Proceedings of Machine Learning Research vol 237:1–36, 2024 35th International Conference on Algorithmic Learning Theory

Semi-supervised Group DRO: Combating Sparsity with Unlabeled Data

Pranjal Awasthi PRANJALAWASTHI@GOOGLE.COM
Google Research

Satyen Kale SATYENKALE@GOOGLE.COM
Google Research

Ankit Pensia ANKITP@IBM.COM
IBM Research

Editors: Claire Vernade and Daniel Hsu

Abstract

In this work we formulate the problem of group distributionally robust optimization (DRO) in a semi-supervised setting. Motivated by applications in robustness and fairness, the goal in group DRO is to learn a hypothesis that minimizes the worst case performance over a pre-specified set of groups defined over the data distribution. In contrast to existing work that assumes access to labeled data from each of the groups, we consider the practical setting where many groups may have little to no amount of labeled data.

We design near optimal learning algorithms in this setting by leveraging the unlabeled data from different groups. The performance of our algorithms can be characterized in terms of a natural quantity that captures the similarity among the various groups and the maximum *best-in-class* error among the groups. Furthermore, for the special case of squared loss and a convex function class we show that the dependence on the best-in-class error can be avoided. We also derive sample complexity bounds for our proposed semi-supervised algorithm.

Keywords: semi-supervised, group DRO, min-max fairness, data sparsity

1. Introduction

Machine learning algorithms are being increasingly deployed in various critical applications e.g., criminal justice, healthcare and finance (Angwin et al., 2019). In such scenarios it is important to ensure that the model is not only accurate overall, but also satisfies additional properties such as robustness and fairness (Szegedy et al., 2014; Buolamwini and Gebru, 2018).

The notion of min-max optimization is an elegant mathematical framework for designing algorithms for the above mentioned criteria. Here one aims to find a classifier that minimizes the maximum of a number of losses. For example, in distributionally robust optimization (DRO) (Sinha et al., 2018; Namkoong and Duchi, 2016; Kuhn et al., 2019), the maximum is taken over expected losses of the classifier over distributions in a certain ball around a given target distribution. Recent works have also formulated the group DRO setting that considers min-max optimization over a finite set of distributions (Sagawa et al., 2019). This setting has many practical applications. For instance, in existing literature on fair machine learning a standard fairness measure known as min-max fairness or *Rawlsian* fairness (Martinez et al., 2020; Diana et al., 2021) is an instance of group DRO where the maximum is taken over the expected losses of the classifier over certain distributions which are associated with groups of users. Apart from the application to min-max fairness Sagawa et al. (2019) apply group DRO to train neural networks that can avoid learning spurious correlations.

Similarly, the work on collaborative PAC learning considers a setting where a finite set of groups are coordinating to learn a hypothesis that is good for all of them (Blum et al., 2017). Motivated by the above applications, in this paper we focus on the setting of group DRO, i.e., min-max optimization over a finite set of distributions.

Much of the prior research on group DRO and min-max optimization in general has been in the *data-rich* regime, i.e., settings where each group has a sufficient number of labeled samples. However, these methods are increasingly being applied to situations where many groups have limited or no labeled data at all. We call such groups *sparse*, and data-rich groups *dense*. As an example consider a large scale recommendation system that consists of content creators and end users. To keep such systems safe for the end users it is critical to have in place classifiers that can automatically filter out harmful content. Typically, a limited number of human raters are used to analyze a small amount of content each day and label it as harmful or not (Deodhar et al., 2022). This labeled data is then used to train the classifier.

Furthermore, such classifiers should not only be good on average but should also have high performance across different slices of the user population that can be defined by attributes such *race, sex, gender, location,* etc., or any combination of them. This can easily result in a min-max optimization problem over hundreds of groups where most of the groups will have a very limited amount of labeled data, if at all. As algorithm designers strive to ensure a certain performance level for an increasing number of groups, this issue of label sparsity is likely to become more prevalent.

In this work, we address the above scenario and initiate a study of group DRO under the semi-supervised setting where one has access to unlabeled data from the groups and a limited amount labeled data from a subset of them. Our contributions are as follows:

1. We formalize the group DRO problem in the semi-supervised setting. As a measure of performance we consider the standard min-max loss (say the $0/1$ classification loss), and the recently proposed min-max regret (Agarwal and Zhang, 2022). We show that some natural approaches for min-max optimization fail in the setting of this paper and that without structural assumptions, the problem becomes impossible due to lack of sufficient data in some groups.
2. We introduce a very mild structural assumption under which we propose a natural two-step procedure that first performs a careful pseudo-labeling of the unlabeled data points followed by invoking an existing algorithm for standard min-max optimization.
3. We show that the proposed algorithm incurs an additive error over the optimal classifier for the min-max problem that can be decomposed into two terms: a) the maximum, over all the groups, of the minimal loss of a classifier from the family of classifiers of interest, and b) a notion of closeness among the groups induced by the structural assumption. We give lower bounds showing that a dependence on both the terms is necessary in the worst case. We also develop sample complexity bounds for our proposed algorithm.
4. We propose and analyze an extension of our main algorithm that algorithmically infers the trade-off among the labeled and the unlabeled data for each group in a data dependent manner.

While the primary contributions are theoretical in nature, Section E also provides some proof-of-concept experimental results for the proposed algorithm.

1.1. Related Work

The notion of min-max optimization has been studied in several contexts. In the design of fair machine learning models, min-max fairness has been proposed as a natural notion of group fairness

(Cotter et al., 2019; Diana et al., 2021; Abernethy et al., 2022; Martinez et al., 2020). The works of Agarwal et al. (2018); Cotter et al. (2019) proposed general algorithms for min-max optimization via connections to no-regret learning in two player games. In addition, the works of Kearns et al. (2018, 2019) consider min-max optimization under exponentially many groups and provide algorithms under the assumptions that groups belong to a hypothesis class of finite VC dimension. The work of Martinez et al. (2020) presents structural results regarding the Pareto optimal min-max classifiers and the recent work of Diana et al. (2021) presents practical algorithms for min-max fairness based on the multiplicative weights update method. There has also been recent work on adaptive sampling methods for min-max optimization (Abernethy et al., 2022; Shekhar et al., 2021; Haghtalab et al., 2022). In particular, the work of Haghtalab et al. (2022) provides near optimal sample complexity bounds for the case of hypothesis sets with a finite Littlestone dimension.

Min-max optimization is also widely studied in the design of robust classifiers. Building upon the framework of robust optimization (Ben-Tal et al., 2009), the work of Duchi et al. (2021) proposed an approach to handling distribution shift where the goal is to design a classifier that optimizes the worst case loss over distribution that are close to a given reference distribution. This setting corresponds to min-max optimization over an infinite set of losses. Later works of Namkoong and Duchi (2016); Levy et al. (2020) presented large scale methods for min-max optimization when the closeness between distributions is measured with certain f-divergence measures.

The works of Gao et al. (2022); Kuhn et al. (2019); Blanchet et al. (2021) study min-max optimization under Wasserstein distances. The work of Chen et al. (2017) presents a general approach for robust optimization over a finite set of non-convex loss functions. Also, the works of Mohri et al. (2019); Cortes et al. (2020) present algorithms to perform min-max optimization over multiple mixtures of a given set of base objectives. Closely related is the line of work on adversarial robustness where the goal is to design classifiers that are robust to test-time adversarial perturbations (Madry et al., 2018; Salman et al., 2019; Feige et al., 2015; Attias et al., 2022; Montasser et al., 2019).

In the setting of semi-supervised group DRO that we consider our proposed algorithms will learn to leverage data from certain dense groups to alleviate the sparsity on the remaining groups. This is broadly related to works on multitask learning and transfer learning where data from one of more tasks is used to learn a good classifier for a new task with small number of labeled samples (Baxter, 2000; Cavallanti et al., 2010; Lounici et al., 2011; Maurer et al., 2016). However a crucial difference is that instead of designing task (or group) specific classifiers, we are interested in a single classifier that does well across all the groups.

1.2. Notation

We consider a supervised learning problem, where features belong to $\mathcal{X}$ and labels belong to $\mathcal{Y}$. We have a collection of "groups" indexed by a set $\mathcal{G}$. Each $g \in \mathcal{G}$ is associated with a distribution P_g over $\mathcal{X} \times \mathcal{Y}$. Let $\mathcal{F}$ be a family of predictive functions mapping $\mathcal{X}$ to $\mathcal{Y}$ and let $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_+$ be a loss function (viz., $\ell(\hat{y}, y)$ is the loss of prediction $\hat{y}$ for true label y). For a function $f : \mathcal{X} \to \mathcal{Y}$ and a distribution P on $\mathcal{X} \times \mathcal{Y}$, we denote the average loss of f on P by

$$\mathsf{Loss}(f, P) := \mathop{\mathbf{E}}_{(x,y)\sim P} \left[\ell(f(x), y)\right]. \tag{1}$$

For a set $S \subseteq \mathcal{X} \times \mathcal{Y}$, we define the notation $\mathsf{Loss}(f, S)$ to be the average loss on S, i.e., $\mathsf{Loss}(f, S) := \frac{1}{|S|} \sum_{(x,y)\in S} \mathsf{Loss}(f(x), y)$. In the learning problem of interest, for each $g \in \mathcal{G}$, we are provided a sample set of (independent) labeled examples $S_g \subseteq \mathcal{X} \times \mathcal{Y}$ drawn from P_g, and a set of (independent) unlabeled examples $U_g \subseteq \mathcal{X}$ drawn from the marginal of P_g over $\mathcal{X}$.

1.3. Metrics of Interest

The optimal min-max loss classifier in $\mathcal{F}$ is the one that minimizes the maximum loss over all the groups (Martinez et al., 2020). The loss of this classifier is then:

$$\mathsf{OPT}_{\mathsf{L}} = \min_{f \in \mathcal{F}} \max_{g \in \mathcal{G}} \mathsf{Loss}(f, P_g).$$

In the setting of interest in this paper, different groups may have varying amounts of inherent noise in their labels—in fact, sparse groups are more likely to have more inherent label noise.[1] The min-max criterion defined above is highly sensitive to the level of label noise across the groups, since the minimal loss of *any* classifier (not just the min-max optimal one) on a given group is at least the inherent label noise in that group. Hence, in the presence of widely different amounts of label noise across the groups, the min-max criterion might potentially yield a bad classifier. To handle this situation, inspired by the work of Agarwal and Zhang (2022), we also consider a different performance measure that removes the inherent label noise in the groups: viz., *regret*. The regret of a classifier f on a distribution P is:

$$\mathsf{Regret}(f, P) := \mathsf{Loss}(f, P) - \inf_{f' \in \mathcal{F}} \mathsf{Loss}(f', P).$$

Since the inherent label noise affects all classifiers equally, the regret notion defined above effectively removes label noise from consideration. In addition, the regret captures more accurately the degradation in performance from not choosing the optimal classifier in $\mathcal{F}$ for the distribution P. We refer the reader to Agarwal and Zhang (2022) for an in-depth discussion of the regret as a performance measure of classifiers. Thus, we also consider the optimal min-max regret classifier: this is the classifier in $\mathcal{F}$ that minimizes the maximum regret over all groups, and its regret is then:

$$\mathsf{OPT}_{\mathsf{R}} = \min_{f \in \mathcal{F}} \max_{g \in \mathcal{G}} \mathsf{Regret}(f, P_g).$$

We provide performance guarantees for our algorithms in terms of both $\mathsf{OPT}_{\mathsf{L}}$ and $\mathsf{OPT}_{\mathsf{R}}$.

1.4. Failure of Standard Approaches with Limited Data

Existing research has studied min-max fairness in the data-rich regime, where each group has a number of labeled samples that is comparable to the complexity of the hypothesis class, e.g., VC dimension or Rademacher complexity (Diana et al., 2021; Sagawa et al., 2019). In such cases, the empirical loss (and regret) uniformly approximates the population-level loss (and regret) for every group. As a result, if one chooses the function that minimizes the maximum empirical loss over all groups, then the returned function $\widehat{f}$ satisfies that $\max_g \mathsf{Loss}(\widehat{f}, P_g)$ is comparable to $\mathsf{OPT}_{\mathsf{L}}$; a similar conclusion holds for regret as well.

[1] For a group g, let $L_g^* := \min_{f \in \mathcal{F}} \mathsf{Loss}(f, P_g)$. The inherent noise in labels in a group g corresponds to L_g^*, which inherently depends on the quality of the training data as follows: If there is no consensus among the human/machine annotators for certain inputs (which typically happens more frequently on sparse groups — the groups with limited samples), the labels have more noise, increasing L_g^*.

If some groups are sparse, then the above approach fails terribly. In particularly, the returned hypothesis may overfit to the limited number of labeled samples in the sparse group. Concretely, we show a simple example below that highlights (i) standard approaches fail if some groups are sparse and (ii) it is still possible to learn a near optimal min-max classifier if the groups are related.

Example 1 *Let $\mathcal{G} = \{1, 2, 3\}$. Let u be an unknown unit vector in $\mathbb{R}^d$ and let z be an unknown sign in $\{+1, -1\}$. The distribution of each group is given below:*

- *For Group 1, (x, y) is distributed as $x \sim \mathcal{N}(0, I)$ and $y = u^\top x$,*
- *For Group 2, (x, y) is distributed as $x \sim \mathcal{N}(0, I)$ and $y = -u^\top x$,*
- *For Group 3, (x, y) is distributed as $x \sim \mathcal{N}(0, \sigma^2 I)$ and $y = zu^\top x$ for some $\sigma^2 \gg 1$.*

We are interested in finding a min-max optimal linear predictor $\beta \in \mathbb{R}^d$, the function class $\mathcal{F}$ is affine, and the loss function is the square loss. Let $\epsilon \in (0, 1)$ be small enough. We assume that for groups 1 and 2 are dense and have $(1 - \epsilon)/2$ fraction of all training examples each, with the remaining ϵ fraction in group 3, which is a sparse group. By simple calculations, we have that $\mathsf{OPT}_\mathsf{L} = \frac{4\sigma^2}{(\sigma+1)^2} \leq 4$.

Note that even though the fraction of training examples in the third group is small, a lot of information of the third group is captured by the first two groups (up to the unknown sign z). In particular, the above example satisfies the structural assumption that we propose in this work for the design of efficient algorithms (see Section 2). Still, as discussed below, the standard approaches fail to adapt to this underlying structure and perform poorly in the data-scarce regime:

- **Maximum Empirical Loss Minimizer** Here we find the candidate β that minimizes the maximum of empirical group-wise loss. However, since the third group does not have enough data, there are many spurious vectors β' that have zero empirical loss on the third group but large max loss. In fact, unless the number of samples in the third group is $\Omega(d)$, there will exist a hypothesis β whose empirical loss is zero on the third group, but with max loss over the distribution scaling with σ^2. This can be seen from Figure 1 where the performance of the empirical minmax optimizer approaches the upper bound on OPT_L as the number of labeled samples from the sparse group approaches the data dimensionality (d is 20 in this case).

- **Ignoring Sparse Groups in Training Data** Another natural approach is to simply ignore groups that are sparse and perform min-max optimization over the dense groups. This is especially appealing if the dense groups capture a lot of information about the sparse ones as is the case in Example 1. However, if we ignore the third (sparse) group while training, by symmetry, any reasonable algorithm will output origin (or an unbiased estimate of it) as $\widehat{\beta}$. However, the max loss incurred by origin is σ^2, which could be much larger than OPT_L as can be seen from Figure 1.

Figure 1 shows that our proposed algorithm achieves near optimal error even when the number of labeled samples is one! In the sequel, we will prove that our proposed algorithm achieves $O(\mathsf{OPT}_\mathsf{L})$ error as long as we observe a constant number of samples from the sparse group (as opposed to linearly many) and $O(d)$ samples overall. See Section A for more details.

2. Structural Assumptions

If there is no relationship between the distributions P_g for different groups, then it is impossible to learn a good classifier on a sparse group — let alone a min-max optimal one — simply because there might not be enough labeled examples in sparse groups and there is no way to infer additional

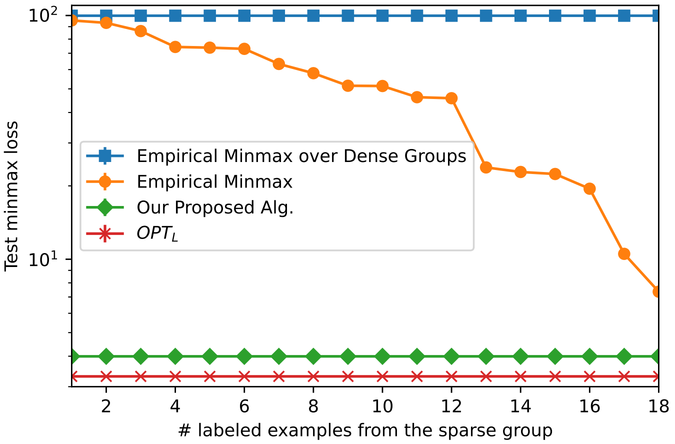

Figure 1: A plot showing that the standard approaches fail in the data scarce regime (number of labeled samples is less than the dimension, 20) even when the group distributions are related. The x-axis shows the number of labeled examples from the sparse group. The y-axis (in log-scale) shows the test max loss of the baselines and our proposed algorithm. See Example 1 for more details.

information via other groups. Thus, to have any hope of learning a min-max optimal classifier we need to assume certain "closeness" relationships between any given sparse group and a dense group allowing us to infer useful information on the sparse group despite the paucity of labeled examples.

One natural closeness assumption is that for every sparse group $g \in \mathcal{G}$, there is a dense group g' such that the distributions P_g and $P_{g'}$ are close in total variation or Wasserstein distance. This is a very strong notion of similarity and is often unrealistic in practice. One way to relax this assumption is to notice that since we are interested only in the performance of classifiers in $\mathcal{F}$ on the groups, it is enough that for every sparse group $g \in \mathcal{G}$, there is a dense group g' such that $\max_{f \in \mathcal{F}} |\mathsf{Loss}(f, P_g) - \mathsf{Loss}(f, P_{g'})| \leq \Delta$, for some closeness parameter $\Delta \geq 0$. This assumption is reminiscent of the notions of discrepancy that are widely studied in literature on domain adaptation (Ben-David et al., 2010; Cortes et al., 2015; Zhang et al., 2020). Although this assumption is considerably weaker than the first one, the fact that the condition needs to hold for *every* $f \in \mathcal{F}$ still imposes strong conditions on the relationship among the two groups. For example, when using the squared loss with a linear function class, the assumption implies that the two groups g and g' have similar covariance matrices of features and similar optimal regressors.

Thus, both of the candidate assumptions discussed above are rather strong notions of similarities and may be unrealistic in practice. In fact, under those assumptions, simply ignoring the sparse group from our optimization problem results in a classifier without a significant drop in the performance from the optimal one. Therefore, we propose a significantly weaker notion of similarity that we will work with:

Algorithm 1 Algorithm for Group DRO: Idealized Setting

```
Require: Groups $\mathcal{G}$, function class $\mathcal{F}$, loss function $\ell$, labeled data $S$, unlabeled data $U$, a partition of $\mathcal{G}$ into dense and sparse groups. Moreover, for each sparse group $g$, the identity of the dense group that attains the minimum $\min_{g':\text{dense}} \mathsf{Loss}(f^*_{g'}, P_g)$ is known.
for each dense group $g \in \mathcal{G}$ do
    Let $\widetilde{f}_g \leftarrow \mathrm{argmin}_{f \in \mathcal{F}} \mathsf{Loss}(f, S_g)$.
    Define empirical regret: for all $f \in \mathcal{F}$, $\mathsf{Regret}_{\mathsf{emp}}(f, g) := \mathsf{Loss}(f, S_g) - \mathsf{Loss}(\widetilde{f}_g, S_g)$.
end for
for each sparse group $g \in \mathcal{G}$ do
    Set $\widehat{g} \leftarrow$ the dense group that attains the minimum $\min_{g':\text{dense}} \mathsf{Loss}(f^*_{g'}, P_g)$.
    Create proxy data from unlabeled data $S^{\text{proxy}} := \{(x, \widetilde{f}_{\widehat{g}}(x), g) : (x, g) \in U\}$.
    Define proxy regret using above: $\forall f \in \mathcal{F}$, $\mathsf{Regret}_{\mathsf{proxy}}(f, g) := \mathsf{Loss}(f, S^{\text{proxy}}_g)$.
end for
Solve the following min-max problem:
    $\widehat{f} = \mathrm{argmin}_{f \in \mathcal{F}} \max\left(\max_{g:\text{dense}} \mathsf{Regret}_{\mathsf{emp}}(f, g), \max_{g:\text{sparse}} \mathsf{Regret}_{\mathsf{proxy}}(f, g)\right)$.
return $\widehat{f}$
```

Assumption 1 (Notion of Similarity: good performance of the optimizer) *For every sparse group $g \in \mathcal{G}$, there is a dense group $g' \in \mathcal{G}$ such that $\mathsf{Loss}(f^*_{g'}, P_g) - \mathsf{Loss}(f^*_g, P_g) \leq \Delta$, where $f^*_{g'}$ and f^*_g are the optimal predictor of groups g' and g, respectively.* [2]

This notion of similarity captures the idea that the marginal distributions of these two groups can be far apart but the conditional distribution of responses is similar. When compared to the notions of discrepancy used in domain adaptation, we only require the optimal predictor for a dense group to perform well on the sparse group. That is, the function $f^*_{g'}$ — the optimal predictor for group g' — continues to perform well on the group g and thus approximates the conditional distribution of responses on g.

Assumption 1 is inspired by the observation that it is often cheap to collect unlabeled data, and use it to approximate the marginal distribution of sparse groups. Thus, the notion of similarity in this sense should focus only on approximating the conditional distribution of the responses.

3. Algorithm in the Idealized Setting

In this section we outline our main algorithm (Algorithm 1) for the problem of min-max optimization in the semi-supervised setting. For simplicity of exposition we will assume the *idealized setting*, i.e., we have access to an infinite amount of labeled examples from the dense groups and an infinite amount of unlabeled examples from the sparse groups and that the sparse groups have no labeled examples. Note that this implies that we know the identity of the dense and the sparse groups. Furthermore, we will also assume that we know apriori the closest dense group for each sparse group. Here closeness refers to the notion defined in Assumption 1. This idealized setting is stated below. In later sections we will generalize our algorithm to settings where such information is not known apriori.

[2] That is, $f^*_{g'} = \mathrm{argmin}_{f \in \mathcal{F}} \mathsf{Loss}(f, P_{g'})$ and $f^*_g = \mathrm{argmin}_{f \in \mathcal{F}} \mathsf{Loss}(f, P_g)$.

Setting 1 (Idealized Setting) *The dense groups have infinite labeled examples. In contrast, the sparse groups have no labeled examples but they have infinite unlabeled examples. Furthermore, each group has a unique minimizer of the loss function, and for each sparse group g, we know the identity of the dense group g' that attains minimum $\min_{g':dense} \mathsf{Loss}(f^*_{g'}, P_g)$.*

Under the above idealized setting our proposed algorithm (Algorithm 1) follows a simple and natural approach that surprisingly leads to near optimal error guarantees. Our algorithm consists of three key steps. In the first step, we compute, for each dense group g, the optimal classifier $\widetilde{f}_g$ obtained via empirical risk minimization over the labeled samples S_g from the group. The second and the third steps involve key technical contributions that enable us to deal with sparse groups. In the second step the algorithm performs a pseudo-labeling of the sparse groups. In particular, for each sparse group g the unlabeled data is labeled using the classifier $\widetilde{f}_{g'}$ where g' is the dense group closest to g (as defined in Setting 1). Let this pseudo-labeled dataset be S_g^{proxy} as constructed in step 7 of Algorithm 1.

In the third step, we define the regret for the sparse group using the proxy dataset. In other words, let $\mathsf{Regret}_{\mathsf{proxy}}(f, g) := \mathsf{Loss}(f, S_g^{\text{proxy}}) - \min_{f \in \mathcal{F}} \mathsf{Loss}(f, S_g^{\text{proxy}})$. Note that since the proxy dataset is labeled via a function in the class $\mathcal{F}$, the second term equates to zero. Furthermore, we will show that if the sparse group is close to the dense group that is being used for constructing the proxy dataset, then using proxy regret only incurs a small additive error that depends on the closeness parameter Δ. This motivates the use of the proxy regret in step 8 of Algorithm 1.

We then output $\widehat{f}$ that optimizes the min-max regret where the standard notion of regret is used for the dense groups and the proxy measure is used for the sparse groups. Note that Algorithm 1 is computationally-efficient as it can be implemented provided one has access to an oracle for performing weighted empirical risk minimization. See Section D for more details. Next we present a generalization analysis of our proposed algorithm in the idealized setting and show that it achieves near optimal guarantees.

4. Generalization Analysis

In this section we will analyze Algorithm 1 in the idealized scenario as described in Setting 1. We will show in later sections that much of the analysis carries over in the more realistic finite sample setting where each dense group has a large number of labeled examples and each sparse group has a large number of unlabeled examples. Our main theorem is stated below for symmetric loss functions, i.e., $\ell(y_1, y_2) = \ell(y_2, y_1)$, that satisfy triangle inequality, i.e., $\ell(y_1, y_2) \le \ell(y_1, y_3) + \ell(y_3, y_2)$ for all $y_1, y_2, y_3 \in \mathcal{Y}$. The following result is proved in Section B.1.

Theorem 2 (Idealized Setting and General Function Class) *Suppose Assumption 1 holds with the parameter Δ. For a group g, let $L^*_g := \min_{f \in \mathcal{F}} \mathsf{Loss}(f, P_g)$. If the loss function ℓ satisfies symmetry and triangle inequality, then the output of Algorithm 1 in the idealized setting (Setting 1) satisfies:*

$$\max_{g \in \mathcal{G}} \mathsf{Loss}(\widehat{f}, P_g) \le \mathsf{OPT}_\mathsf{L} + 2\Delta + 2 \max_g L^*_g \le 3\mathsf{OPT}_\mathsf{L} + 2\Delta \,, \tag{2}$$

$$\max_{g \in \mathcal{G}} \mathsf{Regret}(\widehat{f}, P_g) \le \mathsf{OPT}_\mathsf{R} + 2\Delta + 2 \max_g L^*_g. \tag{3}$$

Remark 3 *We note that our analysis carries over even if the triangle inequality is satisfied up to constants, for example, for squared loss. We refer the reader to Section B.1 for more details.*

4.1. Lower Bounds

Note that the guarantee in Theorem 2 incurs an additive error both in terms of the closeness parameter Δ, and the largest in-group optimal error of any group, i.e., $\max_g L_g^*$. We show that a dependence on both these terms is unavoidable; thus the guarantee of Theorem 2 is essentially the best possible even down to constants. We first present a lower bound below, proved in Section B.2, that shows that the dependence on the maximum L_g^* error is unavoidable, even when Δ is zero.

Theorem 4 *Let $\widehat{f}$ be the output of any algorithm for the idealized setting and the zero-one loss. Further, suppose that $\Delta = 0$ and let $\mathsf{OPT}_\mathsf{R} = \epsilon$ for any $\epsilon \in (0, 0.25)$. There is a choice of distributions such that $\max_g L_g^* = 0.25$, $\mathsf{OPT}_\mathsf{L} = 0.25$ but, with probability at least 0.5, the following two hold: (i) $\max_g \mathsf{Regret}(\widehat{f}, g) \geq 0.5$ and (ii) $\max_g \mathsf{Loss}(\widehat{f}, g) \geq 0.75$. In particular, the term $2\max_g L_g^*$ can not be avoided in Theorem 2 when $\max_g L_g^* = 0.25$ even when $\Delta = 0$.*

Note that the above lower bound constructs instances where Δ is zero. We prove in Section B.3, that in general, the dependence on Δ is also unavoidable.

Theorem 5 *For any value of Δ and any algorithm outputting $\widehat{f}$, there is a learning setup in the idealized setting, where Assumption 1 holds and $\max_{g\in\mathcal{G}} \mathsf{Loss}(\widehat{f}, P_g) = \Omega(\Delta)$.*

4.2. Square Loss and Convex Function Class

In the previous section, we showed that for general function classes the dependence on the maximum prediction error, i.e., $\max_g L_g^*$, is inevitable even if $\mathsf{OPT}_\mathsf{R} = 0$ and $\Delta \to 0$. Here we consider a practically relevant case when the loss function is the squared loss and the function class $\mathcal{F}$ is convex. We show an improved upper bound in this case below (proof given in Section B.4).

Assumption 2 *Let $\ell(y, y') := (y - y')^2$ and $\mathcal{F} : \mathcal{X} \to \mathcal{Y}$ is convex in the function space.*

Theorem 6 *Suppose Assumption 1 holds with the parameter Δ. For a group g, let $L_g^* := \min_{f\in\mathcal{F}} \mathsf{Loss}(f, P_g)$. Suppose the loss function is the squared loss and $\mathcal{F}$ is convex, i.e., Assumption 2 holds. Then, the output of Algorithm 1 in the idealized setting (Setting 1) satisfies*

$$\max_{g\in\mathcal{G}} \mathsf{Regret}(\widehat{f}, P_g) = O\Big(\mathsf{OPT}_\mathsf{R} + \Delta + \max_g \sqrt{L_g^*(\mathsf{OPT}_\mathsf{R} + \Delta)}\Big). \tag{4}$$

Furthermore, the upper bound can be tightened to $O(\mathsf{OPT}_\mathsf{R} + \Delta)$ if, for all sparse groups g, f_g^* lies in the relative interior of $\mathcal{F}$.

If the loss function happens to be bounded, for example, say due to bounded predictors and bounded labels, then the upper bound in Theorem 6 becomes $O\left(\sqrt{\mathsf{OPT}_\mathsf{R} + \Delta}\right)$, which is completely independent of $\max_g L_g^*$ altogether.

5. Finite Sample Analysis and Data-Dependent Choices

In this section, we extend our algorithm and analysis to the more realistic setting, where one only has access to finitely many labeled and unlabeled examples from each group. In particular, the algorithm *does not know* which dense group is closest to a particular sparse group, and the algorithm needs to make data-dependent choices. Thus, for each group, the algorithm must decide in a data dependent manner whether to perform pseudo-labeling for that particular group or not; if yes, then the algorithm also needs to choose the corresponding most informative dense group. Consequently, the algorithm needs to trade-off the effect of data-sparsity with the approximation error incurred by pseudo-labeling. Moreover, there is no clean distinction between what constitutes a sparse group versus a dense group; any hard threshold (for example, a decision based on the comparison between the number of labeled samples and the VC dimension) may be pessimistic. Thus, we would like a data-dependent way of making this decision. In order to analyze this setting we first begin by introducing additional useful notation.

5.1. Notations and Setup

We capture the dependence on the number of labeled and unlabeled samples using the parameters γ_g and τ_g below that can be bounded by standard complexity measures.

Assumption 3 (Uniform Convergence) *For every $\delta > 0$, there are parameters $\gamma_g, \tau_g \in [0, 1]$ such that with probability $1 - \delta$, for all $f, f' \in \mathcal{F}$ and $g \in \mathcal{G}$, we have that*

$$\Big| \mathop{\mathbf{E}}_{(x,y)\sim S_g} [\ell(f(x), y)] - \mathop{\mathbf{E}}_{(x,y)\sim P_g} [\ell(f(x), y)] \Big| \le \gamma_g \tag{5}$$

$$\left| \mathop{\mathbf{E}}_{x\sim U_g} [\ell(f(x), f'(x))] - \mathop{\mathbf{E}}_{x\sim P_g} \Big[\ell(f(x), f'(x))\Big] \right| \le \tau_g. \tag{6}$$

Observe that the first condition is equal to $|\mathsf{Loss}(f, S_g) - \mathsf{Loss}(f, P_g)| \le \gamma_g$. If $\mathcal{F}$ has VC dimension d and the loss function is zero-one loss, then Assumption 3 holds with parameter values $\gamma_g = O\big(\sqrt{\frac{d \log(|\mathcal{G}|/\delta)}{n_{g,s}}}\big)$ and $\tau_g = O\big(\sqrt{\frac{d \log(|\mathcal{G}|/\delta)}{n_{g,u}}}\big)$, where $n_{g,s}$ and $n_{g,u}$ are the number of labeled and unlabeled samples in group g, respectively (Mohri et al., 2018).

Next we strengthen Assumption 1. Since we have access to finitely many samples from a dense group g, we can hope to only approximate the optimal classifier f_g^*. Hence, we need to impose that all near-optimal estimators of a dense group perform well on the sparse group that it is close to. We formally define the set of nearly optimal estimators below, followed by the stronger assumption:

Definition 7 (Set of nearly optimal estimators) *Let $\mathcal{H}(\epsilon, g)$ be the set of functions whose regret on the group g is small, i.e., $\mathcal{H}(\epsilon, g) = \{f \in \mathcal{F} : \mathsf{Regret}(f, g) \le \epsilon\}$.*

Assumption 4 (Pairwise Similarity between Groups) *We say a (dense) group g' is informative for a (sparse) group g with parameter $\Delta_g(g', \epsilon)$ if $\forall f \in \mathcal{H}(\epsilon, g')$, we have $\mathsf{Regret}(f, g) \le \Delta_g(g', \epsilon)$.*

If a group is sparse, we will assume that there exists a dense group g' with small $\Delta_g(g', \epsilon)$. Note that under Assumption 3, the ERM on group g returns a $\widetilde{f}_g \in \mathcal{H}(2\gamma_g, g)$ by Equation (5).

Remark 8 (Weakening Assumption 4) *A milder assumption would be that there exists an f in $\mathcal{H}(\epsilon, g')$ that performs well for g. However, one would ultimately need to identify this f among $\mathcal{H}(\epsilon, g')$ using (limited) labeled data for the sparse group g, which requires the number of samples in g to scale with the statistical complexity of the function class $\mathcal{H}(\epsilon, g')$.*

While our analysis can be easily modified to work with the weaker assumption above, for improved readability we present the main results of this section under Assumption 4.

Finally, we remind the reader that our goal is to extend Theorem 2 in the finite-sample regime when sparse groups have only $\log(|\mathcal{G}|)$ samples (as opposed to the statistical complexity of $\mathcal{F}$).

5.2. Adapting to Δ

Recall that there are two challenges we need to address: First, for any group g, we do not know what the approximation error Δ will be if we perform pseudo-labeling with a different group. Second, we do not know which dense group to choose for a particular sparse group. Suppose for now that we know the closest group for each group (which we will outline shortly), which handles the second challenge. Then to tackle the first challenge, we set each group's objective to be a convex combination of the empirical regret $\mathsf{Regret}_{\mathsf{emp}}(f, g)$ and the proxy regret $\mathsf{Regret}_{\mathsf{proxy}}(f, g)$ with weights α_g and $(1 - \alpha_g)$, respectively for some (data-dependent) $\alpha_g \in [0, 1]$. Setting α_g to be 1 for dense groups and 0 for sparse groups recovers the original objective from Algorithm 1. This leads to the following algorithm (Algorithm 2):

Algorithm 2 Algorithmic Framework for Group DRO: Realistic Setting

```
Require: Groups $\mathcal{G}$, function class $\mathcal{F}$, loss function $\ell$, labeled data $S$, unlabeled data $U$, parameters
    $\gamma_g$ and $\tau_g$ from Assumption 3, a routine to choose $\widehat{f}_g$ and $\alpha_g$ in Lines 6 and 7 below, respectively.
 1: for each group $g \in \mathcal{G}$ do
 2:     Let $\widetilde{f}_g \leftarrow \mathrm{argmin}_{f \in \mathcal{F}} \mathsf{Loss}(f, S_g)$.
 3:     Define empirical regret: for all $f \in \mathcal{F}$, $\mathsf{Regret}_{\mathsf{emp}}(f, g) := \mathsf{Loss}(f, S_g) - \mathsf{Loss}(\widetilde{f}_g, S_g)$.
 4: end for
 5: for each group $g \in \mathcal{G}$ do
 6:     Choose the pseudo-labeling function $\widehat{f}_g$ using the labeled data $S_g$ and $\{\widetilde{f}_{g'} : g' \neq g\}$.
                                                                              ▷ E.g., using Algorithm 3
 7:     Choose $\alpha_g \in [0, 1]$ using the labeled data $S_g$.                      ▷ E.g., using Algorithm 3
 8:     Create proxy data from unlabeled data $S^{\text{proxy}} := \{(x, \widehat{f}_g(x), g) : (x, g) \in U\}$.
 9:     Define proxy regret using unlabeled data: $\forall f \in \mathcal{F}$, $\mathsf{Regret}_{\mathsf{proxy}}(f, g) := \mathsf{Loss}(f, S_g^{\text{proxy}})$.
10: end for
11: Solve the following min-max problem:
            $\widehat{f} = \mathrm{argmin}_{f \in \mathcal{F}} \max_{g \in \mathcal{G}} \left(\alpha_g \mathsf{Regret}_{\mathsf{emp}}(f, g) + (1 - \alpha_g)\mathsf{Regret}_{\mathsf{proxy}}(f, g)\right)$.
12: return $\widehat{f}$
```

Note, we have presented the algorithm in a general manner without specific data dependent choices of α_g and the choice of the closest dense groups. Of course, these choices are critical for the theoretical performance of the the algorithm above. The following technical lemma guides our choices in the final algorithm thereby leading to formal guarantees on the algorithm's performance (Algorithm 3).

Lemma 9 (Generalization Guarantee of Algorithm 2) *Make Assumption 3 and assume that the loss function satisfies symmetry and triangle inequality. In Algorithm 2, for each group g, let $\widetilde{f}_g$ be the ERM (defined in Line 2). Let $\widehat{f}$ be the resulting output in Algorithm 2 for arbitrary (data-dependent) choices of $\widehat{f}_g$ and α_g. Define $\beta := \max_g \left(\alpha_g \big(2\gamma_g + L_g^*\big) + (1 - \alpha_g)\big(\mathsf{Loss}\left(\widehat{f}_g, P_g\right) + \tau_g\big)\right)$*

and $\beta' := \max_g \left(\alpha_g \cdot 2\gamma_g + (1-\alpha_g)\big(\mathsf{Regret}\left(\widehat{f}_g, P_g\right) + \tau_g\big)\right)$*, which is smaller than* β*. Then with probability* $1-\delta$*,*

$$\max_{g \in \mathcal{G}} \mathsf{Loss}\big(\widehat{f}, P_g\big) \leq \mathsf{OPT}_{\mathsf{L}} + 2\beta \quad \textit{and} \quad \max_{g \in \mathcal{G}} \mathsf{Regret}\big(\widehat{f}, P_g\big) \leq \mathsf{OPT}_{\mathsf{R}} + \beta + \beta' \,. \tag{7}$$

Observe that the both sides of the inequalities above are random variables (since $\widetilde{f}_g$ and α_g are data-dependent). The above lemma suggests that to get small error, we should choose (i) the pseudo-labeling function $\widehat{f}_g$ whose loss is small and (ii) the convex weights α_g such that $\alpha_g = 1$ if $2\gamma_g + L_g^* < \mathsf{Loss}\left(\widehat{f}_g, P_g\right) + \tau_g$ and 0 otherwise. However, these values are unknown to the algorithm and must be estimated from the data. In order to tackle (i), we make use of Assumption 4, which imposes that either γ_g is small (i.e., the group is dense) or there exists a $g' \in \mathcal{G} \setminus \{g\}$ whose ERM predictor $\widetilde{f}_{g'}$ performs well on g. Thus, we choose $\widehat{f}_g$ to be $\widetilde{f}_{g'}$ that attains the smallest empirical loss on group g. Although this choice is data-dependent, we know that $\widetilde{f}_{g'}$ was trained on separate data. Thus, a Chernoff-style argument (stated below) implies that the probability of failure is small if each group g has more than $\log(|\mathcal{G}|)$ many samples, which is independent of the complexity of the hypothesis class $\mathcal{F}$.

Lemma 10 *Assume that uniform convergence holds (Assumption 3) and the loss function is bounded in* $[0,1]$*. There exists a constant* $c > 0$ *such that if each group has more than* $c \cdot \log(|\mathcal{G}|/\delta)/\epsilon^2$ *many labeled samples for some* $\epsilon, \delta \in (0,1)$*. Then with probability* $1 - 2\delta$*, for all groups g, the choice of* $\widehat{f}_g$ *in Line 2 of Algorithm 3 satisfies that* $\mathsf{Loss}\left(\widehat{f}_g, P_g\right) \leq L_g^* + 2\epsilon + \min_{g' \neq g}\left(\Delta_g(g', 2\gamma_{g'})\right)$ *and* $\left|\mathsf{Loss}\left(\widehat{f}_g, P_g\right) - \mathsf{Loss}\left(\widehat{f}_g, S_g\right)\right| \leq \epsilon$*.*

Algorithm 3 Algorithm to choose $\widehat{f}_g$ and α_g

```
Require: Group $g$, ERM predictors for all groups $\{\widetilde{f}_{g'} : g' \in \mathcal{G}\}$, labeled data $S_g$, the parameters
    $\gamma_g$ and $\tau_g$ from Assumption 3, the parameter $\epsilon$ from Lemma 10.
1: Set $\widehat{g} \leftarrow \mathsf{argmin}_{g' : g' \neq g} \mathsf{Loss}(\widetilde{f}_g, S_g)$.
2: Let $\widehat{f}_g \leftarrow \widetilde{f}_{\widehat{g}}$ and let $\widehat{L}_g$ be the corresponding minimum loss, $\widehat{L}_g \leftarrow \mathsf{Loss}(\widehat{f}_g, S_g)$.
3: Set $\alpha_g \leftarrow 1$ if $3\gamma_g + \mathsf{Loss}(\widetilde{f}_g) < \widehat{L}_g + \epsilon + \tau_g$ else 0.
```

Next we address the data-dependent choice of α_g (the point (ii) in the paragraph preceding Lemma 10). Here, we need upper-bounds on (i) L_g^*, which can be obtained using the empirical loss of $\widetilde{f}_g$ (up to γ_g factor), and (ii) $\mathsf{Loss}(\widehat{f}_g, P_g)$, which was calculated in the lemma above. Combining everything together, the parameter β in Lemma 9 can be upper bounded by $L_g^* + O\left(\max_g \min\left(\gamma_g, \tau_g + \epsilon + \min_{g' \neq g} \Delta_g(g', 2\gamma_{g'})\right)\right)$, obtaining the following:

Theorem 11 *Assume Assumption 3 and the loss function is symmetric, satisfies triangle inequality, and is bounded in* $[0,1]$*. Suppose each group has* $\Omega(\log(|\mathcal{G}|/\delta)/\epsilon^2)$ *many labeled samples for some* $\epsilon, \delta \in (0,1)$*. Then with probability* $1 - 2\delta$*, the output of Algorithm 2 satisfies (recall* $\Delta_g(\cdot,\cdot)$ *is defined in Assumption 4):*

$$\max_{g \in \mathcal{G}} \mathsf{Loss}\left(\widehat{f}, P_g\right) \leq \mathsf{OPT}_{\mathsf{L}} + 2\max_g L_g^* + \max_{g \in \mathcal{G}} \min\left(4\gamma_g, \min_{g' \neq g}\left(\Delta_g(g', 2\gamma_{g'}) + \tau_g + 4\epsilon\right)\right)$$

$$\max_{g \in \mathcal{G}} \mathsf{Regret}\left(\widehat{f}, P_g\right) \leq \mathsf{OPT}_{\mathsf{R}} + \max_g L_g^* + \max_{g \in \mathcal{G}} \min\left(4\gamma_g, \min_{g' \neq g}\left(\Delta_g(g', 2\gamma_{g'}) + \tau_g + 4\epsilon\right)\right).$$

Here, for each group g, the dominant factors in the excess error (modulo L_g^*, which is unavoidable) are: (i) γ_g, a parameter that depends on the number of labeled samples in a group and the complexity of $\mathcal{F}$ and (ii) Δ, the distance from the nearest dense group as mentioned in Assumption 4. In particular, the last term in Theorem 11, is a maximum over $g \in \mathcal{G}$ of $\min\left(4\gamma_g, \min_{g' \neq g}\left(\Delta_g(g', 2\gamma_{g'}) + \tau_g + 4\epsilon\right)\right)$. For every g, this expression is small if either g is dense (since γ_g is small) or if it is close to a dense group g' and g has many unlabeled samples. Thus, Theorem 11 extends Theorem 2 to the finite sample setting assuming only $\log(|\mathcal{G}|/\delta)/\epsilon^2$ samples per sparse group.

We now discuss the benefits and tightness of Theorem 11: First, a sparse group must have $\operatorname{poly}(\log(|\mathcal{G}), 1/\epsilon)$ many samples to get comparable error in general, matching the sample complexity of Theorem 11 (cf. Section C.4). Second, akin to Theorem 4, an additive dependence on $\Omega(\max_g L_g^*)$ can not be avoided with finitely many samples in sparse groups even when dense groups have infinitely many samples (cf. Theorem 18). Third, we note that the proposed algorithm automatically adapts to the unknown values of $\Delta_g(\cdot,\cdot)$. Finally, the proposed algorithm is computationally-efficient because it can be implemented using weighted-ERM oracle, a standard assumption in this field (cf. Section D).

Implications for group fairness. Note that Theorem 2 and Theorem 11 directly provide concrete guarantees for the min-max (Rawlsian) notion of fairness (Abernethy et al., 2022). As different fairness objectives are often incompatible, these results have no implications on other notions of fairness such as equalized odds or demographic parity (Hardt et al., 2016).

6. Conclusion and Future Work

In this work, we initiate a systematic study of the semi-supervised group DRO in the data-scarce regime. We proposed computationally-efficient and (labeled) sample-efficient algorithms to provably tackle data scarcity under very mild assumptions on the relationships between groups. Starting from an idealized setting, we derive a near-optimal computationally-efficient algorithm, which we then generalize to a more realistic finite-sample setting. An important direction for future work is to perform extensive experiments on real-world data and use the derived insights to modify the notion of similarity and the algorithms. Another important venue is to tackle the losses that do not satisfy (approximate) triangle inequality and symmetry.

References

J. D. Abernethy, P. Awasthi, M. Kleindessner, J. Morgenstern, C. Russell, and J. Zhang. Active sampling for min-max fairness. In *Proc. 39th International Conference on Machine Learning (ICML)*, 2022.

A. Agarwal and T. Zhang. Minimax regret optimization for robust machine learning under distribution shift. In *Proc. 35th Annual Conference on Learning Theory (COLT)*, 2022.

A. Agarwal, A. Beygelzimer, M. Dudik, J. Langford, and H. Wallach. A reductions approach to fair classification. In *Proc. 35th International Conference on Machine Learning (ICML)*, 2018.

J. Angwin, J. Larson, S. Mattu, and L. Kirchner. Machine bias: There's software used across the country to predict future criminals. and it's biased against blacks. 2016. *URL https://www. propublica. org/article/machine-bias-risk-assessments-in-criminal-sentencing*, 2019.

I. Attias, A. Kontorovich, and Y. Mansour. Improved generalization bounds for adversarially robust learning. *Journal of Machine Learning Research*, 23(175):1–31, 2022.

J. Baxter. A model of inductive bias learning. *Journal of artificial intelligence research*, 12:149–198, 2000.

S. Ben-David, J. Blitzer, K. Crammer, A. Kulesza, F. Pereira, and J. W. Vaughan. A theory of learning from different domains. *Machine learning*, 79:151–175, 2010.

A. Ben-Tal, L. El Ghaoui, and A. Nemirovski. *Robust optimization*. Princeton University Press, 2009.

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

A. Blum, N. Haghtalab, A. D. Procaccia, and M. Qiao. Collaborative pac learning. In *Advances in Neural Information Processing Systems 30 (NeurIPS)*, volume 30, 2017.

J. Buolamwini and T. Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Proc. of the Conference on fairness, accountability and transparency*, 2018.

G. Cavallanti, N. Cesa-Bianchi, and C. Gentile. Linear algorithms for online multitask classification. *The Journal of Machine Learning Research*, 11:2901–2934, 2010.

R. S. Chen, B. Lucier, Y. Singer, and V. Syrgkanis. Robust optimization for non-convex objectives. In *Advances in Neural Information Processing Systems 30 (NeurIPS)*, 2017.

C. Cortes, M. Mohri, and A. Muñoz Medina. Adaptation algorithm and theory based on generalized discrepancy. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 169–178, 2015.

C. Cortes, M. Mohri, J. Gonzalvo, and D. Storcheus. Agnostic learning with multiple objectives. In *Advances in Neural Information Processing Systems 33 (NeurIPS)*, 2020.

A. Cotter, H. Jiang, and Karthik S. Two-player games for efficient non-convex constrained optimization. In *Algorithmic Learning Theory*, pages 300–332. PMLR, 2019.

M. Deodhar, X. Ma, Y. Cai, A. Koes, A. Beutel, and J. Chen. A human-ml collaboration framework for improving video content reviews. *arXiv preprint arXiv:2210.09500*, 2022.

E. Diana, W. Gill, M. Kearns, K. Kenthapadi, and A. Roth. Minimax group fairness: Algorithms and experiments. In *Proc. 2021 AAAI/ACM Conference on AI, Ethics, and Society (AIES)*, 2021.

J. Duchi, P. Glynn, and H. Namkoong. Statistics of robust optimization: A generalized empirical likelihood approach. *Mathematics of Operations Research*, 2021.

U. Feige, Y. Mansour, and R. Schapire. Learning and inference in the presence of corrupted inputs. In *Proc. 28th Annual Conference on Learning Theory (COLT)*, 2015.

R. Gao, X. Chen, and A. J. Kleywegt. Wasserstein distributionally robust optimization and variation regularization. *Operations Research*, 2022.

N. Haghtalab, M. I. Jordan, and E. Zhao. On-demand sampling: Learning optimally from multiple distributions. In *Advances in Neural Information Processing Systems 35 (NeurIPS)*, 2022.

M. Hardt, E. Price, and N. Srebro. Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29, 2016.

M. J. Kearns, S. Neel, A. Roth, and Z. S. Wu. Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. In *Proc. 35th International Conference on Machine Learning (ICML)*, 2018.

M. J. Kearns, S. Neel, A. Roth, and Z. S. Wu. An empirical study of rich subgroup fairness for machine learning. In *Proc. of the Conference on Fairness, Accountability, and Transparency, FAT**, 2019.

D. Kuhn, P. M. Esfahani, V. A. Nguyen, and S. Shafieezadeh-Abadeh. Wasserstein distributionally robust optimization: Theory and applications in machine learning. In *Operations Research & Management Science in the Age of Analytics*, pages 130–166. INFORMS, 2019.

D. Levy, Y. Carmon, J. C. Duchi, and A. Sidford. Large-scale methods for distributionally robust optimization. In *Advances in Neural Information Processing Systems 33 (NeurIPS)*, 2020.

K. Lounici, M. Pontil, S. Van De Geer, and A. B. Tsybakov. Oracle inequalities and optimal inference under group sparsity. 2011.

A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu. Towards deep learning models resistant to adversarial attacks. In *Proc. 6th International Conference on Learning Representations (ICLR)*, 2018.

N. Martinez, M. Bertran, and G. Sapiro. Minimax pareto fairness: A multi objective perspective. In *Proc. 37th International Conference on Machine Learning (ICML)*, 2020.

A. Maurer, M. Pontil, and B. Romera-Paredes. The benefit of multitask representation learning. *Journal of Machine Learning Research*, 17(81):1–32, 2016.

M. Mohri, A. Rostamizadeh, and A. Talwalkar. *Foundations of machine learning*. MIT press, 2018.

M. Mohri, G. Sivek, and A. T. Suresh. Agnostic federated learning. In *Proc. 36th International Conference on Machine Learning (ICML)*, 2019.

O. Montasser, S. Hanneke, and N. Srebro. VC classes are adversarially robustly learnable, but only improperly. In *Proc. 32nd Annual Conference on Learning Theory (COLT)*, 2019.

H. Namkoong and J. C. Duchi. Stochastic gradient methods for distributionally robust optimization with f-divergences. In *Advances in Neural Information Processing Systems 29 (NeurIPS)*, 2016.

S. Sagawa, P. W. Koh, T. B. Hashimoto, and P. Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. In *Proc. 8th International Conference on Learning Representations (ICLR)*, 2019.

H. Salman, J. Li, I. Razenshteyn, P. Zhang, H. Zhang, S. Bubeck, and G. Yang. Provably robust deep learning via adversarially trained smoothed classifiers. In *Advances in Neural Information Processing Systems 32 (NeurIPS)*, 2019.

S. Shalev-Shwartz and S. Ben-David. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, 2014. ISBN 978-1-107-29801-9.

S. Shekhar, G. Fields, M. Ghavamzadeh, and T. Javidi. Adaptive sampling for minimax fair classification. In *Advances in Neural Information Processing Systems 34 (NeurIPS)*, 2021.

A. Sinha, H. Namkoong, and J. C. Duchi. Certifying some distributional robustness with principled adversarial training. In *Proc. 6th International Conference on Learning Representations (ICLR)*, 2018.

C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. J. Goodfellow, and R. Fergus. Intriguing properties of neural networks. In *Proc. 2nd International Conference on Learning Representations (ICLR)*, 2014.

R. Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge University Press, 2018.

Y. Zhang, M. Long, J. Wang, and M. I. Jordan. On localized discrepancy for domain adaptation. *arXiv preprint arXiv:2008.06242*, 2020.

Appendix A. Failure of Standard Approaches

In this section, we provide additional details regarding Example 1.

Example 1 *Let $\mathcal{G} = \{1, 2, 3\}$. Let u be an unknown unit vector in $\mathbb{R}^d$ and let z be an unknown sign in $\{+1, -1\}$. The distribution of each group is given below:*

- *For Group 1, (x, y) is distributed as $x \sim \mathcal{N}(0, I)$ and $y = u^\top x$,*
- *For Group 2, (x, y) is distributed as $x \sim \mathcal{N}(0, I)$ and $y = -u^\top x$,*
- *For Group 3, (x, y) is distributed as $x \sim \mathcal{N}(0, \sigma^2 I)$ and $y = zu^\top x$ for some $\sigma^2 \gg 1$.*

We are interested in finding a min-max optimal linear predictor $\beta \in \mathbb{R}^d$, the function class $\mathcal{F}$ is affine, and the loss function is the square loss. Let $\epsilon \in (0, 1)$ be small enough. We assume that for groups 1 and 2 are dense and have $(1 - \epsilon)/2$ fraction of all training examples each, with the remaining ϵ fraction in group 3, which is a sparse group. By simple calculations, we have that $\mathsf{OPT}_\mathsf{L} = \frac{4\sigma^2}{(\sigma+1)^2} \leq 4$.

Calculating the OPT Let β be a candidate vector in $\mathbb{R}^d$. For a group $g \in \{1, 2, 3\}$, recall that $\mathsf{Loss}(\beta, g)$ is equal to $\mathbf{E}[(x^\top \beta - y)^2]$, where (x, y) are distributed as per the group distribution. Thus, $\mathsf{Loss}(\beta, 1) = \|\beta - u\|_2^2$, $\mathsf{Loss}(\beta, 2) = \|\beta + u\|_2^2$, and $\mathsf{Loss}(\beta, 3) = \sigma^2\|\beta - zu\|_2^2$. In particular, it can be seen that $\mathsf{OPT}_\mathsf{L} \leq 4$ as follows: the loss of $\widehat{\beta} = zu$ satisfies that $\max_g \mathsf{Loss}(\widehat{\beta}, g) = \max_g \big((z-1)^2, (z+1)^2, 0\big) \leq 4$ since $z \in \{-1, 1\}$, implying that $\mathsf{OPT}_\mathsf{L} \leq 4$. In fact, the exact optimum can be calculated as follows: let $z = 1$ for simplicity; the case of $z = -1$ is analogous. The maximum loss can then be written as $\max(\sigma^2\|\beta - u\|_2^2, \|\beta - u\|_2^2, \|\beta + u\|_2^2)$. To calculate the minimum over β, we equate the two expressions (and checking that it is indeed the minimum value), we obtain that the $\mathsf{OPT}_\mathsf{L} = \frac{4\sigma^2}{(\sigma+1)^2}$, which increases from 1 to 4 as σ goes from 1 to ∞.

Average Empirical Risk Minimizer Let us first consider the average empirical risk minimizer. Let P be the mixture distribution of these three groups with the given weights. As the loss function is square loss, as $n \to \infty$, the solution converges to $\beta_{\mathrm{OLS}} = \big(\mathbf{E}_P[xx^\top]\big)^{-1}\big(\mathbf{E}_{(x,y)\sim P}[xy]\big)$. It is easy to see that $\mathbf{E}_P[xx^\top] = 0.5\,(1-\epsilon)\,I + 0.5\,(1-\epsilon)\,I + \epsilon\sigma^2 I = (1 + \epsilon\,(\sigma^2 - 1))I$. Turning to $\mathbf{E}_{(x,y)\sim P}[xy]$, we have that it is equal to

$$\begin{aligned}\mathop{\mathbf{E}}_{(x,y)\sim P}[xy] &= 0.5\,(1-\epsilon)\mathop{\mathbf{E}}_{x\sim\mathcal{N}(0,I)}[(u^\top x)x] - 0.5\,(1-\epsilon)\mathop{\mathbf{E}}_{x\sim\mathcal{N}(0,I)}[(u^\top x)x] + \epsilon\mathop{\mathbf{E}}_{x\sim\mathcal{N}(0,\sigma^2 I)}[(zu^\top x)x] \\ &= \epsilon\mathop{\mathbf{E}}_{x\sim\mathcal{N}(0,\sigma^2 I)}[(zu^\top x)x] = \epsilon\sigma^2 zu\,.\end{aligned}$$

Thus, β_{OLS} converges to $\left(\frac{\epsilon\sigma^2 z}{1+\epsilon\sigma^2-\epsilon}\right)u$. The loss of β_{OLS} on the third group is equal to

$$\begin{aligned}\mathop{\mathbf{E}}_{x\sim\mathcal{N}(0,\sigma^2 I)}[(x^\top\beta_{\mathrm{OLS}} - zu^\top x)^2] = \sigma^2\|zu - \beta_{\mathrm{OLS}}\|_2^2 &= \sigma^2\left(1 - \frac{\epsilon\sigma^2}{1+\epsilon\sigma^2-\epsilon}\right)^2 \\ &= \frac{\sigma^2(1-\epsilon)^2}{(1+\epsilon\sigma^2-\epsilon)^2},.\end{aligned}$$

If $\sigma^2\frac{1-\epsilon}{\epsilon}$, then the expression above can be larger than $\sigma^2/4$, which could be much larger than OPT_L if σ is large (and ϵ is small).

Maximum Empirical Loss Minimizer Let $\Sigma_1, \Sigma_2, \Sigma_3$ be the empirical second moment matrices of the three groups, respectively. Then the solution corresponds to $\widehat{\beta}$ that minimizes

$$\widehat{\beta} = \arg\min_{\beta} \max \Big((\beta - u)^\top \Sigma_1 (\beta - u), (\beta + u)^\top \Sigma_2 (\beta + u), (\beta - zu)^\top \Sigma_3 (\beta - zu) \Big). \quad (8)$$

Let us consider the idealistic setting where the first two groups have infinite samples since the main technical roadblock is handling the limited data in the third group. In this idealized setting, we have $\Sigma_1 = I$ and $\Sigma_2 = I$, and the output is thus distributed as follows:

$$\widehat{\beta} = \arg\min_{\beta} \max \Big(\|\beta - u\|_2^2, \|\beta + u\|_2^2, (\beta - zu)^\top \Sigma_3 (\beta - zu) \Big). \quad (9)$$

If the third group has $r \leq d$ samples, which is the regime of interest, then with probability 1, Σ_3 will have rank r. Let U_{d-r} be the null space of Σ_3. By rotational invariance of $\mathcal{N}(0, \sigma^2 I)$, U_{d-r} is distributed uniformly among subspaces of rank $d - r$. We let $P_{U_{n-r}}$ with be projection matrix on the subspace U_{n-r}.

We will now show that for any $r = o(d)$ and $\epsilon > 0$, with high probability $(1 - O(\exp(-\Omega(d\epsilon^2))))$, there is a candidate vector $\widehat{\beta}$ whose empirical (maximum) loss is less than $1 + O(\epsilon + r/d)$ but its population-level loss is $\Omega(\sigma^2)$ (recall that $\mathsf{OPT}_\mathsf{L} \to 4$ as $\sigma \to \infty$). In particular, the empirical loss of $\widehat{\beta}$ will be zero on the sparse group but the population loss of $\widehat{\beta}$ on the sparse group will scale linearly with σ^2.

Let $z = 1$ for simplicity; the case of $z = -1$ is analogous. Let v be the projection of u on U_{n-r}. With high probability, we have that $|u^\top v| = \|v\|_2^2$ is of the order $(1 \pm \epsilon)\frac{d-r}{d} = 1 \pm (\epsilon + r/d)$ (Vershynin, 2018, Chapter 5). Consider $\widehat{\beta} = u - v$. Using Equation (9), the maximum empirical loss of $\widehat{\beta}$ is $\max(\|v\|_2^2, \|2u - v\|_2^2, v^\top \Sigma_3 v) = \max(1, \|2u + v\|_2^2)$, where we use that v belongs to the null space of Σ_3. The second term in the expression is further equal to $\|2u + v\|_2^2 = 4\|u\|_2^2 + \|v\|_2^2 - 4(u^\top v) = 4 - 3\|v\|_2^2$. We can upper bound it as follows: $4 - 3(1 - \epsilon)(1 - r/d) \leq 1 + O(r/d) + \epsilon$. This establishes the claim regarding the maximum empirical loss of this candidate vector. However, the maximum population-level loss of this estimate over the three groups is at least $\sigma^2 \|v\|_2^2$ (achieved by the third group), which is indeed $\Omega(\sigma^2)$ with high probability.

Appendix B. Proofs from Section 4

B.1. Proof of Theorem 2 under Approximate Triangle Inequality

We begin by defining the approximate triangle inequality:

Definition 12 (Approximate Triangle inequality) *We say a loss function $\ell(\cdot, \cdot) : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_+$ satisfies c-approximate triangle inequality for $c \geq 1$ if for all y_1, y_2, y_3:*

$$\ell(y_1, y_2) \leq c \cdot \Big(\ell(y_1, y_3) + \max\left(\ell(y_3, y_2), \ell(y_2, y_3)\right) \Big).$$

The zero-one loss satisfies this condition with $c = 1$, while the square loss satisfies this with $c = 2$. The following theorem is thus a more general version of Theorem 2.

Theorem 13 (Idealized Setting and General Function Class for Approximate Triangle Inequality) *Suppose Assumption 1 holds with the parameter Δ. For a group g, let $L_g^* := \min_{f \in \mathcal{F}} \mathsf{Loss}(f, P_g)$.*

Suppose the loss function ℓ is symmetric and satisfies c-approximate triangle inequality (Definition 12), then the output of Algorithm 1 in the idealized setting (Theorem 1) satisfies:

$$\begin{aligned}\max_{g\in\mathcal{G}} \mathsf{Loss}(\widehat{f}, P_g) &\le c^2 \cdot \mathsf{OPT}_{\mathsf{L}} + c(c+1)\Delta + c(c+1)\max_g L_g^* \\ &\le c(2c+1)\mathsf{OPT}_{\mathsf{L}} + c(c+1)\Delta\,, \qquad \textit{and} \\ \max_{g\in\mathcal{G}} \mathsf{Regret}(\widehat{f}, P_g) &\le c^2\mathsf{OPT}_{\mathsf{R}} + c(c+1)\Delta + (2c^2+c-1)\max_g L_g^*.\end{aligned}$$

Proof We define $\widetilde{\mathsf{Loss}}(f, P_g)$ to be $\mathbf{E}_{x\sim P_g}[\ell(f(x), f^*_{\widehat{g}}(x))]$. By using the approximate triangle inequality, we obtain the following relation between Loss and $\widetilde{\mathsf{Loss}}$ for sparse groups: for all $f \in \mathcal{F}$, we have that

$$\begin{aligned}\mathsf{Loss}(f, P_g) - c\cdot\widetilde{\mathsf{Loss}}(f, P_g) &= \mathop{\mathbf{E}}_{(x,y)\sim P_g}[\ell(f(x), y)] - c \mathop{\mathbf{E}}_{(x,y)\sim P_g}[\ell(f(x), f^*_{\widehat{g}}(x))] \\ &\le c \mathop{\mathbf{E}}_{(x,y)\sim P_g}[\ell(f^*_{\widehat{g}}(x), y)] = c\cdot\mathsf{Loss}(f^*_{\widehat{g}}, P_g) \le c\Delta + cL_g^*, \qquad (10)\end{aligned}$$

where the last inequality uses Assumption 1. Similarly,

$$\begin{aligned}\widetilde{\mathsf{Loss}}(f, P_g) - c\cdot\mathsf{Loss}(f, P_g) &= \mathop{\mathbf{E}}_{(x,y)\sim P_g}[\ell(f(x), f^*_{\widehat{g}}(x))] - c \mathop{\mathbf{E}}_{(x,y)\sim P_g}[\ell(f(x), y)] \\ &\le c \mathop{\mathbf{E}}_{(x,y)\sim P_g}[\ell(y, f^*_{\widehat{g}}(x))] \\ &= c \mathop{\mathbf{E}}_{(x,y)\sim P_g}[\ell(f^*_{\widehat{g}}(x), y)] \le c\Delta + cL_g^*, \qquad (11)\end{aligned}$$

where the last equality uses the symmetry of the loss function.

Observe that the output of the algorithm, $\widehat{f}$, in the idealized setting can equivalently be defined as follows:

$$\widehat{f} = \operatorname*{argmin}_{f\in\mathcal{F}} \max\big(\max_{g:\text{dense}} \mathsf{Regret}(f, P_g), \max_{g:\text{sparse}} \widetilde{\mathsf{Loss}}(f, P_g)\big). \qquad (12)$$

We now proceed as follows to upper bound the maximum loss achieved by the output $\widehat{f}$ of the algorithm: let f^* be the function achieving the $\mathsf{OPT}_{\mathsf{L}}$, then

$$\begin{aligned}\max_{g\in\mathcal{G}} \mathsf{Loss}(\widehat{f}, P_g) &= \max\left(\max_{g:\text{dense}} \mathsf{Loss}(\widehat{f}, P_g), \max_{g:\text{sparse}} \mathsf{Loss}(\widehat{f}, P_g)\right) \\ &\le \max\left(\max_{g:\text{dense}} \mathsf{Regret}(\widehat{f}, P_g) + L_g^*, \max_{g:\text{sparse}} c\cdot\widetilde{\mathsf{Loss}}(\widehat{f}, P_g) + c\Delta + cL_g^*\right) \quad \text{(using (10))} \\ &\le c\max\left(\max_{g:\text{dense}} \mathsf{Regret}(\widehat{f}, P_g), \max_{g:\text{sparse}} \widetilde{\mathsf{Loss}}(\widehat{f}, P_g)\right) + c\Delta + c\max_g L_g^* \\ &\le c\max\left(\max_{g:\text{dense}} \mathsf{Regret}(f^*, P_g), \max_{g:\text{sparse}} \widetilde{\mathsf{Loss}}(f^*, P_g)\right) + c\Delta + c\max_g L_g^* \quad \text{(using (12))} \\ &\le c^2\max\left(\max_{g:\text{dense}} \mathsf{Loss}(f^*, P_g), \max_{g:\text{sparse}} \mathsf{Loss}(f^*, P_g) + \Delta + L_g^*\right) \\ &\qquad + c\Delta + c\max_g L_g^* \quad \text{(using (11))}\end{aligned}$$

$$
\begin{aligned}
&\le c^2 \max_{g\in\mathcal{G}} \mathsf{Loss}(f^*, g) + c(c+1)\Delta + c(c+1)\max_g L_g^* \\
&= c^2 \mathsf{OPT}_{\mathsf{L}} + c(c+1)\Delta + c(c+1)\max_g L_g^*.
\end{aligned}
$$

The final result follows by the fact that $\mathsf{OPT}_{\mathsf{L}} \ge \max_g L_g^*$.

Finally, we consider the maximum regret of $\widehat{f}$ on $\mathcal{G}$. Here, let f^* be the function achieving $\mathsf{OPT}_{\mathsf{R}}$.

$$
\begin{aligned}
\max_{g\in\mathcal{G}} \mathsf{Regret}(\widehat{f}, P_g) &= \max\left(\max_{g:\text{dense}} \mathsf{Regret}(\widehat{f}, P_g), \max_{g:\text{sparse}} \mathsf{Loss}(\widehat{f}, P_g) - L_g^*\right) \\
&\le \max\left(\max_{g:\text{dense}} \mathsf{Regret}(\widehat{f}, P_g), \max_{g:\text{sparse}} c\cdot \widetilde{\mathsf{Loss}}(\widehat{f}, P_g) + c\cdot\Delta + (c-1)L_g^*\right) \quad \text{(using (10))} \\
&\le c\cdot\max\left(\max_{g:\text{dense}} \mathsf{Regret}(\widehat{f}, P_g), \max_{g:\text{sparse}} \widetilde{\mathsf{Loss}}(\widehat{f}, P_g)\right) \\
&\qquad + c\cdot\Delta + \max_g (c-1)L_g^* \\
&\le c\cdot\max\left(\max_{g:\text{dense}} \mathsf{Regret}(f^*, P_g), \max_{g:\text{sparse}} \widetilde{\mathsf{Loss}}(f^*, P_g)\right) \\
&\qquad + c\cdot\Delta + \max_g (c-1)L_g^* \quad \text{(using (12))} \\
&\le c^2\cdot\max\left(\max_{g:\text{dense}} \mathsf{Regret}(f^*, P_g), \max_{g:\text{sparse}} \mathsf{Loss}(f^*, P_g) + \Delta + L_g^*\right) \\
&\qquad + c\cdot\Delta + (c-1)\max_g L_g^* \quad \text{(using (11))} \\
&\le c^2\max\left(\max_{g:\text{dense}} \mathsf{Regret}(f^*, P_g), \max_{g:\text{sparse}} \mathsf{Regret}(f^*, P_g) + \Delta + 2L_g^*\right) \\
&\qquad + c\cdot\Delta + (c-1)\max_g L_g^* \\
&\le c^2\cdot\max_{g\in\mathcal{G}} \mathsf{Regret}(f^*, g) + c(c+1)\Delta + (2c^2 + c - 1)\cdot\max_g L_g^* \\
&= c^2\cdot\mathsf{OPT}_{\mathsf{R}} + c(c+1)\Delta + (2c^2 + c - 1)\max_g L_g^*.
\end{aligned}
$$

■

B.2. Proof of Theorem 4

Theorem 4 *Let $\widehat{f}$ be the output of any algorithm for the idealized setting and the zero-one loss. Further, suppose that $\Delta = 0$ and let $\mathsf{OPT}_{\mathsf{R}} = \epsilon$ for any $\epsilon \in (0, 0.25)$. There is a choice of distributions such that $\max_g L_g^* = 0.25$, $\mathsf{OPT}_{\mathsf{L}} = 0.25$ but, with probability at least 0.5, the following two hold: (i) $\max_g \mathsf{Regret}(\widehat{f}, g) \ge 0.5$ and (ii) $\max_g \mathsf{Loss}(\widehat{f}, g) \ge 0.75$. In particular, the term $2\max_g L_g^*$ can not be avoided in Theorem 2 when $\max_g L_g^* = 0.25$ even when $\Delta = 0$.*

Proof Suppose there are 3 groups, $\mathcal{G} = \{g_1, g_2, g_3\}$, where the group g_3 is sparse, and the function class $\mathcal{F} = \{f_1, f_2, f_3, f_4\}$. Then, we will show that there exist two joint distributions P and Q (over the three groups) such that P and Q have the same distribution over (i) labeled samples on the dense groups and (ii) unlabeled samples on the sparse group. Moreover, both of the dense groups, g_1 and g_2, are valid neighbors of the sparse group, g_3, with $\Delta = 0$ in the sense of Assumption 1 for both choices of the joint distributions P and Q. Thus, no algorithm can differentiate between these cases in the idealized setting, i.e., without using labeled samples from the sparse group.

Furthermore, for the joint distribution P (respectively, Q), there is a single function $f \in \mathcal{F}$, f_3 (respectively, f_4), that has max regret equal to OPT_R, while all other functions in $\mathcal{F}$ have max regret at least 0.5; in fact, f_1 and f_2 have regret 0.75 on either P or Q. Thus, any algorithm when applied to P and Q must choose between f_3 and f_4. However, the performance of f_3 and f_4 is identical on the dense groups under both P and Q, and their behavior differ only on the sparse group. Since we do not observe labeled samples from the sparse group and the marginal distribution of features is identical on P and Q, we obtain that no algorithm can distinguish between P and Q with probability more than 0.5.

In particular, we will show that the functions in $\mathcal{F}$ have the following performance on P and Q:

Table 1: Functions and their expected loss values on different groups under two choices of conditional distributions on group g_3. The first column lists the functions in the function class $\mathcal{F}$. The second and third column list the (average) loss values of the functions on the groups g_1 and g_2 under both P and Q (since these groups are distributed identically under P and Q); Here ϵ is an arbitrarily small positive value in $(0, 0.25)$. Finally, the last two columns give the average loss values of the functions on the group g_3 under P and Q, respectively. Under both P and Q, it can be seen that $\mathsf{OPT}_\mathsf{L} = 0.25$ and $\mathsf{OPT}_\mathsf{R} = \epsilon$, while $\Delta = 0$: Under P (similarly, Q), the neighbor of g_3 is g_1 (g_2) because g_1's (g_2's) optimal classifier, f_1 (f_2), achieves zero regret on g_3.

	Groups			
	g_1 (Dense)	g_2 (dense)	g_3 (sparse)	
Functions			P	Q
f_1	0	1	0.25	0.25
f_2	1	0	0.25	0.25
f_3	ϵ	ϵ	0.25	0.75
f_4	ϵ	ϵ	0.75	0.25

Thus, the desired result follows if we exhibit a functional class and joint distributions P and Q that exhibit the properties in Table 1. Our function class and labels will be binary $\mathcal{Y} = \{0, 1\}$ and $\mathcal{F} \subset \mathcal{X} \to \{0, 1\}$. We assume that the support of g_1 and g_2 is disjoint from the group g_3 and thus the conditional distributions on these groups give no information about the group g_3, and it is easy to construct cases where these functions satisfy the values given in Table 1.

For the group g_3, we assume that the features are uniform on $\{x_1, x_2\}$ in both P and Q for some distinct $x_1, x_2 \in \mathcal{X}$. The only thing that differs between P and Q is the conditional distribution on the (binary) labels. Since the labels are binary, it suffices to define the probability of $y = 1$ for each x under these two distributions on the group g_3. The choice of parameters achieving the desired values is shown in Table 2:

Table 2: The distributions of g_3 under the distributions P and Q. Here, the distribution on $\{x_1, x_2\}$ is uniform under both P and Q. The conditional distribution of y given x is characterized by $\mathbf{E}[y|x]$. The left table also defines the function class $\mathcal{F}$ on $\{x_1, x_2\}$ (this definition is then repeated in the right table). The last columns on both of these tables confirm that these values match the ones given in Table 1.

g_3 under P

	Support		
Functions	x_1	x_2	Loss
$\mathbf{E}[y\|x]$	0.5	0	
f_1	0	0	0.25
f_2	1	1	0.25
f_3	0	1	0.75
f_4	1	0	0.25

g_3 under Q

	Support		
Functions	x_1	x_2	Loss
$\mathbf{E}[y\|x]$	0	0.5	
f_1	0	0	0.25
f_2	1	1	0.25
f_3	0	1	0.25
f_4	1	0	0.75

■

B.3. Proof of Theorem 5

Theorem 5 *For any value of Δ and any algorithm outputting $\widehat{f}$, there is a learning setup in the idealized setting, where Assumption 1 holds and* $\max_{g \in \mathcal{G}} \mathsf{Loss}(\widehat{f}, P_g) = \Omega(\Delta)$.

Proof Suppose there are two groups, i.e., $\mathcal{G} = \{g_1, g_2\}$, and the sparse group is g_2. Furthermore, consider the simple setting of univariate linear regression with square loss, i.e., $\mathcal{F} = \{\beta : \beta \in \mathbb{R}\}$ and $\ell(y, y') = (y - y')^2$. Let the distribution of the group g_1 be point mass on $x = 1$ and $y = 1$. Then, we can see that $f^*_{g_1} = 1$. Let $0 \le \alpha \le 1$ be arbitrary. The marginal distribution of the group g_2 is the point mass on σ for some $\sigma > 1$. Consider two choices of conditional distribution of y for the group g_2. Under the choice 1, $y = x(1+\alpha)$ almost surely and under the choice 2, $y = x(1-\alpha)$ almost surely. Under both the cases, the performance of the optimal predictor of the group g_1, $f^*_{g_1}$, is equal to $(x - y)^2 = (\sigma - \sigma(1 \pm \alpha))^2 = \sigma^2\alpha^2$. That is, $\Delta = \alpha^2\sigma^2$.

Now, let $\widehat{f}$ be any estimator in the function class $\mathcal{F}$. Observe that $\widehat{f}$ can not depend on the conditional distribution of the sparse group. In particular, the estimator is independent of whether $y = x(1 - \alpha)$ or $y = x(1 + \alpha)$. Since $\widehat{f}$ is a linear predictor in one dimension, either $\widehat{f} \le 1$ or $\widehat{f} \ge 1$. Suppose $\widehat{f} \ge 1$, then its loss on the second choice of the conditional distribution (where $y = x(1-\alpha)$) is at least $\sigma^2\alpha^2 = \Delta$. A similar conclusion holds when $\widehat{f} \le 1$. Thus, every algorithm must incur loss of $\Omega(\Delta)$, either with probability $1/2$ or in expectation (if the underlying conditional distribution on the sparse groups is uniform between the two choices given here), on the sparse group. ■

B.4. Proof of Theorem 6

We first state the relations between the true regret and the proxy that we will choose that is adaptive to the structure of the function class.

B.4.1. RELATIONS BETWEEN TRUE REGRET AND A PROXY REGRET FOR SQUARE LOSS

In this section, our focus is on the squared loss. Thus, the regret of a function f on the group g is defined as follows:

$$\mathsf{Regret}(f,g) = \mathop{\mathbf{E}}_{(x,y)\sim P_g}[(y - f(x))^2] - \mathop{\mathbf{E}}_{(x,y)\sim P_g}[(y - f_g^*(x))^2],$$

where f_g^* is the best in-class predictor for group g, i.e., $f_g^* := \operatorname{argmin}_{f\in\mathcal{F}} \mathbf{E}_{(x,y)\sim P_g}[(y-f(x))^2]$. Since the loss function is the squared loss, we know the optimal Bayes predictor for the group g is given by $h_{\mathsf{Bayes},g}(x) = \mathbf{E}_{(x,y)\sim P_g}[y|x]$. Since $h_{\mathsf{Bayes},g}$ might not belong to the function class $\mathcal{F}$, we let E_g denote the approximation error, $E_g = \mathbf{E}_{x\sim P_g}(h_{\mathsf{Bayes},g}(x) - f_g^*(x))^2$.

Since evaluating the regret requires labeled samples (x,y), we study an alternative notion of regret that does not require labeled samples:

$$\mathsf{Regret}_{\mathsf{l}}(f,g) := \mathbf{E}[(f(x) - f_g^*(x))^2]. \tag{13}$$

The following result captures the approximation between Regret and $\mathsf{Regret}_{\mathsf{l}}$:

Proposition 14 (Relations between Regret and $\mathsf{Regret}_{\mathsf{l}}$) *We have the following relations if the loss function is square loss:*

1. *(No structure on function class $\mathcal{F}$)*

$$|\mathsf{Regret}(f,g) - \mathsf{Regret}_{\mathsf{l}}(f,g)| \le 2\sqrt{E_g \cdot \mathsf{Regret}_{\mathsf{l}}(f,g)}.$$

Thus the both of these measures are equal if the problem is well-specified, i.e., $E_g = 0$ or, equivalently, $h_{\mathsf{Bayes},g} \in \mathcal{F}$.

2. *(Convexity) If the function class $\mathcal{F}$ is convex, then we can obtain a tighter lower bound on* $\mathsf{Regret}(f)$.

$$0 \le \mathsf{Regret}(f,g) - \mathsf{Regret}_{\mathsf{l}}(f,g) \le 2\sqrt{E_g \cdot \mathsf{Regret}_{\mathsf{l}}(f,g)}.$$

3. *(Linearity or well-specified) If either (i) the function class $\mathcal{F}$ is a linear subspace, or more generally, f_g^* lies in relative interior of $\mathcal{F}$, or (ii) $h_{\mathsf{Bayes},g} \in \mathcal{F}$, then*

$$\mathsf{Regret}(f,g) = \mathsf{Regret}_{\mathsf{l}}(f,g).$$

Moreover, these guarantees are essentially tight.

Proof We begin by defining important objects. Define $\epsilon_{\mathsf{Bayes}} := \mathbf{E}[(y - h_{\mathsf{Bayes}}(x))^2]$. We start with a simple standard proposition showing that we can ignore the labels y if we know $h_{\mathsf{Bayes},g}$.

Proposition 15 (Folklore) *For any function f, we have that*

$$\mathbf{E}[(y - f(x))^2] = \epsilon_{\mathsf{Bayes}} + \mathbf{E}\left[(h_{\mathsf{Bayes},g}(x) - f(x))^2\right].$$

The result above implies that

$$\mathsf{Loss}(f,g) = \epsilon_{\mathsf{Bayes}} + \mathbf{E}\left[(h_{\mathsf{Bayes},g}(x) - f(x))^2\right] \text{ and } \tag{14}$$

$$\mathsf{Regret}(f,g) = \mathbf{E}\left[(h_{\mathsf{Bayes},g}(x) - f(x))^2\right] - \mathbf{E}\left[(h_{\mathsf{Bayes},g}(x) - f^*(x))^2\right] \tag{15}$$

In particular, $f_g^* = \operatorname{argmin}_{f\in\mathcal{F}} \mathbf{E}[(h_{\mathsf{Bayes},g}(x) - f(x))^2]$.

For any two functions f and g, we define the inner product $\langle f, g\rangle := \mathbf{E}[f(X)g(X)]$ and use $\|f\|_{L_2}^2 := \langle f, f\rangle$. Then, it can be seen that $f_g^* = \operatorname{argmin}_{f\in\mathcal{F}} \|f - h_{\mathsf{Bayes},g}\|_{L_2}^2$.

1. We start as follows:

$$\begin{aligned}\mathsf{Regret}(f,g) &= \|f - h_{\mathsf{Bayes},g}\|_{L_2}^2 - \|f_g^* - h_{\mathsf{Bayes},g}\|_{L_2}^2 \\ &= \langle f - f_g^*, f + f_g^* - 2h_{\mathsf{Bayes},g}\rangle \\ &= \|f - f_g^*\|_{L_2}^2 + 2\langle f - f_g^*, f_g^* - h_{\mathsf{Bayes},g}\rangle \\ &= \mathsf{Regret}_{\mathsf{I}}(f) + 2\langle f - f_g^*, f_g^* - h_{\mathsf{Bayes},g}\rangle. \end{aligned} \tag{16}$$

   By Cauchy-Scwarz inequality, we have that $\langle f - f_g^*, f_g^* - h_{\mathsf{Bayes},g}\rangle \leq \sqrt{E_g \cdot \mathsf{Regret}_{\mathsf{I}}(f)}$.

2. Observe that f_g^* is the Hilbert projection of $h_{\mathsf{Bayes},g}$ on the set $\mathcal{F}$. Thus for each $f \in \mathcal{F}$, we have that $\langle h_{\mathsf{Bayes},g} - f_g^*, f_g^* - f\rangle \geq 0$, leading to the desired result by Equation (16).

3. The case when $h_{\mathsf{Bayes},g} \in \mathcal{F}$ follows from the first part, and the case when f_g^* lies in relative interior of $\mathcal{F}$, then we have that $\langle h_{\mathsf{Bayes},g} - f_g^*, f_g^* - f\rangle = 0$.

Tightness For example, consider $\mathcal{F} = [0,1]$ and let $h_{\mathsf{Bayes},g} = -1$. Then $f_g^* = 0$, and $E_g = 1$, and $\mathsf{Regret}(f) = (f+1)^2 - 1$, whereas $\mathsf{Regret}_{\mathsf{I}}(f) = f^2$. For $f_1 = 1$, we have that $\mathsf{Regret}(f_1) = 3$, whereas $\mathsf{Regret}_{\mathsf{I}}(f_1) = 1$. Thus we have that $1 = \mathsf{Regret}_{\mathsf{I}}(f_1) \leq \mathsf{Regret}(f_1) = 3 = 1 + 2\sqrt{1\cdot 1} = \widetilde{\mathsf{Regret}}(f_1) + 2\sqrt{E_g \cdot \mathsf{Regret}_{\mathsf{I}}(f_1)}$. ■

B.4.2. PROOF OF THEOREM 6

We are now ready to provide the proof using Proposition 14.

Theorem 6 *Suppose Assumption 1 holds with the parameter Δ. For a group g, let $L_g^* := \min_{f\in\mathcal{F}} \mathsf{Loss}(f, P_g)$. Suppose the loss function is the squared loss and $\mathcal{F}$ is convex, i.e., Assumption 2 holds. Then, the output of Algorithm 1 in the idealized setting (Setting 1) satisfies*

$$\max_{g\in\mathcal{G}} \mathsf{Regret}(\widehat{f}, P_g) = O\Big(\mathsf{OPT}_{\mathsf{R}} + \Delta + \max_{g} \sqrt{L_g^*(\mathsf{OPT}_{\mathsf{R}} + \Delta)}\Big). \tag{4}$$

Furthermore, the upper bound can be tightened to $O(\mathsf{OPT}_{\mathsf{R}} + \Delta)$ if, for all sparse groups g, f_g^* lies in the relative interior of $\mathcal{F}$.

Proof [Proof of Theorem 6] Recall that $L_g^* := \min_{f\in\mathcal{F}} \mathsf{Loss}(f, P_g)$ and the notation of $\widetilde{\mathsf{Loss}}(f, P_g) := \mathbf{E}_{x\sim P_g}[\ell(f(x), f_{\widehat{g}}^*(x))]$ from the proof of Theorem 2. Observe that $\widehat{f}$ in Algorithm 1 is defined as follows:

$$\widehat{f} = \operatorname*{argmin}_{f\in\mathcal{F}} \max\left(\max_{g:\text{dense}} \mathsf{Regret}(f, P_g), \max_{g:\text{sparse}} \widetilde{\mathsf{Loss}}(f, P_g)\right). \tag{17}$$

Furthermore, define $\mathsf{Regret}_{\mathsf{I}}(f, P_g) := \mathbf{E}\left[\left(f(x) - f_g^*(x)\right)^2\right]$ that was also used in Proposition 14.

We can now control the deviation between $\mathsf{Regret}_{\mathsf{I}}$ and $\widetilde{\mathsf{Loss}}$ using Δ:

$$\begin{aligned}
\mathsf{Regret}_{\mathsf{I}}(f, P_g) &= \mathbf{E}\left[\left(f(x) - f_g^*(x)\right)^2\right] \\
&\leq 2\,\mathbf{E}\left[\left(f(x) - f_{\widehat{g}}^*(x)\right)^2\right] + 2\,\mathbf{E}\left[\left(f_{\widehat{g}}^*(x) - f_g^*(x)\right)^2\right] && \text{(using } (a+b)^2 \leq 2a^2 + 2b^2\text{)} \\
&= 2\widetilde{\mathsf{Loss}}(f, P_g) + 2\mathsf{Regret}_{\mathsf{I}}(\widehat{f}, P_g) \\
&\leq 2\widetilde{\mathsf{Loss}}(f, P_g) + 2\mathsf{Regret}(f_{\widehat{g}}^*, P_g) && \text{(using Proposition 14 (ii))} \\
&\leq 2\widetilde{\mathsf{Loss}}(f, P_g) + 2\Delta, && (18)
\end{aligned}$$

where the last inequality is by assumption. Similarly, we obtain $\widetilde{\mathsf{Loss}}(f, P_g) \leq 2\mathsf{Regret}_{\mathsf{I}}(f, P_g) + 2\Delta$.

We now proceed as follows: let f^* be the function achieving the $\mathsf{OPT}_{\mathsf{R}}$, then the output $\widehat{f}$ of the algorithm satisfies the following (below, we use two different metrics for the dense and sparse groups: Regret for the dense groups and $0.5\mathsf{Regret}_{\mathsf{I}}$ for the sparse groups):

$$\begin{aligned}
&\max\left(\max_{g:\text{dense}} \mathsf{Regret}(\widehat{f}, P_g),\ \max_{g:\text{sparse}} 0.5\mathsf{Regret}_{\mathsf{I}}(\widehat{f}, P_g)\right) \\
&\leq \max\left(\max_{g:\text{dense}} \mathsf{Regret}(\widehat{f}, P_g),\ \max_{g:\text{sparse}} \widetilde{\mathsf{Loss}}(\widehat{f}, P_g) + \Delta\right) && \text{(using (18))} \\
&\leq \max\left(\max_{g:\text{dense}} \mathsf{Regret}(f^*, P_g),\ \max_{g:\text{sparse}} \widetilde{\mathsf{Loss}}(f^*, P_g)\right) + \Delta && \text{(using (17))} \\
&\leq \max\left(\max_{g:\text{dense}} \mathsf{Regret}(f^*, P_g),\ \max_{g:\text{sparse}} 2\mathsf{Regret}_{\mathsf{I}}(f^*, P_g) + 2\Delta\right) + \Delta \\
&\leq 2\mathsf{OPT}_{\mathsf{R}} + 3\Delta. && (19)
\end{aligned}$$

The above result implies the bound on the maximum regret over dense groups, i.e., for any dense group g, $\mathsf{Regret}(\widehat{f}, g) = O\left(\mathsf{OPT}_{\mathsf{R}} + \Delta\right)$. For the sparse groups, we need one more step. The convexity of $\mathcal{F}$ implies the following using Proposition 14:

$$0 \leq \mathsf{Regret}(f, g) - \mathsf{Regret}_{\mathsf{I}}(f, P_g) \leq 2\sqrt{L_g^* \cdot \mathsf{Regret}_{\mathsf{I}}(f, P_g)}. \qquad (20)$$

If f_g^* lies in the relative interior of $\mathcal{F}$), then we have a stronger guarantee from Proposition 14 that $\mathsf{Regret}(f, g)$ is exactly equal to $\mathsf{Regret}_{\mathsf{I}}(f, P_g)$.

For any sparse group g, combining Equations (20) and (19), we have that

$$\mathsf{Regret}(\widehat{f}, P_g) \leq \mathsf{Regret}_{\mathsf{I}}(\widehat{f}, P_g) + 2\sqrt{L_g^* \mathsf{Regret}_{\mathsf{I}}(\widehat{f}, P_g)}\,.$$

Furthermore, if the function class is linear or the f^* lies in the relative interior of $\mathcal{F}$ for sparse groups, then $\mathsf{Regret}(\widehat{f}, P_g) = O(\mathsf{OPT}_{\mathsf{R}} + \Delta)$. ■

Appendix C. Finite Sample Guarantees: Proofs from Section 5

C.1. Proof of Lemma 9

Proof For each group g, let $\widehat{f}_g$ be arbitrary and potentially depending on the training data, and define $\Delta'_g := \mathsf{Regret}(\widehat{f}_g, P_g)$. For any two functions f and f', define $\widetilde{\mathsf{Loss}}(f, \widehat{f}_g, P_g) := \mathbf{E}_{x \sim P_g}[\ell(f(x), \widehat{f}_g(x))]$. For convenience, we define $\mathsf{Loss}_{\mathsf{proxy}}(f, P_g) := \widetilde{\mathsf{Loss}}(f, \widehat{f}_g, P_g)$.

By using the triangle inequality and the symmetry of the loss function, we obtain the following relation between Loss and $\mathsf{Loss}_{\mathsf{proxy}}$: we have that

$$\begin{aligned} |\mathsf{Loss}(f, P_g) - \mathsf{Loss}_{\mathsf{proxy}}(f, P_g)| &= \left| \mathop{\mathbf{E}}_{(x,y)\sim P_g}[\ell(f(x), y)] - \mathop{\mathbf{E}}_{x \sim P_g}[\ell(f(x), \widehat{f}_g(x))] \right| \\ &\le \mathop{\mathbf{E}}_{(x,y)\sim P_g}[\ell(\widehat{f}_g(x), y)] \\ &= \mathsf{Loss}(\widehat{f}_g, P_g) = \Delta'_g + L^*_g. \end{aligned} \tag{21}$$

For the rest of the proof, we assume that the event from Assumption 3 holds, which happens with probability $1 - \delta$. Thus, we have that for all $f \in \mathcal{F}$ and $g \in \mathcal{G}$:

$$|\mathsf{Regret}_{\mathsf{emp}}(f, g) - \mathsf{Regret}(f, P_g)| \le 2\gamma_g, \tag{22}$$

$$\left|\mathsf{Regret}_{\mathsf{proxy}}(f, g) - \mathsf{Loss}_{\mathsf{proxy}}(f, P_g)\right| \le \tau_g. \tag{23}$$

Combining the last inequality with Equation (21), we obtain

$$\left|\mathsf{Regret}_{\mathsf{proxy}}(f, g) - \mathsf{Loss}(f, P_g)\right| \le \tau_g + \Delta'_g + L^*_g. \tag{24}$$

Let $\alpha_g \in [0, 1]$ be any choice of parameters in the algorithm that potentially depends on the observed samples. We define $\overline{\alpha_g} := 1 - \alpha_g$. Define $\beta = \max_g \left(\alpha_g \left(2\gamma_g + L^*_g\right) + \overline{\alpha_g}\left(\Delta'_g + L^*_g + \tau_g\right)\right)$, which matches the definition from the lemma statement.

Recall that $\widehat{f}$ in Algorithm 2 is defined as follows:

$$\widehat{f} = \mathop{\mathsf{argmin}}_{f \in \mathcal{F}} \max \left(\alpha_g \mathsf{Regret}_{\mathsf{emp}}(f, P_g) + (1 - \alpha_g)\mathsf{Regret}_{\mathsf{proxy}}(f, P_g)\right). \tag{25}$$

We now proceed as follows: let f^* be the function achieving the $\mathsf{OPT_L}$. Then,

$$\begin{aligned} \max_{g \in \mathcal{G}} \mathsf{Loss}(\widehat{f}, P_g) &= \max_{g \in \mathcal{G}} \left(\alpha_g \left(\mathsf{Regret}(\widehat{f}, P_g) + L^*_g\right) + \overline{\alpha_g}\mathsf{Loss}(f, P_g)\right) \\ &\le \max_g \Big(\alpha_g \left(\mathsf{Regret}_{\mathsf{emp}}(\widehat{f}, P_g) + 2\gamma_g + L^*_g\right) \\ &\qquad + \overline{\alpha_g}\left(\mathsf{Regret}_{\mathsf{proxy}}(\widehat{f}, P_g) + \Delta'_g + L^*_g + \tau_g\right)\Big) \\ &\qquad\qquad \text{(using Equations (22) and (24))} \\ &\le \max_g \left(\alpha_g \mathsf{Regret}_{\mathsf{emp}}(\widehat{f}, P_g) + \overline{\alpha_g}\mathsf{Regret}_{\mathsf{proxy}}(\widehat{f}, P_g)\right) + \beta \\ &\qquad\qquad \text{(using definition of } \beta) \\ &\le \max_g \left(\alpha_g \mathsf{Regret}_{\mathsf{emp}}(f^*, P_g) + \overline{\alpha_g}\mathsf{Regret}_{\mathsf{proxy}}(f^*, P_g)\right) + \beta \\ &\qquad\qquad \text{(since } \widehat{f} \text{ minimizes the objective)} \end{aligned}$$

$$\leq \max_g \left(\alpha_g \left(\mathsf{Regret}(f^*, P_g) + \gamma_g\right) + \overline{\alpha_g} \left(\mathsf{Loss}(f^*, P_g) + \Delta'_g + L^*_g + \tau_g\right)\right) + \beta$$

(using Equations (22) and (24))

$$\leq \max_g \left(\alpha_g \left(\mathsf{Loss}(f^*, P_g)\right) + \overline{\alpha_g} \left(\mathsf{Loss}(f^*, P_g)\right)\right) + 2\beta$$

(using definition of β and $\mathsf{Regret}(f, g) \leq \mathsf{Loss}(f, g)$)

$$= \mathsf{OPT}_{\mathsf{L}} + 2\beta.$$

Finally, we consider the maximum regret of $\widehat{f}$ on $\mathcal{G}$. Here, let f^* be the function achieving $\mathsf{OPT}_{\mathsf{R}}$ and define $\beta' = \max_g \left(\alpha_g \left(2\gamma_g\right) + \overline{\alpha_g} \left(\Delta'_g + \tau_g\right)\right) = \beta$.

$$\max_{g \in \mathcal{G}} \mathsf{Regret}(\widehat{f}, P_g) = \max_{g \in \mathcal{G}} \left(\alpha_g \left(\mathsf{Regret}(\widehat{f}, P_g)\right) + \overline{\alpha_g} \left(\mathsf{Loss}(\widehat{f}, P_g) - L^*_g\right)\right)$$

$$\leq \max_g \left(\alpha_g \left(\mathsf{Regret}_{\mathsf{emp}}(\widehat{f}, P_g) + \gamma_g\right) + \overline{\alpha_g} \left(\mathsf{Regret}_{\mathsf{proxy}}(\widehat{f}, P_g) + \Delta'_g + \tau_g\right)\right)$$

(using Equations (22) and (24))

$$\leq \max_g \left(\alpha_g \mathsf{Regret}_{\mathsf{emp}}(\widehat{f}, P_g) + \overline{\alpha_g} \mathsf{Regret}_{\mathsf{proxy}}(\widehat{f}, P_g)\right) + \beta'$$

(using definition of β')

$$\leq \max_g \left(\alpha_g \mathsf{Regret}_{\mathsf{emp}}(f^*, P_g) + \overline{\alpha_g} \mathsf{Regret}_{\mathsf{proxy}}(f^*, P_g)\right) + \beta'$$

(since $\widehat{f}$ minimizes the objective)

$$\leq \max_g \left(\alpha_g \left(\mathsf{Regret}(f^*, P_g) + \gamma_g\right) + \overline{\alpha_g} \left(\mathsf{Loss}(f^*, P_g) + \Delta'_g + L^*_g + \tau_g\right)\right) + \beta'$$

(using Equations (22) and (24))

$$\leq \max_g \left(\alpha_g \left(\mathsf{Loss}(f^*, P_g)\right) + \overline{\alpha_g} \left(\mathsf{Loss}(f^*, P_g)\right)\right) + \beta' + \beta$$

$$= \mathsf{OPT}_{\mathsf{R}} + \beta + \beta'.$$

■

C.2. Proof of Lemma 10

Proof Fix a group g. Observe that $\widetilde{f}_{g'}$ for $g' \neq g$ are independent of S_g, the training data of the group g. By applying Hoeffding's inequality (which is applicable due to the loss function being bounded) and a union bound, we obtain that if $n \geq c \log(|\mathcal{G}|/\delta)/\epsilon^2$, then with probability $1 - \delta$, for each $g \in \mathcal{G}$ and $g' \in \mathcal{G} \setminus \{g\}$, it holds that $|\mathsf{Loss}(\widetilde{f}_{g'}, S_g) - \mathsf{Loss}(\widetilde{f}_{g'}, P_g)| \leq \epsilon$.

This directly applies the second claim that $|\mathsf{Loss}(\widehat{f}_g, P_g) - \mathsf{Loss}(\widehat{f}_g, S_g)| \leq \epsilon$ since $\widehat{f}_g$ belongs to one of $\{\widetilde{f}_{g'} : g' \neq g\}$. For the first claim, we note that on the same event as before, starting with the conclusion of the second claim above, we obtain the following series of inequalities:

$$\begin{aligned} \mathsf{Loss}(\widehat{f}_g, P_g) &\leq \mathsf{Loss}(\widehat{f}_g, S_g) + \epsilon \\ &= \min_{g' \neq g} \mathsf{Loss}(\widetilde{f}_{g'}, S_g) + \epsilon \\ &\leq \min_{g' \neq g} \mathsf{Loss}(\widetilde{f}_{g'}, P_g) + 2\epsilon \\ &\leq \min_{g' \neq g} L^*_g + \Delta_g(g', 2\gamma_{g'}) + 2\epsilon, \end{aligned}$$

where the last step follows from the definition of Δ_g in Assumption 4. This completes the proof. ■

C.3. Proof of Theorem 11

Proof We will now combine the guarantees of Lemmas 9 and 10 and assume that the events in Lemmas 9 and 10 and Assumption 3 hold simultaneously, which happens with probability at least $1 - 2\delta$. By Lemma 9, it suffices to upper bound the parameter β with high probability.

First, for all groups g, by the event in Assumption 3, $|L_g^* - \mathsf{Loss}(\widetilde{f}_g, S_g)| \leq \gamma_g$ since $\widetilde{f}_g$ minimizes the empirical error. Similarly, by Lemma 10, we have that for all groups g, $\mathsf{Loss}(\widehat{f}, P_g) \leq \mathsf{Loss}(\widehat{f}, S_g) + \epsilon$. Thus, we obtain the following bound on β:

$$\begin{aligned}\beta &= \max_g \left(\alpha_g \left(2\gamma_g + L_g^*\right) + (1-\alpha_g)\left(\mathsf{Loss}\left(\widehat{f}_g, P_g\right) + \tau_g\right)\right) \\ &\leq \max_g \left(\alpha_g \left(2\gamma_g + \gamma_g + \mathsf{Loss}(\widetilde{f}_g, S_g)\right) + (1-\alpha_g)\left(\mathsf{Loss}(\widehat{f}, S_g) + \epsilon + \tau_g\right)\right) \\ &\leq \max_g \left(\alpha_g \left(3\gamma_g + \mathsf{Loss}(\widetilde{f}_g, S_g)\right) + (1-\alpha_g)\left(\mathsf{Loss}(\widehat{f}, S_g) + \epsilon + \tau_g\right)\right).\end{aligned}$$

Algorithm 3 now chooses α_g to minimize the expression for each g. Let $\widetilde{\beta}$ be the resulting expression, which can be further upper bounded as follows using the same inequalities as above:

$$\begin{aligned}\widetilde{\beta} &= \max_g \min\left(3\gamma_g + \mathsf{Loss}(\widetilde{f}_g, S_g), \mathsf{Loss}(\widehat{f}, S_g) + \epsilon + \tau_g\right) & (26) \\ &\leq \max_g \min\left(4\gamma_g + L_g^*, \mathsf{Loss}(\widehat{f}, P_g) + 2\epsilon + \tau_g\right) & (27) \\ &\leq \max_g \min\left(4\gamma_g + L_g^*, L_g^* + \min_{g' \neq g} \Delta_g(g', 2\gamma_{g'}) + 4\epsilon + \tau_g\right). & (28)\end{aligned}$$

This completes the proof. ■

C.4. Lower Bound on Sample Complexity

In this section, we will show a lower bound on the sample complexity for sparse groups, which holds even in a special simple case. Consider the case where there are $|\mathcal{G}|$ groups, where only the group $g_0 \in \mathcal{G}$ is sparse. Define $\mathcal{G}' = \mathcal{G} \setminus \{g_0\}$ to be the set of dense groups. For simplicity, assume that all of the dense groups have infinite labeled samples and the sparse group g' has infinite unlabeled samples. Thus, $\gamma_g = 0, \tau_g = 0$ for all $g \in \mathcal{G}'$; additionally, $\tau_{g_0} = 0$. We further assume the sparse group is perfectly approximated by an (unknown) dense group $g' \in \mathcal{G}'$, i.e., there exists a group $g' \in \mathcal{G}'$ such that $\Delta_{g_0}(g', 0)) = 0$.

Moreover, we assume that for each group $g \in \mathcal{G}$, $L_g^* = 0$, i.e., there is a perfect classifier for each group. Consequently, the guarantee of Theorem 11 says that if the sparse group has $\Omega(\log(|\mathcal{G}|)/\epsilon^2)$ samples, then the output of Algorithm 3, $\widehat{f}$, achieves

$$\max_{g \in \mathcal{G}} \mathsf{Loss}(\widehat{f}, g) \leq \mathsf{OPT_L} + O(\epsilon).$$

The following simple claim shows that a polynomial dependence on both $\log |\mathcal{G}|$ and ϵ is necessary to achieve this guarantee for the zero-one loss.

Theorem 16 *Let ℓ be the zero-one loss and $c \in (0,1)$ be a small enough constant. Let $\epsilon \in (0,c)$ and let $\widehat{f}$ be the output of any fixed algorithm (that may depend on ϵ) belonging to $\mathcal{F}$. Then for any $\tau \in (0,\epsilon/2)$, there exists a learning setup satisfying the constraints defined above such that $\mathsf{OPT_L} = \tau$, and if the number of labeled samples from the sparse group is less than $c(\log|\mathcal{G}|)/\epsilon$, then with probability at least 0.15, $\max_{g\in\mathcal{G}} \mathsf{Loss}(\widehat{f},g) \geq \epsilon \geq \mathsf{OPT_L} + \epsilon/2$.*

Proof We prove this lower bound in three steps:

Distribution over the sparse group. Consider the following hard instance: suppose that the class $\mathcal{F}$ shatters a set $S \subset \mathcal{X}'$ of cardinality $d := \lceil|\log\mathcal{F}|\rceil$; Such a set exists Mohri et al. (2018). We will assume that the sparse group is supported on S.

We will now apply the following lower bound on the sample complexity of PAC Learning:

Lemma 17 (PAC Sample Complexity Lower Bound) *(Mohri et al., 2018; Shalev-Shwartz and Ben-David, 2014) Let $\mathcal{H}$ be any function class that shatters a set S of size d. Let $\epsilon \in 0, 1/2)$ be small enough. Let P be a fixed arbitrary marginal distribution over S. Let $\widehat{f}$ be any algorithm that has access to the marginal distribution P and outputs a classifier in $\mathcal{H}$. Then, given less than $\frac{d-1}{64\epsilon}$ many labeled samples from a labeled distribution whose marginal distribution is P and there exists a unique function in $\mathcal{H}$ with zero error, the output $\widehat{f}$ satisfies that with probability at least $1/15$, $\mathsf{Loss}(\widehat{f},P) \geq \epsilon$.*

Thus, applying this lower bound, we know that unless one takes $\Omega((\log|\mathcal{H}|)/\epsilon)$ many samples, the error is at least $\Omega(\epsilon)$ even though there exists a function in $\mathcal{H}$ with zero error.

We will now embed this hard instance of PAC learning in our setting by defining the distributions of the dense groups and how the function class $\mathcal{F}$ is defined on these groups. In particular, $|\mathcal{F}|$ will be equal to $|\mathcal{G}| - 1$ for us, implying a lower bound of $\Omega((\log|\mathcal{G}|)/\epsilon)$ on the sample complexity.

Distribution of the dense groups. For the dense groups, given infinite labeled samples, we can simply assume access to the joint distribution of the labels and responses. Furthermore, these groups are supported on disjoint domains. The dense groups will be supported on the set $\mathcal{X}\backslash S$ of cardinality at least $2(|\mathcal{G}| - 1)$, where each dense group is supported on the set of two points.

We define the function class $\mathcal{F}$ to be of cardinality of cardinality $|\mathcal{G}| - 1$ defined shortly. Since the groups are supported on disjoint domains, we are free to define the function class $\mathcal{F}$ on the dense groups (without any constraint that arises from assuming that the function class shatters a large set in the sparse group). We will now extend this function class to the support of the dense groups as follows:

The group distribution of a dense group $g \in \mathcal{G}'$ is as follows:

- The distribution P_g is supported on two points: (x_g, y_g) and (x'_g, y'_g). The probability of (x_g, y_g) is $1-\tau$ and probability of (x'_g, y'_g) is τ.
- For all functions f in $\mathcal{F}$, $y_g = f(x_g)$. Moreover, there is a single function $f_g \in \mathcal{F}$ such that $y'_g = f_g(x'_g)$.

Consequently, each group has a unique optimal classifier f_g with perfect prediction; moreover, all other classifiers have average error equal to τ. Since each dense group is identified by its optimal classifier, we obtain that each function has zero error on one dense group and error τ on every other dense group.

Verifying that it is a valid instance. We now verify that this instance satisfies all the conditions described before the lemma statement. First of all, the algorithm has access to the marginal distribution over the sparse group and the complete distribution over the dense groups.

Additionally, each group has a unique function in $\mathcal{F}$ that has zero error. There exists a function f^* in $\mathcal{F}$ that has zero error over the sparse group and the loss of f^* on every other group is bounded by τ. Coupling with the observation that each function $f \in \mathcal{F}$ has error τ on some dense group, we obtain that $\mathsf{OPT}_\mathsf{L} = \tau$.

■

C.5. Dependence on $\max_g L_g^*$ in Finite-Sample Regime

In this section, we extend the lower bound of Theorem 4, which holds when there are no labeled samples from the sparse group, to the finite sample regime, where a sparse group has finitely many samples. In particular, We will show that given any amount of finite labeled data from sparse groups, we must incur an additive dependence on $c \max_g L_g^*$ in our setting for an absolute constant c.

Theorem 18 (Dependence on $\max_g L_g^*$ in Finite-Sample Regime) *There exists a constant $c > 0$ ($c = 1/2$ works) such that the following holds: Let $\mathcal{G} = \{g_1, g_2, g_3\}$ and suppose that the group g_3 has k labeled samples for an arbitrary $k \in \mathbb{N}$ and infinite unlabeled samples, while groups g_1 and g_2 have infinite labeled samples.*[3] *Let the loss be zero-one loss. For every $\epsilon \in (0, 1/4)$ and $k \in \mathbb{N}$, there is a function class $\mathcal{F}$ such that for every proper learning algorithm $\widehat{f}$ there is a choice of distributions such that the following holds: $\Delta = 0$ (in fact, both dense groups are close to g_3 as per Assumption 1), $\max_g L_g^* = 0.25$, $\mathsf{OPT}_\mathsf{L} = 0.25$, and $\mathsf{OPT}_\mathsf{R} = \epsilon$, but, with probability at least $\Omega(1)$, the following two hold: (i) $\max_g \mathsf{Regret}(\widehat{f}, g) \geq c \max_g L_g^*$ and (ii) $\max_g \mathsf{Loss}(\widehat{f}, g) \geq \mathsf{OPT}_\mathsf{L} + c \max_g L_g^*$.*

In particular, with any finite amount of unlabeled data, the term $\Omega(\max_g L_g^*)$ is inherent in Theorem 11 even when $\Delta = 0$.[4]

Proof Following the proof structure of Theorem 4, we consider a setting of three groups $\{g_1, g_2, g_3\}$, where g_1 and g_2 are dense groups, with their unique optimizers f_1 and f_2. Instead of simply having f_3 and f_4 to confuse the learner, we will (shortly) define multiple functions. The sparse group g_3 will now be supported on k points $H = \{x_1, \dots, x_k\}$ uniformly, for some large even k. Observe that the proof in Theorem 4 considered only $k = 2$.

To show a lower bound, we can assume that we have complete knowledge of the joint distribution of features and labels on the dense groups. We now consider different choices of conditional distribution of y on the sparse group generated as follows (as opposed to only two, P and Q, in Theorem 4): Let $\mathcal{H}$ be the set of subsets of H that have cardinality equal to $k/2$. For each $C \in \mathcal{H}$, define the candidate joint distribution P_C to have conditional expectation $\mathbf{E}[y|x] = 0$ if $x \in C$, otherwise 0.5; Additionally, define $f_C(x) = 1$ if $x \in C$ and 0 otherwise.

We now consider the function class $\mathcal{F} = \{f_1, f_2\} \bigcup \{f_C : C \in \mathcal{H}\}$, where f_1 and f_2 are defined to be uniformly zero. On the dense groups g_1 and g_2, the functions $\{f_C : C \in \mathcal{H}\}$ are defined

[3] Thus, this learning setup is even easier that the finite sample regime in Theorem 11, and thus the lower bound is more powerful.

[4] Observe that given infinite labeled data from the groups g_1 and g_2, Assumption 4 reduces to Assumption 1.

arbitrarily so that their average values are ϵ for an arbitrarily small ϵ.[5] Thus f_1 and f_2, which are the (unique) optimal classifiers of the dense groups, achieve the minimum loss, $1/4$, on the sparse group g_3 for all the candidate distributions $\{P_C : C \in \mathcal{H}\}$; Recall that f_1 (similarly, the function f_2) achieves unit loss on the group g_2 (group g_1, respectively). For two subsets C, C' in $\mathcal{H}$, the loss of the function f_C on the joint distribution $P_{C'}$ is equal to

$$\frac{0.5(|(C')^{\complement}|) + 1(|C' \cap C|)}{k} + \frac{|C' \cap C^{\complement}|)}{k} = \frac{1}{4} + \frac{|C' \cap C|}{k}\,. \tag{29}$$

Thus for each candidate distribution P_C, there are three functions, $f_1, f_2, f_3 \in \mathcal{F}$, that have the minimum loss $1/4$, and all other functions have loss values bigger than $1/4$. Thus, $\mathsf{OPT}_{\mathsf{L}} = 1/4$, $\mathsf{OPT}_{\mathsf{R}} = \epsilon$, $\max_g L_g^* = 0.25$, and $\Delta = 0$, and a learner algorithm must output a function whf from $\{f_C : c \in \mathcal{H}\}$ to get $\max_g \mathsf{Loss}(\widehat{f}, g) < 1$. We will now show that any learner must, with large constant probability, incur loss at least $\mathsf{OPT}_{\mathsf{L}} + c \max_g L_g^*$ on the sparse group (and hence max regret at least $c \max_g L_g^*$) when given $o(k)$ many labeled samples from the sparse group. Since k is arbitrary (independent of the group size), we get that the additive dependence on $c \max_g L_g^*$ is inherent.

Let $\widehat{f}$ be the output of the algorithm, belonging in $\{f_C : C \in \mathcal{H}\}$; Alternatively, we may think of the output of the algorithm as $\widehat{f} = f_{\widehat{C}}$ for some $\widehat{C} \in \mathcal{H}$. If the algorithm achieves $\mathsf{Loss}(f_{\widehat{C}}, g_3) \leq 0.25 + c \max_g L_g^* = 0.25 + 0.25c$, then the output $\widehat{C}$ satisfies $|\widehat{C} \cap C_*| \leq ck/4$, where C_* corresponds to the (unknown) conditional distribution on the sparse group P_{C_*}.

Now, consider the setting where C_* is unknown to the learner and chosen uniformly from $\mathcal{H}$. Then, standard information-theoretic arguments using Fano's inequality implies that to get $|\widehat{C} \cap C_*| \leq k/8$ requires $\Omega(k)$ samples (Mohri et al., 2018). Intuitively, this is because each sample (x_i, y_i) from the sparse group (at best) can tell us whether a particular x_i is in C_* or not (based on the value of y_i). Thus, we obtain that an additive dependence on $0.5 \max_g L_g^*$ is needed for all proper learning algorithms. ■

Appendix D. Algorithms for Min-max Optimization using Weighted-ERM Oracle

For completeness, we mention how to perform minmax regret optimization given a weighted ERM oracle. That is, how to compute $\widehat{f}$, defined as

$$\min_{f \in \mathcal{F}} \max_{g \in \mathcal{G}} \mathsf{Regret}(f, S_g).$$

Recall that this is exactly what is needed in Algorithms 1 and 2, where for sparse groups, one uses the proxy data to define the regret.

Let us define the weighted-ERM oracle:

Definition 19 (Weighted ERM Oracle for a Function Class $\mathcal{F}$) *Let $\mathcal{W} = \{(x_i, y_i, z_i) : i \in [n]\}$ be a weighted dataset, where x_i are the features, y_i are the labels, and $z_i \in \mathbb{R}_+$ are the weights. We say $\mathcal{O}$ is a weighted ERM oracle for function class $\mathcal{F}$ and the loss function $\ell(\cdot, \cdot)$ if given any weighted dataset $\mathcal{W}$, the oracle returns a function $\widehat{f}$ that optimizes $\min_{f \in \mathcal{F}} \sum_{i \in n} z_i \ell(f(x_i, y_i)$.*

We now present the following algorithm for minmax regret optimization from Agarwal and Zhang (2022); see also Agarwal et al. (2018); Diana et al. (2021):

[5] This can be achieved by increasing the domain of groups g_1 and g_2, if needed.

Algorithm 4 Min-max optimization using Weighted-ERM Oracle

```
Require: Function class $\mathcal{F}$, set of groups $\mathcal{G}$, loss function $\ell(\cdot,\cdot)$, learning rate $\eta_t$, iteration count $T$, a weighted ERM oracle $\mathcal{O}_{\mathcal{F},\ell}$ for $\mathcal{F}$ and $\ell(\cdot,\cdot)$ (see Definition 19), labeled data points for each group: $\{\mathcal{S}_g : g \in \mathcal{G}\}$.
for each $g \in \mathcal{G}$ do
    Set $\widehat{L}_g \leftarrow \min_{f\in\mathcal{F}} \mathsf{Loss}(f, S_g)$.
end for
Initialize $\lambda_0 \in \mathbb{R}_+^{|\mathcal{G}|}$ as $\lambda_0(g) = \frac{1}{|\mathcal{G}|}$.
for $t = 1, 2, \dots, T$ do
    Calculate the weighted dataset $\mathcal{W}_t$ (cf. Definition 19), where the weight of a point $(x, y)$ belonging to the group $g$ is equal to $\lambda_t(g)$.
    Set $f_t$ to be the output of $\mathcal{O}$ on the weighted data $\mathcal{W}_t$, i.e., $f_t \leftarrow \mathcal{O}_{\mathcal{F},\ell}(\mathcal{W}_t)$
    for each $g \in \mathcal{G}$ do                                  ▷ Update $\lambda_t$
        $\lambda_{t+1}(g) := \lambda_t(g) \cdot \exp\left(\eta_t \cdot \left(\mathsf{Loss}(f_t, S_g) - \widehat{L}_g\right)\right)$.
    end for
end for
return the average $\widehat{f}(\cdot) := (1/T)\sum_{t=1}^{T} f_t(\cdot)$.
```

Lemma 20 (Agarwal and Zhang (2022)) *For any T and $\eta = \sqrt{(\log|\mathcal{G}|)/T}$, the returned output $\widehat{f} = \frac{1}{T}\sum_{t=1}^{T} f_t$ from Algorithm 4 satisfies:*

$$\sum_{t=1}^{T} \frac{1}{T} \sup_{g\in\mathcal{G}} \mathsf{Regret}(f_t, g) \le \min_{f\in\mathcal{F}} \max_{g\in\mathcal{G}} \mathsf{Regret}(f, g) + O\left(\sqrt{\frac{\log|\mathcal{G}|}{T}}\right). \tag{30}$$

Remark 21 *We note that it is easy to modify the algorithm and analysis here to adapt to various settings such as:*

- *Optimizing* $\max_{g\in\mathcal{G}} \mathsf{Loss}(f, g)$ *instead of* $\max_{g\in\mathcal{G}} \mathsf{Regret}(f, g)$.
- *Optimizing the maximum loss subject to the constraint that the average loss over groups is bounded. in particular, for a tunable parameter $\gamma \ge 0$, we want to solve*

$$\min_{f\in\mathcal{F}} \sum_{g\in\mathcal{G}} w_g \mathsf{Loss}(f, g)$$
$$\text{subject to} \max_{g\in\mathcal{G}} \mathsf{Loss}(f, g) \le \gamma.$$

We refer the reader to Diana et al. (2021); Abernethy et al. (2022) for further details.

Appendix E. Experimental Results

In this section we evaluate the empirical performance of our algorithm on synthetic and real world datasets. For comparison we include two natural baselines, the empirical risk minimizer (ERM), i.e., the minimizer of the average empirical loss, and the empirical min-max optimizer, i.e., the algorithm that simply invokes an existing method for min-max optimization. We first describe our results on the synthetic dataset.

Synthetic dataset We consider a linear regression problem in 100 dimensions. The data is partitioned into 15 groups, where 5 groups are dense and 10 groups are sparse. Letting the number of labeled examples to be n, each dense group has $(0.99/5) \cdot n$ number of samples, while each sparse group has only $0.001 \cdot n$ many samples. For each group, the true distribution follows a linear model with Gaussian covariates and Gaussian responses. The data is generated as follows. We first sample a random "central" vector w^* of unit norm. For each dense group i, we generate $y = w_i^* \cdot x + \epsilon_i$, where ϵ_i is a mean zero random Gaussian noise variable and x is drawn from $N(0, I)$. Furthermore, w_i^* is a random norm 1 vector that is $\delta = 0.01$ close to w^*. For a sparse group i, we select one of the dense groups at random and set w_i^* to be a random vector that is 0.0001 close to the optimal regressor of the dense group. Furthermore, we set $\Sigma = 4I$ and perform a rank one update to remove 0.01 fraction of the variance from the direction of $w^* - w_i^*$. This rank-one update ensures that the error of w^* on the sparse group will be much smaller than it would have been otherwise.

We first consider the setting with $n = 5000$. Thus each dense group has roughly 1000 samples while each sparse group has only 5 samples. As shown in Figure 2, natural baselines such as ERM or the algorithm that minimizes max-loss on training data, do not generalize well in the data scarce regime. However, our proposed algorithm achieves near-optimal error even when the number of labeled samples per sparse group is only 5 although the inherent dimensionality is 100!

We next investigate the effect of labeled examples on our proposed algorithm. We vary n from 1000 to 5000 and the plot the performance of the proposed algorithm when the number of unsupervised data is fixed (8000). Note that even when $n = 1000$ and thus each dense group has roughly 200 samples, there are enough samples in each dense group to fit the model on itself in isolation since the dimensionality is 100. Thus the large error for small n is only due to lack of samples in sparse groups. As we can see from Figure 3, as soon as we have 3 labeled samples per sparse group, our algorithm beats the baselines achieves close to asymptotic optimal performance.

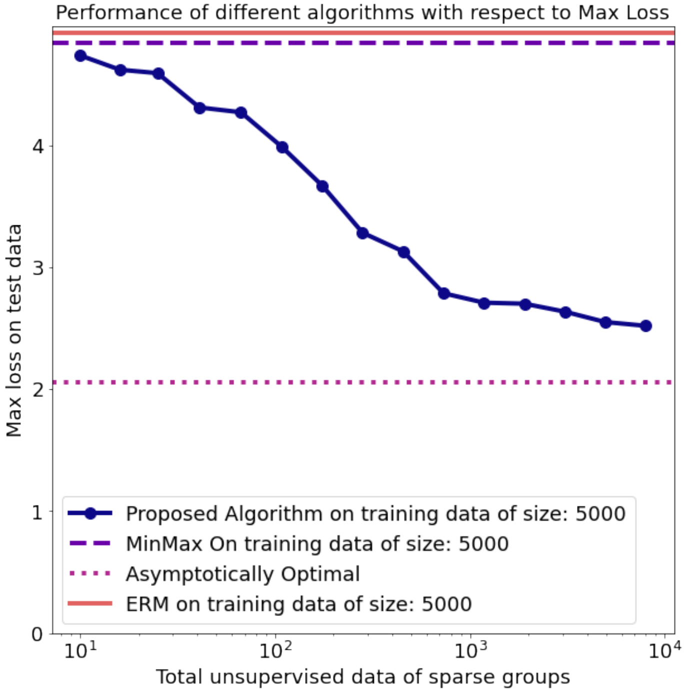

Figure 2: In this figure, we compare the performance of different algorithms on a synthetic training data. The value of Δ, the regret of the optimal model of the closest dense group of each sparse group, is approximately 0.2. The y-axis plots the max-loss on test data, i.e., generalization to unseen data. Even though the optimal asymptotic error is approximately 2, natural baseslines do not generalize to unseen samples. However, Algorithm 1 performs close to optimal as the number of unsupervised examples increase. With roughly, 100 unlabeled samples per group (which corresponds to a total of 1000 unlabeled samples), which is equal to the dimensionality, the algorithm achieves near-optimal error. Recall that the number of labeled samples per sparse group is only 5!

Real Dataset

Here we consider a dataset from the DomainBed benchmark [6] that contains popular datasets for evaluating the performance of various algorithms for robustness and domain adaptation tasks. In particular we evaluate our proposed algorithm on the Colored MNIST dataset. The dataset consists of three domains each containing a disjoint set of digits that are labeled either red or blue. The domains differ by how much the color is correlated with the true label. We consider each domain $d \in \{0, 1, 2\}$ as a group in our setup and we pick group 0 to be the sparse one. We create several versions of the Colored MNIST dataset where an α fraction of the data from group 0 is considered to be unlabeled, and 1% of the data from the remaining groups is considered unlabeled. For each value of α we compare our proposed algorithm for semi supervised DRO with the empirical risk minimization (ERM) and the Group DRO baselines, where the baselines are run on only the provided labeled data. The results are shown in Figure 4. As can be seen for smaller values of the

[6] https://github.com/facebookresearch/DomainBed

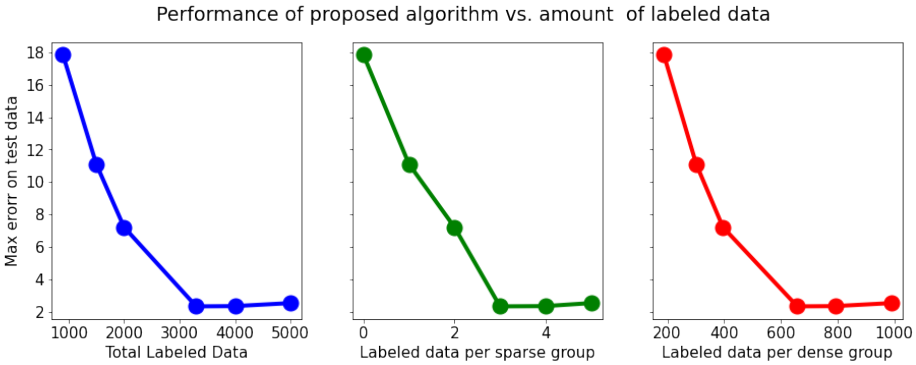

Figure 3: We plot the performance of our algorithm as the amount of labeled samples increase. We vary the amount of labeled data, n, and fix the amount of unsupervised data of sparse groups. For each value of n — the amount of total labeled data — each of the 5 dense groups has $0.99n/5$ samples and each of the 10 dense groups has $0.001n$ samples. In all of these plots, the y-axis corresponds to the max-loss on the test data. The left plot shows the total amount of labeled data, the center plot shows the amount of labeled data per sparse group, and the right plot shows the amount of labeled data per dense group. Note that even when $n = 1000$ and thus each dense group has roughly 200 samples, there are enough samples to fit the model on each dense in isolation — the dimensionality is 100. Thus the large error for small n comes only due to lack of samples in sparse groups. As we can see, as soon as we have 3 labeled samples per sparse group, our algorithm achieves close to asymptotic optimal performance.

sparsity parameter our algorithm matches the performance of Group DRO and as the sparsity parameter increases our semi supervised approach outperforms Group DRO by effectively leveraging the unlabeled data.

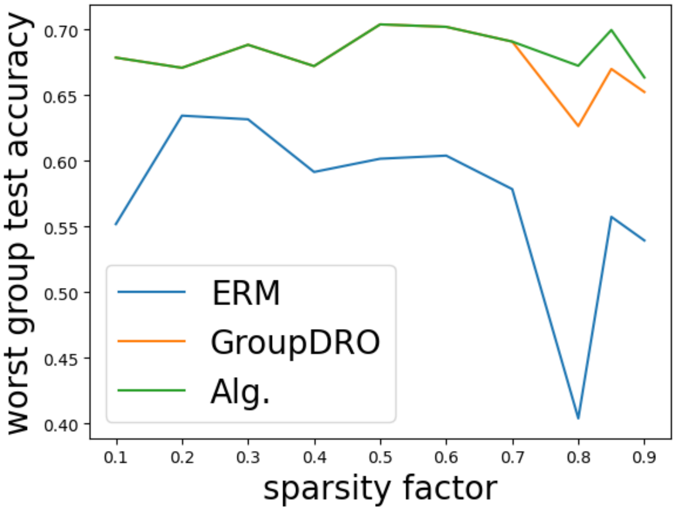

Figure 4: Performance of our proposed algorithm as compared to the baselines as a function of the amount of sparsity in group 0 for the Colored MNIST dataset.