathcal{Q}_\gamma} \mathop{\mathbb{E}}_{f \sim P_F} \mathbb{E}_{P_{X,Y|G Q_G}}[\epsilon(f(X), Y)] \leq \\ &\max_{Q \in \mathcal{Q}_\gamma} \mathop{\mathbb{E}}_{f \sim P_F} \textstyle\sum_g Q_g \mathbb{E}_{D_g^a}[\epsilon(f(X), Y)] + \sqrt{\frac{\log \frac{|\mathcal{G}|}{\delta}}{2\bar{n}}}. \end{aligned} \tag{26}$$

Proof. In this scenario the samples of D_g^a on round $T + 1$ are independent of the resulting distribution $P_F = U_{[\{f^1, \ldots, f^T\}]}$. Therefore, the error samples $\epsilon^i = \mathop{\mathbb{E}}_{f \sim P_F} \epsilon(f(x_i, y_i)); x_i, y_i \in D_g^a$ are also i.i.d.. Using the bounded differences inequality, we get, with probability at least $1 - \delta/G$

$$\max_{Q \in \mathcal{Q}_\gamma} \mathop{\mathbb{E}}_{f \sim P_F} \mathbb{E}_{P_{X,Y|G Q_G}}[\epsilon(f(X), Y)] \leq \quad \max_{Q \in \mathcal{Q}_\gamma} \mathop{\mathbb{E}}_{f \sim P_F} \textstyle\sum_g Q_g \mathbb{E}_{D_g^a}[\epsilon(f(X), Y)] + \sqrt{\frac{\log \frac{|\mathcal{G}|}{\delta}}{2|D_g^a|}}. \tag{27}$$

Where the main difference w.r.t. the previous result is that we no longer need to apply a union bound over all possible distributions P_F since we can guarantee independence of the error samples themselves. The rest of the proof proceeds identically to the preceding proof.

Proposition 3.8 For any loss function $\ell(u)$ satisfying Assumption 3.7 with smoothness parameter β, the tempered loss $L(u, q)$ also satisfies 3.7 with smoothness parameter β for any parameter $q \in (0, 1]$. It additionally satisfies $\frac{dqL(u;q)}{dq} > 0 \, \forall q \in (0, 1]$.

Proof. We first verify that $L(u, q)$ satisfies Assumption 3.7. We observe that $L(u, q) > 0 \forall u$; this is trivial for $u < 0$, since $\ell(u)$ itself satisfies Assumption 3.7, for $u \geq 0$, we additionally observe

$$\begin{aligned} L(u, q) &= \ell(u[1 - s(u)(1 - q^{-1})]) \\ &= \ell(u[\tfrac{1}{q} + \tfrac{q-1}{q} e^{-\tau u}]), \\ \tfrac{1}{q} + \tfrac{q-1}{q} e^{-\tau u} &\geq 0 \; \forall q > 0. \end{aligned} \tag{28}$$

We now check the derivatives w.r.t. margin u:

$$\frac{dL(u, q)}{du} = \begin{cases} \ell'(u), \; u < 0, \\ \ell'(u[\frac{1}{q} + \frac{q-1}{q} e^{-\tau u}]) \frac{1 + (q-1) e^{-\tau u}(1 - \tau u)}{q}, \; u \geq 0, \end{cases} . \tag{29}$$

We now verify that $\frac{dL(u,q)}{du} < 0$. For $u < 0$ this is immediate; for $u \geq 0$ we additionally observe that, for $q \in (0, 1]$

$$\begin{aligned} 1 + (1 - q) e^{-\tau u}(\tau u - 1) &= \underbrace{\tau u (1 - q) e^{-\tau u}}_{>0} + \underbrace{1 - e^{-\tau u}}_{>0} + \underbrace{e^{-\tau u} q}_{>0} \\ &> 0, \end{aligned} \tag{30}$$

since $\tau > 0$ and $u > 0$.

To check for β-smoothness, we immediately observe that $L(u, q)$ is continuous and differentiable for $u \neq 0$; for the special case $u = 0$, we have

$$\begin{array}{rcll} \lim_{u\to\gamma^-} L(u,q) & = & \lim_{u\to\gamma^+} L(u,q) & = \ell(\gamma), \\ \lim_{u\to\gamma^-} \frac{dL(u,q)}{du} & = & \lim_{u\to\gamma^+} \frac{dL(u,q)}{du} & = \ell'(\gamma). \end{array} \tag{31}$$

To compute the Lipschitz constant of $L'(u, q)$, we observe that $|L'(u, q)| \leq \beta, \forall u < 0$, and likewise $|L'(u, q)| \leq \beta \max_{u>0} \frac{|1+e^{-\tau u}(q-1)(1-\tau u)|}{q} \forall u < \gamma = \beta$, thus $L(u, q)$ is β-smooth.

The limits in Assumption 3.7 are easily verifiable from the base properties of $\ell(u)$:

$$\begin{array}{rcll} \lim_{u\to\infty} L(u,q) & = & \lim_{u\to\infty} \ell(\frac{u}{q}) & = 0, \\ \lim_{u\to\infty} L'(u,q) & = & \lim_{u\to\infty} \frac{1}{q}\ell'(\frac{u}{q}) & = 0, \\ \lim_{u\to-\infty} L(u,q) & = & \lim_{u\to\infty} \ell(u) & \neq 0. \end{array} \tag{32}$$

The last property of Assumption 3.7 we need to verify is if $-L'(u, q)$ has a tight exponential tail. This is immediately verifiable since $-\ell'(u)$ has a tight exponential tail, and $\lim_{u\to\infty} L'(u, q) = \frac{1}{q}\ell'(\frac{u}{q})$.

Lastly, we verify that $\frac{dL(u,q)}{dq} > 0 \, \forall q \in (0, 1], u \geq 0$, that is

$$\frac{dL(u,q)}{dq} = \begin{cases} 0, \ u < 0, \\ \ell'(\frac{u}{q}[(q-1)e^{-\tau u} + 1])[\frac{u}{q^2}(e^{-\tau u} - 1)], \ u \geq 0. \end{cases} \tag{33}$$

To complete the proof, we observe that

$$\begin{array}{r} \ell'(\frac{u-\gamma}{q}[(q-1)e^{-\tau u} + 1] + \gamma) < 0, \\ \frac{u}{q^2}(e^{-\tau u} - 1) < 0, \end{array} \tag{34}$$

which proves $\frac{dL(u,q)}{dq} > 0$ as required.

Proposition 3.9 Given a binary classification dataset $\mathcal{D} = \{(x_i, y_i, g_i)\}_{i=1}^n \in \mathcal{X} \times \{-1, 1\} \times \mathcal{G}$, a fixed representation $\phi : \mathcal{X} \to \mathbb{R}^d$ such that $\exists \theta \in \mathbb{R}^d : y_i \theta^T \phi(x_i) \geq 0 \ \forall i = 1, ..., n$ (data is linearly separable under ϕ), and a group distribution Q. Then, the limit of the gradient descent iterates of the linear classifier $f_\theta(x) = \theta^T \phi(x)$ for the inner loop objective in Eq.(5) with base loss $\ell(u)$ satisfying Assumption 3.7 is

$$\begin{array}{rl} \lim_{t\to\infty} \frac{\theta^t}{||\theta^t||_2} & \to \frac{\hat{\theta}}{||\hat{\theta}||_2}, \\ \hat{\theta} & := \arg\min_{\theta\in\mathbb{R}^d} ||\theta||_2^2 \\ & \quad \text{s.t. } \{y_i \theta^T \phi(x_i) \geq Q_{g_i}\}_{i=1}^n. \end{array} \tag{35}$$

Proof. Let $x_i' = \frac{\phi(x_i)}{Q_{g_i}}$ be the scaled, embedded i-th sample; we can build an equivalent dataset $D' = \{(\frac{\phi(x_i)}{Q_{g_i}}, y_i)\}_{i=1}^n = \{x_i', y_i)\}_{i=1}^n$. The dataset D' is also linearly separable since $Q_g \in (0, 1] \forall g \in \mathcal{G}$. Consider the dataset loss

$$\mathcal{L}(\theta, Q) = \sum_{i=1}^N \frac{Q_{g_i}}{p_{g_i}} L(y_i \theta^t \phi(x_i); Q_{g_i}). \tag{36}$$

Following Proposition 3.9, we can use Lemma 1 in [Soudry et al., 2018] to state that the gradient descent iterates

of $L(\theta, Q)$, $\theta(t)$ satisfy

$$\begin{aligned} \lim_{t\to\infty} \mathcal{L}(\theta(t), q) &= 0, \\ \lim_{t\to\infty} ||\theta(t)||_2^2 &= \infty, \\ \lim_{t\to\infty} y_i \theta^t(t)\phi(x_i) &= \infty, \\ \lim_{t\to\infty} y_i \theta^t(t) \frac{\phi(x_i)}{Q_{g_i}} &= \infty. \end{aligned} \tag{37}$$

From the preceding observations, it suffices to analyze the behaviour of $L(u, q)$ in the $u >> 0$ regime without loss of generality ($L(u, q) \simeq \ell(\frac{u}{q})$). In this regime, the composite loss $L(u, q)$ satisfies assumptions 1 through 4 in [Soudry et al., 2018] and we can thus follow the same reasoning as in Theorem 2 there (where constant multiplicative factors outside of the loss function like Q_{g_i}/p_{g_i} that are independent of the score u can be safely ignored) to state

$$\theta^t = \hat{\theta} \log(\frac{\eta}{B}\frac{t}{K}) + r(t) + \tilde{\theta}, \tag{38}$$

where B, K are the batch size and number of batches in dataset D' respectively, $\eta < 2 \max_g \frac{Q_g}{p_g} \beta^{-1} \sigma_{\max}^{-2}(\phi(X))$ is the SGD learning rate, and $\hat{\theta}$ is the max-margin solution

$$\begin{aligned} \hat{\theta} \quad &: \arg\min ||\theta||_2^2, \\ &\quad s.t. \quad y_i \theta^T x_i' \geq 1, \, \forall i \in [n] \\ &= \arg\min ||\theta||_2^2, \\ &\quad s.t. \quad y_i \theta^T \phi(x_i) \geq Q_{g_i}, \, \forall i \in [n]. \end{aligned} \tag{39}$$

Additionally,

$$\tilde{\theta} \quad : \eta e^{-\tilde{\theta}^T x_n'} = \alpha_n \frac{p_{g_n}}{q_{g_n}} \forall n \in \mathcal{S}, \tag{40}$$

with $\mathcal{S}$ the support set of (x_i', y_i) and α_n its support coefficients. The residual vector $r(t)$ has a bounded norm, which implies that

$$\lim_{t\to\infty} \frac{\theta^t}{||\theta^t||} \to \hat{\theta}. \tag{41}$$

Proposition 3.10 In the setting of Proposition 3.9, where the representation function ϕ is $M-$Lipschitz, the solution is guaranteed to asymptotically satisfy

$$\begin{aligned} &\min_{\bar{x}} y_i \theta^T \phi(\bar{x}) \geq 0 \qquad\qquad \forall (x_i, y_i, g_i) \in \mathcal{D}, \\ &s.t. \, ||\bar{x} - x_i||_2^2 \leq \frac{Q(g_i)^2}{||\hat{\theta}(Q)||_2^2} \frac{1}{M}. \end{aligned} \tag{42}$$

That is, the model is robust to ℓ_2 input perturbations at least up to $\frac{Q(g_i)^2}{||\hat{\theta}||_2^2}\frac{1}{M}$ on its training samples.

Proof. First we examine the following simplified problem

$$\begin{aligned} &\min_{\bar{\phi}} y_i \theta^T \bar{\phi}, \\ &s.t. \, ||\bar{\phi} - \phi_i||_2^2 \leq C^2, \end{aligned} \tag{43}$$

with the corresponding Lagrangian

$$\min_{\bar{\phi}} \max_{\xi \geq 0} y_i \theta^T \bar{\phi} + \tfrac{\xi}{2}(||\bar{\phi} - \phi_i||_2^2 - C^2), \tag{44}$$

By setting the derivative w.r.t. $\bar{\phi}$ of the Lagrangian to 0, we obtain

$$\begin{aligned} \tfrac{\partial \mathcal{L}}{\partial \xi} &= \theta + \xi(\bar{\phi} - \phi_i), \\ &= 0, \\ \bar{\phi} &= \phi_i - y_i \tfrac{\theta}{\xi}. \end{aligned} \tag{45}$$

The constraint needs to be active for the module of $\bar{\phi}$ to be finite, therefore, we solve for the constraint $||\bar{\phi} - \phi_i||_2^2 = C^2$ and recover

$$\begin{aligned} ||\bar{\phi} - \phi_i||_2^2 &= C^2 \\ \frac{||\theta||_2^2}{\xi^2} &= C^2, \\ \bar{\phi} &= \phi_i - y_i \frac{C}{||\theta||_2}\theta. \end{aligned} \tag{46}$$

Plugging this back into the value of $y_i\theta^T\bar{\phi}$ we recover

$$\begin{aligned} y_i\theta^T\bar{\phi} &= y_i\theta^T\phi_i - C||\theta||_2, \\ \lim_{t\to\infty} \frac{y_i\theta_t^T\bar{\phi}}{||\theta_t||_2} &= \lim_{t\to\infty} y_i \frac{\theta_t^T}{||\theta||_2}\phi_i - C \\ &= y_i \frac{\hat{\theta}^T}{||\hat{\theta}||_2}\phi_i - C \\ &\geq \frac{Q_{g_i}}{||\hat{\theta}||_2} - C \\ &\geq 0 \qquad\qquad\qquad\qquad\qquad if\, C \leq \frac{Q_{g_i}}{||\hat{\theta}||_2}. \end{aligned} \tag{47}$$

Here we added the time dependency on parameter θ_t and additionally used Proposition 3.9 to state that $\frac{\theta}{||\theta||_2} \xrightarrow[t\to\infty]{} \frac{\hat{\theta}}{||\hat{\theta}||_2}$. To finalize the proof, we observe that if $||x - x_i||_2^2 \leq \frac{Q_{g_i}^2}{||\hat{\theta}||_2^2}\frac{1}{M}$, then $||\phi(x) - \phi(x_i)||_2^2 \leq \frac{Q_{g_i}^2}{||\hat{\theta}||_2^2}$, and therefore $y_i\theta^T\phi(x) \geq 0$.

B Extended Experiments and Results

B.1 Experimental Details

B.1.1 Datasets

We used the following image classification datasets

Waterbirds [Sagawa et al., 2019] Based on the CUB dataset [Wah et al., 2011], this dataset contains photographs of birds, classified as *waterbirds* if they belong to either the seabird or waterfowl category, otherwise, the birds are labeled as *landbirds*, this is used as the target label of the classifier. The background images have been artificially replaced, and all images have a randomized background belonging to water backgrounds or land background, with 95% of waterbirds placed on water backgrounds, and 95% of landbirds placed on land backgrounds. The combination of these background categories and the target attribute constitute our 4 demographic groups.

CIFAR-10. [Krizhevsky et al., 2009] Based on the TinyImages dataset [Wah et al., 2011], this dataset contains 10 different classes, which are used as both labels and groups.

HAM-10K. [Tschandl et al., 2018] Contains more than 10k dermatoscopic images from 7 classes of skin lesions; with lesion frequencies varying between 67% and 1.1%. This is a highly unbalanced classification problem and we use the type of skin lesion as the target variable and group.

For tabular data classification we used

UCI ADULT [Dua et al., 2017] is a tabular dataset with census information of more than 26k individuals. The goal is to predict if an individual has low or high income (above 50K). We consider 8 groups corresponding to the product of (binarized) ethnicity, gender and income (target variable).

German Credit [Hofmann, 1994] is a tabular dataset containing information of 1k individuals that requested a bank credit. The goal is to predict if an individual has good or bad credit risk. We consider 8 groups corresponding to the product of personal status (married, non-married), binary gender and credit risk (binary target variable).

B.1.2 Methods

All image classification experiments are conducted over a ImageNet pretrained ResNet32 architecture [He et al., 2016] as in [Sagawa et al., 2019]. In all cases we do not rely on hyper-parameter tuning to select regularization strength and instead use a standard 10^{-4} ℓ_2 weight decay penalty. We used a sgd optimizer with momentum 0.9 and learning rate 10^{-2} or 10^{-3} (we chose the one with best performance in validation). We

implemented a linear warmup scheduler for the first 5 epochs starting at 0.1 of the base learning rate followed by a step decay of 0.1 every 1/3 of the total number of epochs. We used random crop and horizontal flip for image augmentations.

For the tabular datasets we used a two hidden layer MLP of 512x512. In all cases categorical variables were converted to one-hot encoding. We used same training setting and scheduler as in the image classification tasks except the learning rate which was set to 10^{-4}. Details are summarized in Table 2.

Table 2: Experimental hyperparameters

Dataset	Input Size	Classifier	Epochs/Batch	Optimizer (sgd + momentum=0.9)	Scheduler
Waterbirds	128x128x3	Resnet32 pretrained	300/128	lr=1e-2, ℓ_2-weight decay=1e-4	Linear warmup + Step decay
HAM10K	128x128x3	Resnet32 pretrained	300/128	lr=1e-3, ℓ_2-weight decay=1e-4	Linear warmup + Step decay
CIFAR10	32x32x3	Resnet32 pretrained	300/128	lr=1e-3, ℓ_2-weight decay=1e-4	Linear warmup + Step decay
UCI Adult	84	MLP 512x512	500/128	lr=1e-4, ℓ_2-weight decay=1e-4	Linear warmup + Step decay
German Credit	37	MLP 512x512	500/128	lr=1e-4, ℓ_2-weight decay=1e-4	Linear warmup + Step decay

For GDRO and Bi-level GDRO approaches we explored the following range of values for the MWU parameter $\eta \in \{0.0005, 0.001, 0.005, 0.01, 0.05, 0.1\}$ and selected those with the more stable behaviour based on CE loss. For VS and Polytail losses we used the loss parameters that suggested in the respective works. A summary of these is available in Table 3. In all cases an adversary update (MWU with parameter η on the group priors) was performed after an epoch of SGD. Algorithm 1 and subsequent theory results make use of all T classifiers; to deal with this cumbersome requirement in practice, we follow [Chen et al., 2017] and [Sagawa et al., 2019] (GDRO) where the deployed model is the final classifier. For each dataset we performed 5 experiments where we changed the random seed affecting the model initialization, data sampling and dataset split. If the dataset provided test partition the random seed only affected the validation and train split.

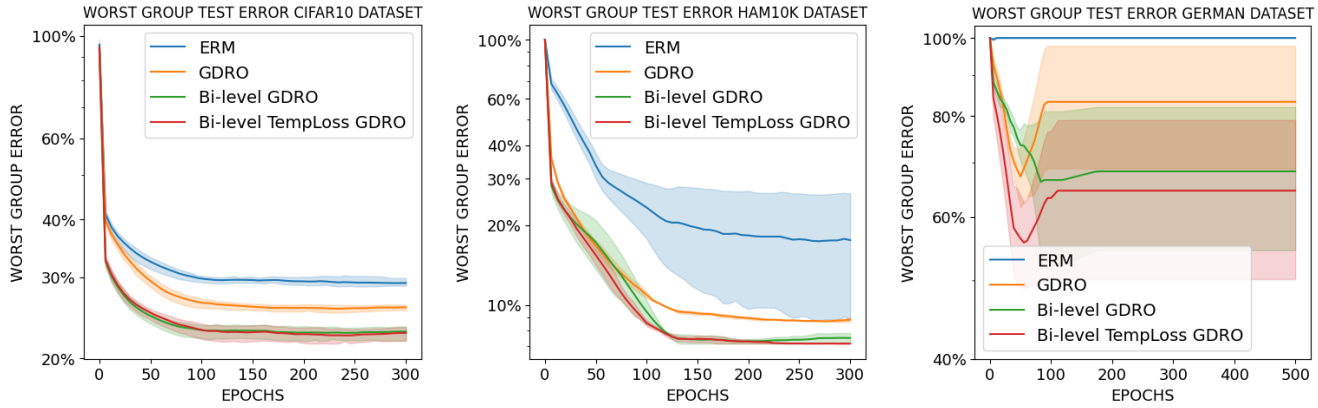
Table 3: Best set of parameters for each method.

Method	η	Loss Parameters
GDRO	0.1	-
GDRO VSLoss	0.1	VSLoss: $\gamma = 0.3$ (Section B.1 in [Kini et al., 2021])
GDRO Polytail	0.1	Polytail: α=1, $\beta = 1$ (Section 5 in [Wang et al., 2021])
Bi-level GDRO (All)	0.01/0.05 Waterbird/0.0005 German	CE, VS, Polytail params same as GDRO

TempLoss implementation and parameters. We make the following simplifications for a simple and efficient implementation of the (cross-entropy) TempLoss. First, we observe that $\lim_{\tau\to\infty} s(u) = \lim_{\tau\to\infty} 1 - e^{-\tau u} = 1 \ \forall u > 0$. Taking $\tau = \infty$ in Eq.(13) we get $v(u,q)\begin{cases} u \text{ if } u \leq 0, \\ \frac{u}{q} \text{ if } u > 0, \end{cases}$. Additionally, we observe that $\frac{u(f(x),y)}{q} = \frac{1}{q}[f_y(x) - f_{\bar{y}}(x) - \log\sum_{y'\neq y} e^{f_{y'}(x) - f_{\bar{y}}(x)}] \simeq u(\frac{f(x)}{q}, y)$, where $\bar{y} = \arg\max_{y'\neq y} f_{y'}(x)$. With this, we can simply approximate the cross-entropy TempLoss function as

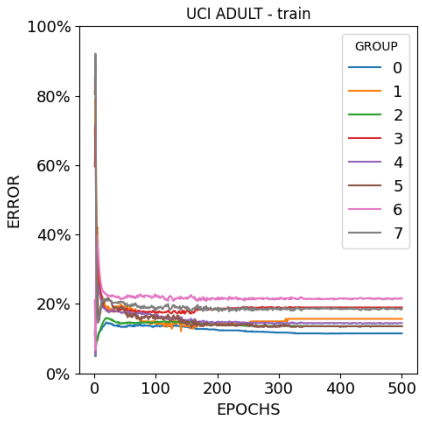
$$L(u,q) = \begin{cases} \ell(f(x),y) & \text{if } \ell(f(x),y) \leq \ell(u=0), \\ \ell(\frac{f(x)}{q}, y) & \text{if } o.w., \end{cases} \tag{48}$$

B.2 Additional Results

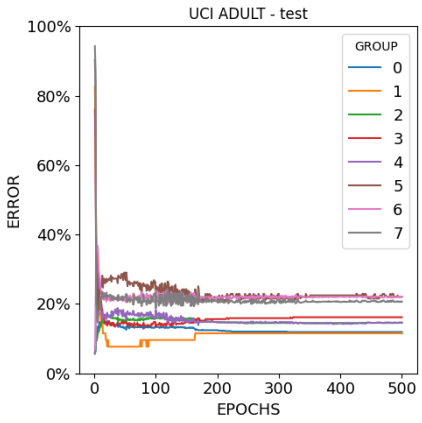


(a) CIFAR10 (b) HAM10K (c) German

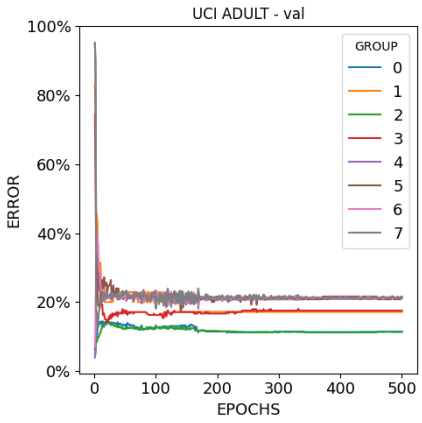
Figure 4: Evolution of worst group test error across training epochs on the CIFAR10, HAM10K and German datasets; deviations computed over 5 splits. We compare ERM; GDRO (adversary uses training group losses to update group priors); our proposed Bi-level GDRO, were the adversary uses the error on a held-out dataset to update group priors; and Bi-level TempLoss GDRO, where the learner additionally minimizes our proposed TempLoss. Bi-level GDRO outperforms regular GDRO, especially over the last training epochs. The TempLoss modification on the learner's objective further improves performance. Results are consistent across several datasets as shown in Section 5.



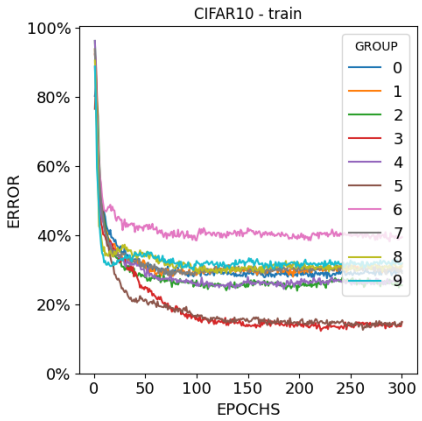
(a) Train group errors, Adult dataset



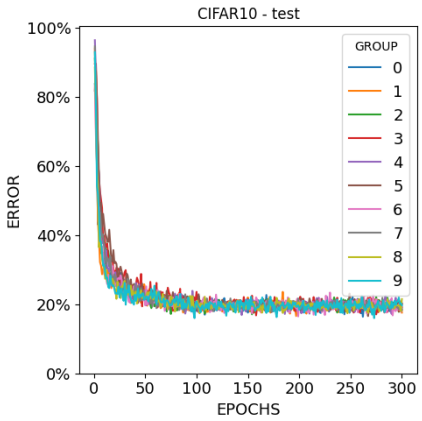
(b) Test group errors, UCI Adult



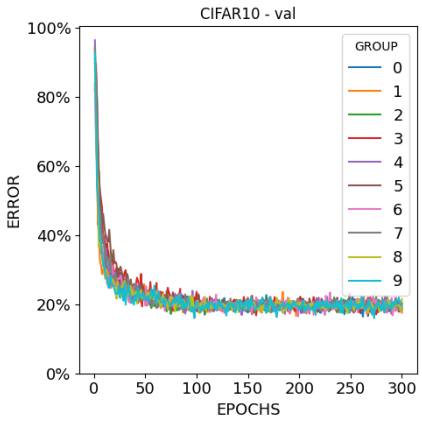
(c) Validation group errors, UCI Adult



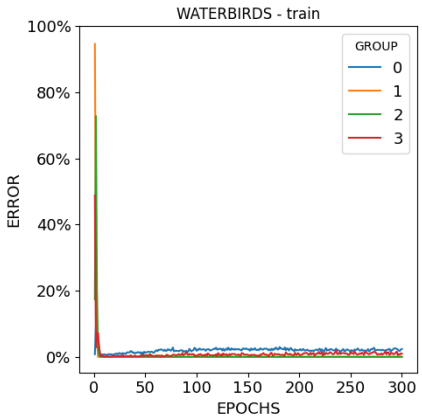
(d) Train group errors, CIFAR10



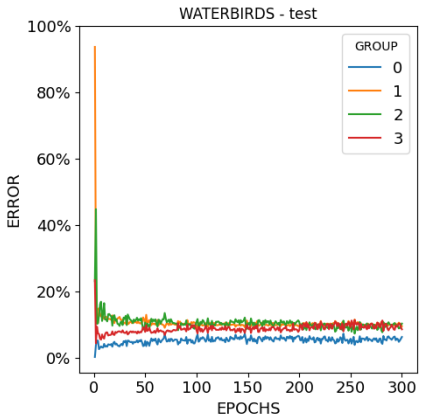
(e) Test group errors, CIFAR10



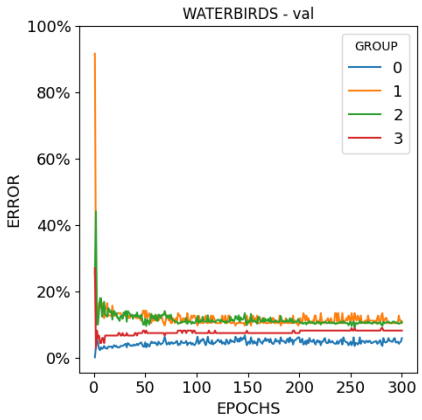
(f) Validation group errors, CIFAR10



(g) Train group errors, Waterbirds



(h) Test group errors, Waterbirds



(i) Validation group errors, Waterbirds

Figure 5: Evolution of all group errors on Adult, CIFAR10, and Waterbirds datasets across training epochs for one training run of Bi-level TempLoss GDRO. Train errors do not accurately reflect out-of-sample (test) errors for all groups. However, we observe validation errors to accurately track this out-of-sample error, in spite of their use as part of the adversarial update in Algorithm 1. This empirical result supports the idea of 'near-independence' of the validation errors, and motivate our choice to decouple training and adversarial datasets in our Bi-level formulation.

C Related Work

Minimax Baselines

GDRO [Sagawa et al., 2019] and similar approaches such as Minimax Group Fairness [Martinez et al., 2020, Diana et al., 2020] minimize the worst-group loss. However, when applied to neural networks, these methods rely on strong regularization techniques to provide good generalization on unseen data [Sagawa et al., 2019]. Just train twice by [Liu et al., 2021] is a two-stage algorithm that involves ERM followed by a second round of training with a weighted objective where the wrongly classified samples are up-weighted with the same coefficient. JTT faces the same challenges mentioned in our work, it relies on strong l2 regularization and early stopping to avoid empty error set, see their Section 4. The work by [Zhou et al., 2022] proposed a Bi-level formulation for importance weight learning in the context of out of distribution learning. Their proposed framework directly maps the space of importance weights into the model parameter space.

GDRO [Sagawa et al., 2019], [Martinez et al., 2020] and [Diana et al., 2020] share the minimax objective as well as implementation similarities ([Sagawa et al., 2019] and [Diana et al., 2020] use MWU as a potential adversarial update, [Martinez et al., 2020] uses a different projection to update the group priors), differing primarily on the assumptions made on the learner (e.g., [Diana et al., 2020] implementation uses logistic regression). Therefore, we considered that either method gives a representative view of the existing baselines, and GDRO was suitable for our experiments on neural networks since it already integrates gradient descent on the learner model updates.

The MWU algorithm (as implemented in [Sagawa et al., 2019, Diana et al., 2020]) comes with no regret guarantees independently of the learner's specific optimization strategy, and minimax guarantees as long as the learner achieves an α-approximate solution. Other Bi-level approaches such as [Blondel et al., 2022] or [Zhou et al., 2022] require the explicit or implicit differentiation of the model parameters w.r.t. the group priors $\mathcal{Q}$ in the inner loop which can be challenging for large models. These may require the outer level objective to be differentiable which for our purposes would imply the use of a surrogate loss for the adversary's objective instead of the current error objective.

Importance Weights on Interpolating Classifiers.

Several works have empirically shown that neural networks trained with SGD tend to be non-responsive to sample re-weighting in the asymptotic training regime [Sagawa et al., 2019, Byrd and Lipton, 2019, Xu et al., 2021], and proven for linear classifiers on linearly separable representation functions of the data. For the SGD-trained linear classifiers, [Nacson et al., 2019b] proved that the classifier converges towards a scaled max-margin solution, with [Xu et al., 2021] showing this to be true for neural networks trained with importance weights.

Prior work has attempted to characterize and change the asymptotic behaviour of the learned classifier with a variety of approaches. The work in [Buet-Golfouse, 2021] analyzed dataset conditions under which the classifier may fail to exhibit this max-margin behaviour. In [Wang et al., 2021], importance weights are applied to polynomial-tailed losses, which are asymptotically responsive to importance weights, but lack the interpretability of the max-margin finite sample solution and require the use of loss functions that are not directly related to standard statistical objectives such as minimal-entropy prediction. The work in [Cao et al., 2019] proposes the use of an additive term on the margin of the correct label to affect the effective per-label margin of the final classifier. The margin-additive correction term is applied after a predefined burn-in phase for numerical stability, and is computed based on the number of samples per label, which limits its use for minimax optimization. Authors in [Menon et al., 2020] propose a similar approach where a weighted label prior is added to the softmax crossentropy loss, increasing margins on low frequency labels. However, additive corrections to the classifier's output logits do not modify the exponential tail behaviour and must be combined with regularization techniques to be effective.

Additionally, the work in [Ye et al., 2020] identifies the need for multiplicative, rather than additive, modifications to the learned margins. The authors in [Kini et al., 2021] show that the multiplicative margins may be unstable during early training, before a good classification baseline is established so incorporate both additive and multiplicative correction terms to the logits in what they call the VS loss. Like with the preceding works, the margin corrections are decided *a-priori* and are not connected to the importance weights, both additive and multiplicative weights are maintained per label and group. [Narasimhan and Menon, 2021] proposes a loss with calibration guarantees based on logit adjustment. However, this method directly adjusts the per-label prior, but is not immediately clear how this can be extended to incorporate generic group adversarial priors in setting where

the groups do not match the classification labels. In settings where the groups match the classes, this loss can be used as the inner loop objective in our Bi-level GDRO formulation. The loss proposed by [Lu et al., 2022] shares similarities with our TempLoss and VS-loss. The main differences with our work are the absence of importance weighting, tempering being present for all samples - not just misclassified ones - and the tempering parameters are not learned like ours to optimize minimax error but instead are fixed based on the relative group priors present in the training dataset.

In our work, we smoothly interpolate between unmodified margins, and multiplicative-weighted margins on samples that are already well classified, which greatly simplifies the training procedure since no burn-in phase is needed, and only per-group weights are needed. We are also able to learn the correct weighting procedure to produce minimax-optimal classifiers within a set of target distributions. The work by [Kini et al., 2021], which proposes the VS loss is the most closely related to our definition with the difference that we explicitly formalize the connection with the adversarial weights in the GDRO setting, and thus do not require to make any prior assumptions on which group will be most disadvantaged. Moreover, our adaptation does not require any burn-in phase or separate (additive) corrections.

D Analysis of Convergence and Generalization in the Over-parametrized Regime

Here we extend the analysis of over-parametrization and memorization proposed in [Sagawa et al., 2020] in the context of our proposed Bi-level TempLoss objective.

The analysis consists of the following steps. We first describe a simple data generation process with binary labels and two groups that allows for even a linear model to memorize and over-fit to the training set; this data generation process has a single (core) feature that allows accurate prediction of the target label for both groups, and will further have a confounding (spurious) attribute that can greatly assist in the classification of samples from the majority groups while being detrimental to the minority group. Then, we show that a (linear) model trained with TempLoss with a fixed adversarial prior Q can be analyzed in terms of the max margin solution. In this setting we analyze two solutions, one that uses the core feature and generalizes well to both groups, and one that uses the spurious feature and memorization of the minority groups during training to produce a model that only has good test performance on the majority group. We compare the relative norms of these solutions as a function of the adversarial prior Q to show that TempLoss can steer the model towards one or the other solution. Finally we show that the adversary in Bilevel TempLoss has accurate group error estimates and therefore can learn the correct prior Q to maximize worst group accuracy.

Data Generation. Consider the following binary classification scenario where $y = \{-1, 1\}$ is the target label $a = \{-1, 1\}$ is a spurious attribute, and the input features x are described by the following data generation process:

$$\begin{aligned} y &\sim \text{Rademacher}(0.5), \\ a \mid y &\sim y \times \text{Rademacher}(\rho), \\ x^{\text{core}} \mid y &\sim \mathcal{N}(y, \sigma^2_{\text{core}}), \\ x^{\text{spu}} \mid a &\sim \mathcal{N}(a, \sigma^2_{\text{spu}}), \\ x^{\text{noise}} &\sim \mathcal{N}(0, \tfrac{\sigma^2_{\text{noise}}}{N} I_N), \\ x &= [x^{\text{core}}, x^{\text{spu}}, x^{\text{noise}}]. \end{aligned} \tag{49}$$

That is, the target labels y are equally likely to be -1 or 1, and the spurious attribute a takes the same value as y with probability ρ. The input features $x \in \mathbb{R}^{N+2}$ include a single feature x^{core} that relates to the target variable y, a spurious feature x^{spu} that can be used to predict a but not y, and several entries of uncorrelated noise. This dataset is characterized by two groups, the majority group (maj) satisfies $y = a$, while the minority group (min) satisfies $y = -a$.

For the purposes of the analysis, we assume we have access to a training dataset with n samples $n = n_{maj} + n_{min}$, where n_{maj} denotes samples from the majority group ($y = a$) and n_{min} denotes the converse. We additionally assume σ^2_{spu}) is small with $\sigma^2_{\text{spu}}) < \sigma^2_{\text{core}})$, and $N \gg n$, in this scenario, the training dataset is linearly separable with high probability (a model can "memorize" a training point via its noise component).

Model Learning. We therefore analyze a linear model ($\phi^T x$) trained with TempLoss for a given (adversarial) prior Q over the majority and minority groups, this setting satisfies the conditions of Proposition 3.9, that is to say, for a given (adversarial) prior Q over the majority and minority groups, the resulting model parameters can be characterized by the (weighted) max margin solution

$$\begin{aligned} \lim_{t\to\infty} \frac{\theta^t}{||\theta^t||_2} &\to \frac{\hat{\theta}}{||\hat{\theta}||_2}, \\ \hat{\theta} &:= \arg\min_{\theta\in\mathbb{R}^(N+2)} ||\theta||_2^2 \\ &\quad \text{s.t. } \{y_i\theta^T x_i) \geq Q_{g_i}\}_{i=1}^n. \end{aligned} \tag{50}$$

Max Margin Cost and Error Comparison. We focus on the analysis of the max margin solution $\hat{\theta}$. Following the analysis in [Sagawa et al., 2020], we decompose $\hat{\theta} = [\hat{\theta}^{\text{core}}, \hat{\theta}^{\text{spu}}, \hat{\theta}^{\text{noise}}]$

Note that the test error for a linear model parametrized by $\hat{\theta}$ for both groups can be computed as

$$\begin{aligned} \sigma_\theta^2 &:= (\hat{\theta}^{\text{core}})^2\sigma_{\text{core}}^2 + (\hat{\theta}^{\text{spu}})^2\sigma_{\text{spu}}^2 + ||\hat{\theta}^{\text{noise}}||^2\sigma_{\text{noise}}^2, \\ y\theta^T x|a,y &\sim \mathcal{N}(\hat{\theta}^{\text{core}} + ay\hat{\theta}^{\text{spu}}; \sigma_\theta^2), \\ Pr[y\theta^T x \leq 0 \mid y = a] &= \phi(-\tfrac{\hat{\theta}^{\text{core}}+\hat{\theta}^{\text{spu}}}{\sqrt{\sigma_\theta^2}}), \\ Pr[y\theta^T x \leq 0 \mid y \neq a] &= \phi(-\tfrac{\hat{\theta}^{\text{core}}-\hat{\theta}^{\text{spu}}}{\sqrt{\sigma_\theta^2}}), \end{aligned} \tag{51}$$

where ϕ denotes the cdf operator of a standard distribution. This formalizes the notion that positive values of $\hat{\theta}^{\text{core}}$ benefit classification for both groups, while positive values of $\hat{\theta}^{\text{spu}}$ benefit the majority group at the detriment of the minority group. Therefore, it is straightforward to observe that the **solution that minimizes worst group error must satisfy $\hat{\theta}^{\text{spu}} = 0$.**

We continue the analysis by noting that the representer theorem allows us to express $\hat{\theta}^{\text{noise}}$ as

$$\hat{\theta}^{\text{noise}} = \sum_{i\in[n]} \alpha_i x_i^{\text{noise}}. \tag{52}$$

This decomposition has a few interesting consequences, since we assume N to be large, we have $||x_i^{\text{noise}}||_2^2 = \sigma_{\text{noise}}^2$ with high probability, and since we also assume $N \gg n$ then with high probability any two noise samples in the training dataset satisfy $< x_i^{\text{noise}}, x_j^{\text{noise}} >= 0$ (i.e., the noise components of the samples are orthogonal). This implies both of the following

$$\begin{aligned} ||\hat{\theta}^{\text{noise}}||_2^2 &= \sigma_{\text{noise}}^2 \textstyle\sum_{i\in[n]} \alpha_i^2, \\ (\hat{\theta}^{\text{noise}})^T x_i^{\text{noise}} &= \sigma_{\text{noise}}^2 \alpha_i. \end{aligned} \tag{53}$$

In other words, we can increase the margin of any sample x_i by setting α_i to be sufficiently high without affecting the prediction in any other training sample w.h.p., and the (norm) cost of memorizing sample i with strength α_i (and margin increase $\alpha_i\sigma_{\text{noise}}^2$) is $\alpha_i\sigma_{\text{noise}}^2$.

Now consider the following sets of solutions

$$\begin{aligned} \Theta_{\text{use-spu}} &:= \{\theta : \theta \text{ is a separator}, \theta^{\text{core}} = 0\}, \\ \Theta_{\text{use-core}} &\{\theta : \theta \text{ is a separator}, \theta^{\text{spu}} = 0\}. \end{aligned} \tag{54}$$

Where we take similar definitions to [Sagawa et al., 2020] but here 'separator' indicates a solution that linearly separates each data point with margin at least Q_{g_i}

To compare the norms of these two types of solutions, we extend propositions 1 and 2 in [Sagawa et al., 2020] to the setting of weighted margin norms. We provide the proof sketch similar to [Sagawa et al., 2020] and note that the full proof is a straightforward (but lengthy) adaptation of theirs. For convenience we denote Q_{maj} as the (adversarial) prior of the majority group ($y = a$) and $Q_{\text{min}} = 1 - Q_{\text{maj}}$ the prior of the minority group ($y \neq a$)

Proposition Under mild conditions (see Theorem 1 [Sagawa et al., 2020]) there exists a (weighted) separator $\theta_{\text{use-spu}} \in \Theta_{\text{use-spu}}$ and constant $\gamma_1 > 0, \eta > 0$ such that

$$\begin{aligned} ||\theta_{\text{use-spu}}||_2^2 \leq & \quad (\gamma_1 Q_{\text{maj}})^2 + \frac{n_{\text{min}}}{\sigma^2_{\text{noise}}}(Q_{\text{min}} + \eta\gamma_1 Q_{\text{maj}}), \\ ||\theta_{\text{use-spu}}||_2^2 \leq & \quad (\gamma_1 Q_{\text{min}})^2 + \frac{n_{\text{maj}}}{\sigma^2_{\text{noise}}}(Q_{\text{maj}} + \eta\gamma_1 Q_{\text{min}}), \end{aligned} \tag{55}$$

Proof sketch For simplicity, we focus on the first inequality. Since σ^2_{spu} is small, w.h.p there exists some constant γ_1 such by setting $\theta^{\text{spu}}_{\text{use-spu}} = \gamma_1$, the majority points have a margin > 1, therefore, by setting $\theta^{\text{spu}}_{\text{use-spu}} = \gamma_1 Q_{\text{maj}}$ all majority samples satisfy their margin condition. However, for the minority training points, the attribute a is anti-correlated with the label, and w.h.p. there exists some constant η that depends on σ^2_{spu} such that the decrease in the margin due to $\theta^{\text{spu}}_{\text{use-spu}} = \gamma_1 Q_{\text{maj}}$ is at most $-\eta\gamma_1 Q_{\text{maj}}$ w.h.p.. To satisfy the margin condition on minority samples, it suffices to set $\alpha_i = y_i \frac{Q_{\text{min}} + \eta Q_{\text{maj}}}{\sigma^2_{\text{noise}}}$ and the bound on the norm follows.

The second inequality can be similarly derived for a different set of constants γ_1', η' by considering a similar scenario where $\theta^{\text{spu}} = -\gamma_1'$ and the majority samples are the ones being memorized (this is potentially a viable solution for very low values of Q_{maj}). The proof can be concluded by taking maximums over the two sets of constants.

Proposition Under mild conditions (see Theorem 1 [Sagawa et al., 2020]) there exists a (weighted) separator $\theta_{\text{use-core}} \in \Theta_{\text{use-core}}$ and constant $\gamma_3 > 0$ such that

$$||\theta_{\text{use-core}}||_2^2 \geq \gamma_3 \frac{n}{\sigma^2_{\text{noise}}} \max\{Q_{\text{maj}}, Q_{\text{min}}\} \tag{56}$$

Proof sketch This is a direct application of Proposition 3 from [Sagawa et al., 2020] which states that there exists a parameter θ_* with norm $||\theta_*||_2^2 \leq \gamma_3 \frac{n}{\sigma^2_{\text{noise}}}$ s.t. $\theta^{\text{spu}}_* = 0$ and the margin of each training sample is > 1. The key insight of this proof is that there is a constant fraction of samples s.t. $x^{\text{core}} y \leq 1$ (i.e., the core feature is noisy enough that some samples are misclassified by it), therefore, this fraction of samples needs to be memorized to linearly separate the training set.

Depending on the problem parameters, there are scenarios in which carefully chosen values of $Q_{\text{min}}, Q_{\text{max}}$ yield $||\theta_{\text{use-core}}||_2^2 \leq ||\theta_{\text{use-spu}}||_2^2$. Therefore, a model trained with the correct adversarial prior on TempLoss will be able to achieve the optimal minimax error rates across groups.

Adversary Convergence and Error Estimates. The convergence of Algorithm 1 was already shown in general in Theorem 3.4. Here, we simply observe that the adversary's error estimates on its held-out dataset are fully unaffected by memorization, since any sample x that is part of the adversary's, but not the model learner's, dataset, satisfies $\theta^T x = \theta_{\text{core}} x_{\text{core}} + \theta_{\text{spu}} x_{\text{spu}}$ w.h.p. (since the adversary's samples are orthogonal to the model learner's samples w.h.p).

Therefore, the adversary's samples satisfy

$$\begin{aligned} Pr[y\theta^T x \leq 0 \mid y = a] = & \quad \phi(-\frac{\hat{\theta}^{\text{core}} + \hat{\theta}^{\text{spu}}}{\sqrt{\sigma^2_{\hat{\theta}}}}), \\ Pr[y\theta^T x \leq 0 \mid y \neq a] = & \quad \phi(-\frac{\hat{\theta}^{\text{core}} - \hat{\theta}^{\text{spu}}}{\sqrt{\sigma^2_{\hat{\theta}}}}), \end{aligned} \tag{57}$$

and would therefore correctly modify the prior Q until minimax error is achieved.