To Pool or Not To Pool: Analyzing the Regularizing Effects of Group-Fair Training on Shared Models

Cyrus Cousins
University of Massachusetts Amherst

I. Elizabeth Kumar
Brown University

Suresh Venkatasubramanian
Brown University

Abstract

In fair machine learning, one source of performance disparities between groups is overfitting to groups with relatively few training samples. We derive group-specific bounds on the generalization error of welfare-centric fair machine learning that benefit from the larger sample size of the majority group. We do this by considering group-specific Rademacher averages over a restricted hypothesis class, which contains the family of models likely to perform well with respect to a fair learning objective (e.g., a power-mean). Our simulations demonstrate these bounds improve over a naïve method, as expected by theory, with particularly significant improvement for smaller group sizes.

1 INTRODUCTION

It is well-known that learned models can have performance or outcome disparities on underrepresented or disadvantaged groups in a distribution (Buolamwini and Gebru, 2018; Obermeyer et al., 2019). Research suggests that these disparities are the result of a complex interaction between the training procedure, model class, and training data (Chen et al., 2018).

Group-based welfare-centric machine learning attempts to mitigate disparities by optimizing *aggregations of per-group risk values*, rather than average overall loss. In other words, the task is to approximate $\operatorname{argmin}_{h \in \mathcal{H}} \mathrm{M}\big(\mathrm{R}(h, \mathcal{D}_1), \ldots, \mathrm{R}(h, \mathcal{D}_g)\big)$ for some *malfare function* $\mathrm{M}(\cdot)$, where $\mathrm{R}(h, \mathcal{D}_i)$ is the risk (average loss) of group i under model h. Such objectives produce models that fairly compromise among groups in various ways. The malfare function determines the fairness concept; for example, $\boldsymbol{w}$-weighted risk minimization is equivalent to optimizing *utilitarian malfare* $\mathrm{M}_1(\mathcal{S}; \boldsymbol{w}) = \mathcal{S} \cdot \boldsymbol{w}$, and taking $\mathrm{M}(\cdot)$ to be the *maximum* produces the *minimax-optimal* h^*, a.k.a., the *egalitarian* or *Rawlsian* fair model.

However, training with *empirical risk* is susceptible to "overfitting to fairness," wherein models overfit small or high-risk *minority groups.* Cousins (2021, 2022, 2023b) shows that generalization error (overfitting) of the *overall objective* decreases with *each group's* sample size, but the current SOTA generalization bounds *for group* i depend only on *group* i's sample size. We address this discrepancy; in particular, we show that in fair learning, each group i effectively learns over a "restricted class" of models that are reasonably likely given the training data *for all groups* $j \neq i$, thus we bound their generalization error via Rademacher averages *of the restricted class*, improving over existing bounds based on the original hypothesis class.

We begin by introducing notation and preliminary concepts (section 2.1) and situating our approach with respect to existing literature (section 2.2). We derive group-specific bounds on the generalization error of jointly trained models, which benefit from the larger sample size of the majority group (section 3). These techniques also translate to improved bounds on the generalization error of the malfare objective itself. Additionally, we experimentally verify our methods on synthetic linear and logistic regression tasks, finding that our bounds better describe the overfitting behavior of fair-learning methods than SOTA analysis (section 4). Our analysis allows us to resolve key real-world problems, such as when multiple groups benefit from pooling data to train a single (shared) model. All proofs are relegated to appendix A.

2 BACKGROUND

We now introduce notation and preliminary concepts, followed by a brief review of related work.

2.1 Preliminaries

We assume a standard supervised learning setting. Given domain label space $\mathcal{Y}$, domain $\mathcal{X}$, and codomain $\mathcal{Y}'$, we assume a *hypothesis class* $\mathcal{H} \subseteq \mathcal{X} \to \mathcal{Y}'$ and

Proceedings of the 27th International Conference on Artificial Intelligence and Statistics (AISTATS) 2024, Valencia, Spain. PMLR: Volume 238. Copyright 2024 by the author(s).

loss function $\ell : \mathcal{Y}' \times \mathcal{Y} \to \mathbb{R}$. Now, suppose a sample $(\boldsymbol{x}, \boldsymbol{y}) = \boldsymbol{z} \in (\mathcal{X} \times \mathcal{Y})^m$ or instance distribution $\mathcal{D}$ over $\mathcal{X} \times \mathcal{Y}$. We define the *empirical risk* of hypothesis h as

$$\hat{\mathrm{R}}(h, \boldsymbol{z}) \doteq \frac{1}{m} \sum_{i=1}^{m} \ell(h(\boldsymbol{x}_i), \boldsymbol{y}_i) \enspace ,$$

and the true risk over the distribution $\mathcal{D}$ as

$$\mathrm{R}(h, \mathcal{D}) \doteq \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}} [\ell(h(x), y)] \enspace .$$

A standard supervised learning task then identifies the *empirical risk minimizer*

$$\hat{h} \doteq \operatorname*{argmin}_{h \in \mathcal{H}} \hat{\mathrm{R}}(h, \boldsymbol{z})$$

as a proxy for the *true risk minimizer*

$$h^* \doteq \operatorname*{argmin}_{h \in \mathcal{H}} \mathrm{R}(h, \mathcal{D}) \enspace .$$

This framework encapsulates simple supervised settings where $\mathcal{Y} = \mathcal{Y}'$, such as least-squares regression or hard binary classification, but it also contains more sophisticated supervised learning settings, like probabilistic classification or conditional density estimation.

Group-Fair Learning This work considers *group-fair learning*, in which we assume not one instance distribution $\mathcal{D}$, but rather g *per-group* instance distributions $\mathcal{D}_{1:g}$, and per-group samples $\boldsymbol{z}_i \sim \mathcal{D}_i^{\boldsymbol{m}_i}$ where $\boldsymbol{m}_i$ is the sample size for group i. The distribution $\mathcal{D}_i$ encapsulates the situations encountered by members of each group i, which may vary in $\mathcal{X}$ (situations encountered by each group), as well as their conditional labels $\mathcal{Y}|\mathcal{X}$ (responses or labels to a given situation).

To treat groups fairly, we consider objectives that consider the risk of all groups. In particular, we assume a cardinal *malfare function* $\mathrm{M}(\cdot) : \mathbb{R}^g \to \mathbb{R}$, and we then seek the *empirical malfare minimizer*

$$\begin{aligned}\hat{h} &\doteq \operatorname*{argmin}_{h \in \mathcal{H}} \mathrm{M}\big(i \mapsto \hat{\mathrm{R}}(h, \boldsymbol{z}_i)\big) \\ &= \operatorname*{argmin}_{h \in \mathcal{H}} \mathrm{M}\big(\hat{\mathrm{R}}(h, \boldsymbol{z}_1), \hat{\mathrm{R}}(h, \boldsymbol{z}_2), \dots, \hat{\mathrm{R}}(h, \boldsymbol{z}_g)\big)\end{aligned}$$

as a proxy for the *true malfare minimizer*

$$h^* \doteq \operatorname*{argmin}_{h \in \mathcal{H}} \mathrm{M}\big(i \mapsto \mathrm{R}(h, \mathcal{D}_i)\big) \enspace .$$

On Malfare Functions The choice of malfare function $\mathrm{M}(\cdot)$ directly encodes how one wishes to make tradeoffs between various groups at various levels of risk. The malfare function is thus a fundamental fair-learning hyperparameter that must be selected to achieve a modeler's desired fairness properties, i.e., choosing a malfare function is equivalent to choosing a fairness concept.

Two popular choices are the *utilitarian malfare* (weighted average), which generally weights the risk of each group proportional to their size, and the *egalitarian malfare*, which seeks to lift up the most disadvantaged groups by minimizing the maximum risk. These are in some sense two extremes of a spectrum: utilitarian malfare only weights groups, and does not distinguish between high-risk and low-risk groups (no equitable redistribution), whereas egalitarian malfare considers only the risk of each group, offering preferential treatment to those most in need (no consideration of non-minimal groups). It is known that both of the above malfare functions belong to a general class of such functions.

Definition 1 (Power-Mean Malfare). *Suppose some* $p \geq 1$, *positive* probability measure $\boldsymbol{w} \in \triangle_g$, *and non-negative* risk vector $\mathcal{S} \in \mathbb{R}^g_{0+}$. *We define the* weighted power-mean *as*

$$\mathrm{M}_p(\mathcal{S}; \boldsymbol{w}) \doteq \sqrt[p]{\sum_{i=1}^{g} \boldsymbol{w}_i \mathcal{S}_i^p} \enspace , \enspace \mathrm{M}_\infty(\mathcal{S}; \boldsymbol{w}) \doteq \max_{i \in 1, \dots, g} \mathcal{S}_i \enspace . \quad (1)$$

Both utilitarian and egalitarian malfare arise as power-mean special-cases $p = 1$ and $p = \infty$, respectively.

The power-mean class is axiomatically justified (Debreu, 1959; Gorman, 1968; Cousins, 2021, 2023b), which motivates its use in a variety of learning and allocation settings (Barman et al., 2020; Cousins et al., 2022; Viswanathan and Zick, 2023; Cousins et al., 2023a,b). Fairness and robustness are closely linked, and Cousins (2023a) also motivates power-means, as well as Gini malfare, and other malfare classes, from the perspective of robustness. This work is neutral to the choice of malfare function; we only seek to show that our methods may be applied to any malfare concept that meets certain broad criteria.

We generally assume that $\mathrm{M}(\cdot)$ is *monotonic*, i.e., that increasing any group's risk never decreases malfare. Furthermore, *convex* malfare functions are convenient for optimization, and in section 3.3 we utilize this property to efficiently bound Rademacher averages. Finally, several of our bounds have algebraically convenient corollaries if we assume *Lipschitz continuity*, i.e., small changes to risk yield small changes to malfare. The power-mean malfare family, as well as other malfare classes, such as the Gini class (Weymark, 1981; Gajdos and Weymark, 2005) or the *utilitarian-maximin* class (Deschamps and Gevers, 1978; Bossert and Kamaga, 2020; Schneider and Kim, 2020), each arise uniquely from their own sets of axioms. Each assume some type of *monotonicity*, *transfer principles*, such as the Pigou-Dalton (Pigou, 1912; Dalton, 1920), which incentivize *equitable redistribution* of harm and give rise to convexity, as well as some concept of *continuity*, which coupled with functional analysis of the resultant class, give rise to Lipschitz continuity. Our criteria for malfare functions are thus quite reasonable.

Statistical Background The Rademacher average is a key statistical tool used to bound the *supremum* *deviation* of empirical means from their expectations (Bartlett and Mendelson, 2002). Denote the *loss class* $\ell \circ \mathcal{H} \doteq \{(x, y) \mapsto \ell(h(x), y) \mid h \in \mathcal{H}\}$, and define Rademacher averages as follows.

Definition 2 (Rademacher Averages)**.** *Let* $\boldsymbol{\sigma}_{1:m}$ *be a vector of* m *i.i.d.* $\mathrm{Unif}(\pm 1)$ *random variables. The* empirical Rademacher average *is then*

$$\hat{\mathfrak{R}}_m(\ell \circ \mathcal{H}, \boldsymbol{z}) \doteq \mathop{\mathbb{E}}_{\boldsymbol{\sigma}} \left[\sup_{h \in \mathcal{H}} \frac{1}{m} \sum_{i=1}^{m} \boldsymbol{\sigma}_i \ell(h(\boldsymbol{x}_i), \boldsymbol{y}_i) \right] \enspace ,$$

i.e., the maximum correlation of any $h \in \mathcal{H}$ *with noise on a sample* $\boldsymbol{z} \in (\mathcal{X} \times \mathcal{Y})^m$*, and the* Rademacher average *is its expectation over i.i.d. samples from* $\mathcal{D}$*, i.e.,*

$$\mathfrak{R}_m(\ell \circ \mathcal{H}, \mathcal{D}) \doteq \mathop{\mathbb{E}}_{\boldsymbol{z} \sim \mathcal{D}^m} \left[\hat{\mathfrak{R}}_m(\ell \circ \mathcal{H}, \boldsymbol{z}) \right] \enspace .$$

Assuming *bounded loss range* r, let $\boldsymbol{\varepsilon}_i \doteq r\sqrt{\frac{\ln \frac{1}{\delta}}{2\boldsymbol{m}_i}}$ and $\hat{\boldsymbol{\eta}}_i \doteq 2\hat{\mathfrak{R}}_{\boldsymbol{m}_i}(\ell \circ \mathcal{H}, \boldsymbol{z}_i) + 2\boldsymbol{\varepsilon}_i$. For any failure probability δ, $h \in \mathcal{H}$, and group i, sampling error is bounded as

$$\mathop{\mathbb{P}}_{\boldsymbol{z}_i \sim \mathcal{D}_i^{\boldsymbol{m}_i}} \left(\left| \hat{\mathrm{R}}(h, \boldsymbol{z}_i) - \mathrm{R}(h, \mathcal{D}_i) \right| > \boldsymbol{\varepsilon}_i \right) < 2\delta \enspace . \qquad (2)$$

Moreover, considering all $h \in \mathcal{H}$ simultaneously, we have for each group i that

$$\mathop{\mathbb{P}}_{\boldsymbol{z}_i \sim \mathcal{D}_i^{\boldsymbol{m}_i}} \begin{pmatrix} 2\mathfrak{R}_m(\ell \circ \mathcal{H}, \mathcal{D}_i) > 2\hat{\mathfrak{R}}_m(\ell \circ \mathcal{H}, \boldsymbol{z}_i) + \boldsymbol{\varepsilon}_i \\ \bigvee \sup_{h \in \mathcal{H}} \left| \hat{\mathrm{R}}(h, \boldsymbol{z}_i) - \mathrm{R}(h, \mathcal{D}_i) \right| > \hat{\boldsymbol{\eta}}_i \end{pmatrix} < 3\delta \enspace . \qquad (3)$$

Equations (2) and (3) are used throughout for various hypotheses and hypothesis classes, both in the above forms, and as 1-tailed variants. These "textbook results" are now standard in learning theory[1] (Shalev-Shwartz and Ben-David, 2014; Mitzenmacher and Upfal, 2017).

The quantity $|\hat{\mathrm{R}}(h, \boldsymbol{z}_i) - \mathrm{R}(h, \mathcal{D}_i)|$ of (2) is the *absolute deviation* between the empirical risk and the expected risk for each individual $h \in \mathcal{H}$, and it bounds the *estimation error* (i.e., error due to sampling) of any such function. The quantity $\sup_{h \in \mathcal{H}} |\hat{\mathrm{R}}(h, \boldsymbol{z}_i) - \mathrm{R}(h, \mathcal{D}_i)|$ of (3) is known as the *supremum deviation* over the loss class $\ell \circ \mathcal{H}$, and it bounds the *generalization error*, both due to sampling error and due to selection bias (training), of the learned $\hat{h}$.

From (3) and a union-bound over groups, following Cousins (2021, 2022, 2023b), we probabilistically bound each group's generalization error (training-true risk gap) as

$$\mathop{\mathbb{P}}_{\boldsymbol{z}_{1:g}} \left(\forall i \colon \left| \hat{\mathrm{R}}(\hat{h}, \boldsymbol{z}_i) - \mathrm{R}(\hat{h}, \mathcal{D}_i) \right| \leq \hat{\boldsymbol{\eta}}_i \right) \geq 1 - 3g\delta \enspace . \qquad (4)$$

Moreover, if $\mathrm{M}(\mathcal{S})$ is monotonically increasing in $\mathcal{S}$, then the malfare generalization error obeys

$$\mathop{\mathbb{P}}_{\boldsymbol{z}_{1:g}} \begin{pmatrix} \mathrm{M}\big(i \mapsto \hat{\mathrm{R}}(\hat{h}, \boldsymbol{z}_i) - \hat{\boldsymbol{\eta}}_i\big) & \text{(Empcl. LB)} \\ \leq \mathrm{M}\big(i \mapsto \mathrm{R}(\hat{h}, \mathcal{D}_i)\big) & \text{(True Malfare)} \\ \leq \mathrm{M}\big(i \mapsto \hat{\mathrm{R}}(\hat{h}, \boldsymbol{z}_i) + \hat{\boldsymbol{\eta}}_i\big) & \text{(Empcl. UB)} \end{pmatrix} \geq 1 - 3g\delta \enspace , \qquad (5)$$

i.e., the true malfare of $\hat{h}$ is sandwiched by upper and lower bounds in terms of empirical malfare. Finally, using also a union bound over (2), the gap between the true risk of the empirical malfare minimizer $\hat{h}$ and the true malfare minimizer h^* is

$$\mathop{\mathbb{P}}_{\boldsymbol{z}_{1:g}} \begin{pmatrix} \mathrm{M}\big(i \mapsto \mathrm{R}(\hat{h}, \mathcal{D}_i) - \hat{\boldsymbol{\eta}}_i\big) \\ \leq \mathrm{M}\big(i \mapsto \mathrm{R}(h^*, \mathcal{D}_i) + \boldsymbol{\varepsilon}_i\big) \end{pmatrix} \geq 1 - 5g\delta \enspace . \qquad (6)$$

2.2 Related work

This work follows others in group-based welfare-centric fair machine learning. This often takes the form of *Rawlsian* or *egalitarian* learning, also known as *minimax fair learning*, wherein $\mathrm{M}(\cdot)$ is the maximum function, and the goal is to minimize the maximum (over groups) average loss (Diana et al., 2021; Shekhar et al., 2021; Abernethy et al., 2022; Martinez et al., 2020; Lahoti et al., 2020; Cortes et al., 2020; Shekhar et al., 2021; Dong and Cousins, 2022), which is a form of *distributionally robust optimization* (Hu et al., 2018; Oren et al., 2019; Sagawa et al., 2019). Most such works only consider performance over the training set, but the Seldonian learner framework (Thomas et al., 2019) explicitly requires trained models be probably approximately optimal w.r.t. some constrained nonlinear objective. Similarly, the fair-PAC learning framework (Cousins, 2021, 2023b) considers malfare minimization with power-mean objectives.

Due to the nonlinearity of $\mathrm{M}(\cdot)$, existing work bounds generalization errror *separately* for each group j, and applies assumed Lipschitz or Hölder continuity properties of $\mathrm{M}(\cdot)$ to bound the overall objective (Cousins, 2021, 2022, 2023b). In this work, we show sharper bounds on the generalization error of malfare objectives, but we also seek to bound each group's generalization error.

Multitask learning There is overlap between group fair learning (GFL) and multitask learning (MTL). This work shows that GFL reduces generalization error for all groups (particularly smaller groups), which is essentially the motivation for MTL. In both cases, we have g distributions (per-group in GFL, per-task in MTL) and some objective that considers each distribution through $\mathrm{R}(h, \mathcal{D}_i)$. To our knowledge, there is no published work in multitask learning on objectives that treat tasks nonlinearly, i.e., the objective is always (Caruana, 1997; Zhang and Yang, 2018, 2021)

$$\hat{h} \doteq \mathop{\mathrm{argmin}}_{h \in \mathcal{H}} \sum_{i=1}^{g} \frac{1}{\boldsymbol{m}_i} \sum_{j=1}^{\boldsymbol{m}_i} \ell_i(h(\boldsymbol{x}_{i,j}), \boldsymbol{y}_{i,j}) \enspace .$$

Existing MTL analysis bounds generalization error by

[1] Constants vary between sources, depending on definitions and derivations. Our probabilistic statements use 2-tailed Hoeffding (1963) bounds and 3 applications of McDiarmid's (1989) inequality, with optimal bounded differences, for Rademacher averages with no absolute value inside the supremum.

considering all data at once (Zhang et al., 2020; Zhang and Yang, 2021); assuming m samples each for g groups, VC dimension, Rademacher averages, etc. bound total estimation error as $O\sqrt{\ln\frac{1}{\delta}/mg}$. Such methods do not apply in our setting, as we seek *per-group generalization bounds* and treat *nonlinear objectives*, thus ultimately we do not expect bounds of this order.

To pool or not to pool Some work directly addresses the tradeoff between training pooled versus separate models for groups. Dwork et al. (2018) define the *cost-of-coupling* as

$$\max_{\mathcal{D}}\left(\min_{h\in\mathcal{H}}\sum_{i=1}^{g}\mathrm{R}(h,\mathcal{D}_i)-\sum_{i=1}^{g}\min_{h\in\mathcal{H}}\mathrm{R}(h,\mathcal{D}_i)\right) \quad ,$$

i.e., worst-case difference between the sum risk of the optimal shared model $\hat{h}$, vs. sum risk of optimal per-group models $\hat{h}_{1:g}$. When this quantity is positive, training with pooled data may require tradeoffs in accuracy across groups. They then introduce *transfer learning* methods to train per-group classifiers $\hat{h}_{1:g}$ while leveraging available data where appropriate. Similarly to our work, this results in improved VC-theoretic groupwise bounds on generalization error than fully separated training. However, the goal of our learning framework is still to learn a joint model, avoiding thorny questions of disparate treatment. Wang et al. (2021) also examine the tradeoff, where the metric of interest or *benefit of splitting* is based on an egalitarian notion of fairness. They largely focus on the infinite-samples or known-distributions settings; however, they provide VC-theoretic generalization bounds on the benefit of splitting. These are necessarily worst-case (over possible distributions), and specific to binary classification, whereas we provide data-dependent Rademacher average bounds applicable to a broad range of supervised and unsupervised settings.

3 BOUNDING GENERALIZATION ERROR IN FAIR TRAINING

The generalization error analysis of section 2.1 does not take into account the fact that learning is not equally likely to produce any $h \in \mathcal{H}$. In this section, we present a sharper analysis that reflects this, both in per-group generalization error bounds, and in the overall generalization error of a malfare objective.

Our approach is to take the core idea of localization (Bartlett et al., 2005) — restricting the function class of interest to a subset that with high probability contains the function that will be learned — and generalize it to apply in multi-group fair learning settings. In Section 3.1 we argue that, for each group i, with high probability, the learned function $\hat{h}$ belongs to some $\mathcal{H}_i^* \subseteq \mathcal{H}$. We bound generalization error over $\mathcal{H}_i^*$, with $\hat{\mathfrak{R}}_{\boldsymbol{m}_i}(\mathcal{H}_i^*, \boldsymbol{z}_i)$, where often $\hat{\mathfrak{R}}_{\boldsymbol{m}_i}(\mathcal{H}_i^*, \boldsymbol{z}_i) \ll \hat{\mathfrak{R}}_{\boldsymbol{m}_i}(\mathcal{H}, \boldsymbol{z}_i)$.

The analysis depends on the group index i, since while analyzing group i, we can treat the training samples $\boldsymbol{z}_j$ as constant for each $j \neq i$, but the class $\mathcal{H}_i^*$ must not depend on $\boldsymbol{z}_i$ for vital technical reasons (see proof of theorem 4; we require $\mathcal{H}_i^*$ to be established independently from $\boldsymbol{z}_i$ in $\hat{\mathfrak{R}}_{\boldsymbol{m}_i}(\mathcal{H}_i^*, \boldsymbol{z}_i)$). We thus establish a theoretical hypothesis class that directly depends on the training sample for each $j \neq i$, but depends on the distribution for group i instead of its training sample.

Unfortunately, $\mathcal{H}_i^*$ is a theoretical object (not actually known, as it depends on $\mathcal{D}_i$). Thus, we have little recourse but to relax to dependence on purely empirical quantities. We thus establish in Section 3.2 an empirical class $\hat{\mathcal{H}}_i$, which depends on $\boldsymbol{z}_i$ instead of $\mathcal{D}_i$. At first glance this seems to violate core statistical precepts, but through careful construction, we show that $\hat{\mathcal{H}}_i$ acts merely as a probabilistic proxy for $\mathcal{H}_i^*$.

Finally, we must actually *estimate* the relevant Rademacher bounds. In Section 3.3 we illustrate how this can be done for linear hypothesis classes using Monte-Carlo Rademacher averaging.

3.1 Theoretical Restricted Classes

When bounding the generalization error of group i, we want to construct a restricted hypothesis class leveraging information given by the remaining group samples, in particular their empirical risks. However, we can't directly use the group i sample $\boldsymbol{z}_i$, so instead we bound empirical risk $\hat{\mathrm{R}}(h, \boldsymbol{z}_i)$ in terms of $\mathrm{R}(h, \mathcal{D}_i)$. Intuitively, we want this restricted class to be the set of all $h \in \mathcal{H}$ that could reasonably be the function we learn from *all data* (the empirical malfare minimizer $\hat{h}$), where the restricted class is constructed after observing only the data $\boldsymbol{z}_j$ for all groups $j \neq i$.

Similar techniques are common in learning theory and the study of localization, where a *theoretical class* is constructed based on the (unknown) distribution(s), and subsequently an *empirical class* that is with high probability a superset which can be built from the data. Our approach, however, is unique in that it is in some sense *half-empirical*, as the theoretical class depends on the distribution $\mathcal{D}_i$ of one group, and the samples $\boldsymbol{z}_j$ from all groups $j \neq i$. We do this instead of constructing a "fully theoretical" class using only the distributions $\mathcal{D}_{1:g}$, as well as an empirical variant based on all training samples $\boldsymbol{z}_{1:g}$, which would be substantially larger.

First, let $\boldsymbol{\varepsilon}_i \doteq r\sqrt{\frac{\ln\frac{1}{\delta}}{2\boldsymbol{m}_i}}$ and $\boldsymbol{\eta}_i \doteq 2\mathfrak{R}_{\boldsymbol{m}_i}(\ell \circ \mathcal{H}, \mathcal{D}_i) + \boldsymbol{\varepsilon}_i$, where $\boldsymbol{m}_i$ is the sample size for group i and r is the range of loss values in $\ell \circ \mathcal{H}$. Recall that the empirical malfare minimization task is to select

$$\hat{h} \doteq \operatorname*{argmin}_{h'\in\mathcal{H}} \mathrm{M}(j \mapsto \hat{\mathrm{R}}(h', \boldsymbol{z}_j)) \quad ,$$

but since we can't observe sample i yet, we (pessimistically) upper-bound the objective value (w.h.p.) as

$$\inf_{h'\in\mathcal{H}} \mathrm{M}(j\mapsto \hat{\mathrm{R}}(h',\boldsymbol{z}_j);\boldsymbol{w}) \le$$
$$\inf_{h'\in\mathcal{H}} \mathrm{M}\left(j\mapsto \begin{cases} j\neq i & \hat{\mathrm{R}}(h',\boldsymbol{z}_j) \\ j=i & \mathrm{R}(h',\mathcal{D}_i)+\boldsymbol{\varepsilon}_i \end{cases}\right) \ ,$$

and (optimistically) lower-bound the empirical malfare of all $h\in\mathcal{H}$, w.h.p. simultaneously, as

$$\mathrm{M}(j\mapsto \hat{\mathrm{R}}(h,\boldsymbol{z}_j);\boldsymbol{w}) \ge$$
$$\mathrm{M}\left(j\mapsto \begin{cases} j\neq i & \hat{\mathrm{R}}(h,\boldsymbol{z}_j) \\ j=i & \mathrm{R}(h,\mathcal{D}_i)-\boldsymbol{\eta}_i \end{cases}\right) \ .$$

Via this analysis, we then construct our theoretical class, which with high probability shall contain the empirical malfare minimizer $\hat{h}$, as the subset $\mathcal{H}^*_i\subseteq\mathcal{H}$ constrained to h such that

$$\mathrm{M}\left(j\mapsto \begin{cases} j\neq i & \hat{\mathrm{R}}(h,\boldsymbol{z}_j) \\ j=i & \mathrm{R}(h,\mathcal{D}_i)-\boldsymbol{\eta}_i \end{cases}\right) \le$$
$$\inf_{h'\in\mathcal{H}} \mathrm{M}\left(j\mapsto \begin{cases} j\neq i & \hat{\mathrm{R}}(h',\boldsymbol{z}_j) \\ j=i & \mathrm{R}(h',\mathcal{D}_i)+\boldsymbol{\varepsilon}_i \end{cases}\right) \ . \tag{7}$$

This construction is valid (formalized in theorem 3), as we took any h that optimistically could outperform a pessimistic estimate of the empirical objective.

The LHS is a "best case" estimate of the empirical malfare of a candidate hypothesis, whereas the RHS is a "worst case" estimate of the minimal empirical malfare, because we want our restricted hypothesis class to be large enough to contain any $h\in\mathcal{H}$ that might be the empirical malfare minimizer. In particular, the LHS uses a Rademacher average bound (3), as the bound must apply to all $h\in\mathcal{H}$, but a simple tail-bound term (2) suffices on the RHS, as we are comparing to a bound involving some specific h' (not dependent on the data $\boldsymbol{z}_i$).

Intuitively, for utilitarian malfare, $\hat{\mathcal{H}}_i$ describes models that definitely perform well for groups $j\neq i$, and will probably perform well for group i. Some malfare functions, such as power-means, are undefined for negative risk values, and the LHS risk lower bounds $\mathrm{R}(h,\mathcal{D}_i)-\boldsymbol{\eta}_i$ may be negative. However, if we assume risk (or loss) is nonnegative, we may use the risk lower bound $\max(0,\mathrm{R}(h,\mathcal{D}_i)-\boldsymbol{\eta}_i)$, which preserves convexity, continuity, and even differentiability if $p<\infty$ except around the point $\mathbf{0}$.

Observe now that, conditioning on $\boldsymbol{z}_j$ for each $j\neq i$, with high probability over choice of $\boldsymbol{z}_i$, empirical malfare minimization yields some $\hat{h}\in\mathcal{H}^*_i$. Therefore, for all intents and purposes, learning occurs over $\mathcal{H}^*_i$, and we may thus use Rademacher averages over this restricted class to bound generalization error for group i. Formally put, we have the following result.

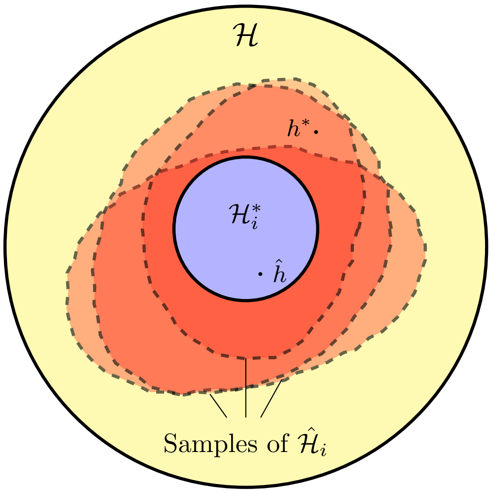

Figure 1: Visualization of unrestricted class $\mathcal{H}$, theoretical restricted class $\mathcal{H}^*_i$, and samples of empirical restricted class $\hat{\mathcal{H}}_i$ (varying $\boldsymbol{z}_i$). One possible empirical malfare minimizer $\hat{h}$ (contained by $\hat{\mathcal{H}}_i$ and $\mathcal{H}^*_i$ with high probability), as well as the true malfare minimzer h^* (which may fall outside of $\mathcal{H}^*_i$ or $\hat{\mathcal{H}}_i$ due to overfitting to groups other than i) are also shown.

Theorem 3 (Theoretical Group-Regularized Malfare Bounds)**.** *Suppose a monotonic* malfare function $\mathrm{M}(\cdot):\mathbb{R}^g\to\mathbb{R}$, hypothesis class $\mathcal{H}\subseteq\mathcal{X}\to\mathcal{Y}'$, loss function $\ell:\mathcal{Y}'\times\mathcal{Y}\to\mathbb{R}$, *per-group* distributions $\mathcal{D}_{1:g}$ *over* $\mathcal{X}\times\mathcal{Y}$, *and per-group* samples $\boldsymbol{z}_{1:g}$, *with* $\boldsymbol{z}_j\sim\mathcal{D}_j^{\boldsymbol{m}_j}$ *for each group j. Fix any group index i, and take $\mathcal{H}^*_i$ defined as in (7). The following then hold.*

1. *With probability at least $1-2\delta$ over choice of $\boldsymbol{z}_i$, it holds that $\hat{h}\in\mathcal{H}^*_i$.*
2. *With probability at least $1-4\delta$ over choice of $\boldsymbol{z}_i$,*
$$\left|\mathrm{R}(\hat{h},\mathcal{D}_i)-\hat{\mathrm{R}}(\hat{h},\boldsymbol{z}_i)\right| \le 2\mathfrak{R}_{\boldsymbol{m}_i}(\ell\circ\mathcal{H}^*_i,\mathcal{D}_i)+\boldsymbol{\varepsilon}_i \ .$$

3.2 Empirical Restricted Classes

$\mathcal{H}^*_i$ is an object only of theoretical interest (it is not actually known, since it depends on $\mathcal{D}_i$). Consequently, without more information, $\mathfrak{R}_{\boldsymbol{m}_i}(\ell\circ\mathcal{H}^*_i,\mathcal{D}_i)$, and thus the bounds of theorem 3, can not be computed. We remedy this issue here, relaxing dependence on the distribution $\mathcal{D}_i$ by replacing it with dependence on the training sample $\boldsymbol{z}_i$ and thus establishing a new *empirically restricted hypothesis class.*

Note that we can't simply substitute $\hat{\mathrm{R}}(h,\boldsymbol{z}_i)$ for $\mathrm{R}(h,\mathcal{D}_i)$, as theorem 3 clearly requires the restricted hypothesis class $\mathcal{H}^*_i$ to be fixed before observing the training data $\boldsymbol{z}_i$. We account for this by indirectly using $\hat{\mathrm{R}}(h,\boldsymbol{z}_i)$ to bound $\mathrm{R}(h,\mathcal{D}_i)$. In particular, take $\hat{\boldsymbol{\eta}}_i\doteq 2\hat{\mathfrak{R}}_{\boldsymbol{m}_i}(\ell\circ\mathcal{H},\boldsymbol{z}_i)+2\boldsymbol{\varepsilon}_i$, and take $\boldsymbol{\varepsilon}_i\doteq r\sqrt{\frac{\ln\frac{1}{\delta}}{2\boldsymbol{m}_i}}$, as in (3). Now, we construct our empirical class $\hat{\mathcal{H}}_i$, which with high probability shall contain the theoretical class $\mathcal{H}_i^*$, as the subset $\hat{\mathcal{H}}_i \subseteq \mathcal{H}$ constrained to h such that

$$\mathrm{M}\left(j \mapsto \begin{cases} j \neq i & \hat{\mathrm{R}}(h, \boldsymbol{z}_j) \\ j = i & \hat{\mathrm{R}}(h, \boldsymbol{z}_i) - 2\hat{\boldsymbol{\eta}}_i \end{cases} \right) \leq$$
$$\inf_{h' \in \mathcal{H}} \mathrm{M}\left(j \mapsto \begin{cases} j \neq i & \hat{\mathrm{R}}(h', \boldsymbol{z}_j) \\ j = i & \hat{\mathrm{R}}(h', \boldsymbol{z}_i) + 2\boldsymbol{\varepsilon}_i \end{cases} \right) \quad . \tag{8}$$

Note that (8) matches (7), except risks and Rademacher averages are bounded in terms of their empirical counterparts. In particular, on the LHS, w.h.p., for all $h \in \mathcal{H}$ it holds $\hat{\mathrm{R}}(h, \boldsymbol{z}_i) - 2\hat{\boldsymbol{\eta}}_i \leq \mathrm{R}(h, \mathcal{D}_i) - \boldsymbol{\eta}_i$, and on the RHS, w.h.p., $\hat{\mathrm{R}}(h', \boldsymbol{z}_i) + 2\boldsymbol{\varepsilon}_i \geq \mathrm{R}(h', \mathcal{D}_i) + \boldsymbol{\varepsilon}_i$. Figure 1 visualizes the difference between $\hat{\mathcal{H}}_i$ and $\mathcal{H}_i^*$, as well as other key players.

Observe now that, with high probability, $\mathcal{H}_i^* \subseteq \hat{\mathcal{H}}_i$, therefore we can employ the theoretical properties of $\mathcal{H}_i^*$ while being able to compute everything from a sample using $\hat{\mathcal{H}}_i$. The following theorem makes precise this statement, and should be viewed as an *empirical counterpart* to theorem 3.

Theorem 4 (Empirical Group-Regularized Malfare Bounds)**.** *Suppose as in theorem 3. The following then hold for $\hat{\mathcal{H}}_i$ defined as in (8).*

1. With probability at least $1 - 4\delta$ over choice of $\boldsymbol{z}_i$, it holds that $\hat{h} \in \mathcal{H}_i^* \subseteq \hat{\mathcal{H}}_i$.
2. With probability at least $1 - 6\delta$, it holds that

$$\left| \mathrm{R}(\hat{h}, \mathcal{D}_i) - \hat{\mathrm{R}}(\hat{h}, \boldsymbol{z}_i) \right| \leq 2\hat{\mathfrak{R}}_{\boldsymbol{m}_i}(\ell \circ \hat{\mathcal{H}}_i, \boldsymbol{z}_i) + 2\boldsymbol{\varepsilon}_i \quad .$$

Theorem 4 satisfies our primary goal of showing per-group generalization bounds for fair learning that leverage information from other groups. In particular, when $\hat{\mathcal{H}}_i \subset \mathcal{H}$, we obtain sharper generalization bounds, which quantifies the intuition that training a shared model is less susceptible to overfitting than training per-group models. Theorem 4 part 2 should be contrasted with (4), which gives a similar guarantee using Rademacher averages of the *unrestricted class* $\mathcal{H}$. Corollary 5 now applies these bounds to improve the state-of-the-art generalization guarantees for (nonlinear) malfare objectives, which would otherwise depend on Rademacher averages of $\hat{\mathcal{H}}_i$ rather than $\mathcal{H}$, cf. (5).

Corollary 5 (Empirical Malfare Generalization Bounds)**.** *Suppose as in theorem 4. Suppose also that there exists some $\lambda > 0$ and norm $\|\cdot\|_{\mathrm{M}}$ such that $\mathrm{M}(\cdot)$ is λ-$\|\cdot\|_{\mathrm{M}}$ Lipschitz continuous, i.e., $\forall \mathcal{S}, \mathcal{S}'$: $\mathrm{M}(\mathcal{S} + \mathcal{S}') \leq \mathrm{M}(\mathcal{S}) + \lambda \|\mathcal{S}'\|_{\mathrm{M}}$. We then have:*

1. With probability at least $1 - 5g\delta$, the true malfare of $\hat{h}$ is bounded by

$$\begin{aligned} &\mathrm{M}(j \mapsto \mathrm{R}(\hat{h}, \mathcal{D}_j)) \\ &\quad \leq \mathrm{M}\Big(j \mapsto \hat{\mathrm{R}}(\hat{h}, \boldsymbol{z}_j) + 2\hat{\mathfrak{R}}_{\boldsymbol{m}_j}(\hat{\mathcal{H}}_j, \boldsymbol{z}_j) + 2\boldsymbol{\varepsilon}_j\Big) \\ &\quad \leq \mathrm{M}\Big(j \mapsto \hat{\mathrm{R}}(\hat{h}, \boldsymbol{z}_j)\Big) + \lambda \Big\| j \mapsto 2\hat{\mathfrak{R}}_{\boldsymbol{m}_j}(\hat{\mathcal{H}}_j, \boldsymbol{z}_j) + 2\boldsymbol{\varepsilon}_j \Big\|_{\mathrm{M}} \ . \end{aligned}$$

2. With probability at least $1 - 6g\delta$, we bound the suboptimality of $\hat{h}$ as

$$\begin{aligned} &\mathrm{M}\left(j \mapsto \mathrm{R}(\hat{h}, \mathcal{D}_j)\right) \\ &\qquad \leq \mathrm{M}\left(j \mapsto \mathrm{R}(h^*, \mathcal{D}_j) + 2\hat{\mathfrak{R}}_{\boldsymbol{m}_j}(\hat{\mathcal{H}}_j, \boldsymbol{z}_j) + 3\boldsymbol{\varepsilon}_j\right) \\ &\implies \left| \mathrm{M}\left(j \mapsto \mathrm{R}(h^*, \mathcal{D}_j)\right) - \mathrm{M}\left(j \mapsto \mathrm{R}(\hat{h}, \mathcal{D}_j)\right) \right| \\ &\qquad \leq \lambda \Big\| j \mapsto 2\hat{\mathfrak{R}}_{\boldsymbol{m}_j}(\hat{\mathcal{H}}_j, \boldsymbol{z}_j) + 3\boldsymbol{\varepsilon}_j \Big\|_{\mathrm{M}} \ . \end{aligned}$$

The first inequality of parts 1 & 2 of corollary 5 is sharper, but the second is generally more analytically convenient. In particular, any power-mean malfare function $\mathrm{M}_p(\cdot; \boldsymbol{w})$ obeys

$$\mathrm{M}_p(\mathcal{S} + \mathcal{S}'; \boldsymbol{w}) - \mathrm{M}_p(\mathcal{S}; \boldsymbol{w}) \leq \mathrm{M}_p(\mathcal{S}'; \boldsymbol{w}) \leq \|\mathcal{S}'\|_\infty \ , \tag{9}$$

thus we bound malfare generalization error in terms of the generalization error of each group.

Naturally, one may ask how sharp this localization strategy is. We now show an example where theorem 4 improves slow $\mathbf{O}(\frac{1}{\sqrt{m}})$ convergence rates to fast $\mathbf{O}(\frac{1}{m})$ convergence rates. Consider unit-range 0-dimensional linear regression, i.e., mean estimation under square loss ℓ, with $g = 1$. Thus we have

$$\ell \circ \mathcal{H}_r = \left\{ \ell(h_c(x), y) = (c - y)^2 \,\middle|\, c \in [-r, r] \right\}$$

with $r = 1$. Take *constant probability distribution* $\mathcal{D} = 0$, thus $\boldsymbol{y} = \boldsymbol{0}$. From random walk theory, we have

$$\begin{aligned} \hat{\mathfrak{R}}_m(\ell \circ \mathcal{H}_r, \boldsymbol{y}) &= \mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\left[\sup_{c \in [-r,r]} \frac{1}{m} \sum_{i=1}^m \boldsymbol{\sigma}_i (\boldsymbol{y}_i - c)^2 \right] \\ &= \mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\left[\sup_{c \in [-r,r]} \frac{1}{m} \sum_{i=1}^m \boldsymbol{\sigma}_i c^2 \right] \\ &= \frac{r^2}{2} \mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\left[\left| \frac{1}{m} \sum_{i=1}^m \boldsymbol{\sigma}_i \right| \right] \approx r^2 \sqrt{\frac{1}{2\pi m}} \ . \end{aligned}$$

To construct $\hat{\mathcal{H}}$, observe that we have $\hat{\mathrm{R}}(c, \boldsymbol{y}) = c^2$, thus via (8) we restrict s.t. $c^2 \leq 4\hat{\mathfrak{R}}_m(\ell \circ \mathcal{H}_r, \boldsymbol{y}) + 6\varepsilon \approx \sqrt{\frac{8}{\pi m}} + 6\sqrt{\frac{\ln \frac{1}{\delta}}{2m}} \implies |c| \leq r \in \mathbf{\Theta}\sqrt[4]{\frac{1}{m}}$. We thus have

$$\hat{\mathfrak{R}}_m(\ell \circ \hat{\mathcal{H}}, \boldsymbol{y}) \approx r^2 \sqrt{\tfrac{1}{2\pi m}} \in \mathbf{\Theta}\left(\tfrac{1}{m}\right) \ ,$$

which asymptotically improves $\hat{\mathfrak{R}}_m(\ell \circ \mathcal{H}, \boldsymbol{y}) \approx \sqrt{\frac{1}{2\pi m}}$.

3.3 Monte-Carlo Rademacher Averages of Linear Hypothesis Classes

We now present a method to estimate Rademacher averages for linear hypothesis classes using Monte-Carlo sampling. We start by noting that, in general if $\ell(\hat{y}, y) = f(g(\hat{y}, y))$ and f is λ-Lipschitz-continuous, then we have, for any $\boldsymbol{z} \in (\mathcal{X} \times \mathcal{Y})^m$, that

$$\hat{\mathfrak{R}}_m(\ell \circ \mathcal{H}, \boldsymbol{z}) \leq \lambda \hat{\mathfrak{R}}_m(g \circ \mathcal{H}, \boldsymbol{z}) \ . \tag{10}$$

For this reason, we formulate the Rademacher averages of both linear least-squares regression and logistic regression as follows. Take $\mathcal{H} = \{h_{\boldsymbol{\beta}}(\boldsymbol{x}) \doteq \boldsymbol{\beta} \cdot \boldsymbol{x} \,|\, \boldsymbol{\beta} \in \boldsymbol{B}\}$ and loss function $\ell(\hat{y}, y) = f(g(\hat{y}, y))$, where for least-squares regression, $g(\hat{y}, y) = \hat{y} - y$ and $f(u) = u^2$. This is λ-Lipschitz continuous, assuming bounded $\boldsymbol{B}$, $\mathcal{X}$, and $\mathcal{Y}$, with $\lambda = 2\sup_{\boldsymbol{B},\mathcal{Y},\mathcal{X}} |\boldsymbol{x} \cdot \boldsymbol{\beta} - y|$.[2] For logistic regression, in which $\mathcal{Y} = \pm 1$, we have $g(\hat{y}, y) = \hat{y} \cdot y$ and $f(u) = \ln(1 + \exp(u))$, which is 1-Lipschitz.

Estimation Standard methods for bounding Rademacher averages of linear regression classes start by bounding the Rademacher average of $\mathcal{H}$ itself (Shalev-Shwartz and Ben-David, 2014). However, this method is loose (Cousins and Riondato, 2020), and seems especially so for irregular weight spaces (i.e., those not defined by simple p-norms), which known analytic methods can not handle.

Instead, we directly estimate the Rademacher average of the function family $g \circ \mathcal{H}$ directly using Monte-Carlo estimation. That is to say, given sampled Rademacher random variables $\boldsymbol{\sigma} \in (\pm 1)^{n \times m}$ and data sample $\boldsymbol{z} \in (\mathcal{X} \times \mathcal{Y})^m$, we compute

$$\hat{\mathfrak{R}}^n_m(g \circ \mathcal{H}, \boldsymbol{z}; \boldsymbol{\sigma}) \doteq \frac{1}{n}\sum_{k=1}^{n} \sup_{\boldsymbol{\beta} \in W} \frac{1}{m}\sum_{j=1}^{m} \boldsymbol{\sigma}_{k,j} g(\boldsymbol{x}_j \cdot \boldsymbol{\beta}, \boldsymbol{y}_j) \ . \tag{11}$$

This fully data-dependent method gracefully tolerates arbitrary data distributions and parameter spaces, and is loose only in a small amount of Monte-Carlo error and the contraction inequality (Cousins and Riondato, 2020). In practice, we use $\lambda \hat{\mathfrak{R}}^n_{\boldsymbol{m}_i}(g \circ \mathcal{H}, \boldsymbol{z}_i; \boldsymbol{\sigma})$ as a plug-in estimate of $\lambda \hat{\mathfrak{R}}_{\boldsymbol{m}_i}(g \circ \mathcal{H}, \boldsymbol{z}_i)$, which then bounds $\hat{\mathfrak{R}}_{\boldsymbol{m}_i}(\ell \circ \mathcal{H}, \boldsymbol{z}_i)$ via (10). We similarly estimate and bound Rademacher averages over our restricted hypothesis classes as $\lambda \hat{\mathfrak{R}}^n_{\boldsymbol{m}_i}(g \circ \hat{\mathcal{H}}_i, \boldsymbol{z}_i; \boldsymbol{\sigma})$.

Lemma 6 (Convex Optimization for Monte-Carlo Rademacher Averages)**.** *Suppose the* parameter space $\boldsymbol{B}$ *of* $\mathcal{H}$ *is a convex set,* loss $\ell(h_{\boldsymbol{\beta}}(\boldsymbol{x}), y)$ *is convex in* $\boldsymbol{\beta} \in \boldsymbol{B}$ *for all* $\boldsymbol{x} \in \mathcal{X}$*,* $y \in \mathcal{Y}$*, and* malfare $\mathrm{M}(\cdot) : \mathbb{R}^g \to \mathbb{R}$ *is quasiconvex and monotonically increasing in each argument. Then the parameter spaces of* $\hat{\mathcal{H}}_i$ *and* $\mathcal{H}^*_i$ *are convex sets.*

Moreover, if $g \circ \mathcal{H}$ *is an affine function family, then* $\hat{\mathfrak{R}}^n_m(g \circ \hat{\mathcal{H}}_i, \boldsymbol{z}_i; \boldsymbol{\sigma})$ *reduces to maximizing a linear function over a convex set. Similarly, if we strengthen the quasiconvexity assumption on* $\mathrm{M}(\cdot)$ *to* convexity, *then EMM reduces to minimizing a convex objective over the convex set* $\boldsymbol{B}$*.*

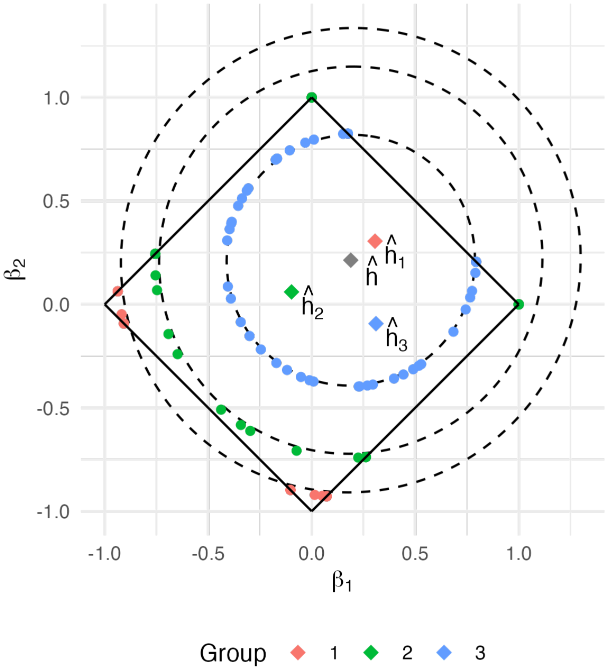

Figure 2: Rademacher average samples in the parameter space of $\hat{\mathcal{H}}_i$ for each group $i \in \{1, 2, 3\}$.

Table 1: Sample sizes $\boldsymbol{m}_{1:3}$, parameter vectors $\boldsymbol{\beta}_{1:3}$, and Monte-Carlo empirical Rademacher averages (MCERA) for both $\mathcal{H}$ and $\hat{\mathcal{H}}_{1:3}$.

Group	$\boldsymbol{m}_i$	True $\boldsymbol{\beta}$	MCERA	
ID			$\mathcal{H}$	$\hat{\mathcal{H}}_i$
1	6500	(0.3,0.3)	0.047	0.046
2	3000	(-0.1,0.1)	0.075	0.046
3	500	(0.3,0)	0.183	0.135

Visualizing $\hat{\mathcal{H}}_i$ in least-squares regression For least-squares regression, under utilitarian malfare, the restricted hypothesis constraint of $\hat{\mathcal{H}}_i$ is an ellipsoid (under egalitarian welfare, it is an *intersection* of ellipsoids). We visualize a simple example in figure 2, with parameters and results described in table 1.

Taking $\boldsymbol{B} \doteq \left\{\boldsymbol{\beta} \in \mathbb{R}^2 \,\middle|\, \|\boldsymbol{\beta}\|_1 \leq 1\right\}$ to be the unit ℓ_1 ball, we sample $(\boldsymbol{x}, y)$ as $\boldsymbol{x} \sim \mathrm{Unif}([-1, 1]^2)$, $y = \boldsymbol{x} \cdot \boldsymbol{\beta}_i + \mathrm{Unif}([-1, 1])$, where each group has slightly different data generating parameters $\boldsymbol{\beta}_i$. In figure 2, taking $\delta = 0.1$, we plot the $n = 100$ values of $\boldsymbol{\beta}$ which realize each supremum of (11) for some Rademacher sample $\boldsymbol{\sigma}_k$. These points necessarily lie on either (the corner of) the ℓ_1 constraint boundary of $\boldsymbol{B}$ or the restricted hypothesis constraint boundary of $\hat{\mathcal{H}}_i$, illustrated by the concentric ellipses, which represent constant upper-bounds of weighted utilitarian malfare over the whole dataset and are centered around $\hat{h}$.

[2] In practice, we compute the Lipschitz constant over $\mathcal{H}$, rather than over $\hat{\mathcal{H}}_i \subseteq \mathcal{H}$, which would require computing the diameter of $\hat{\mathcal{H}}_i$ or bounding the range of $g \circ \hat{\mathcal{H}}_i$.

Note that for the smallest group, 3, the fact that $\hat{h}$ must perform well on the other two groups under weighted utilitarian malfare shrinks $\hat{\mathcal{H}}_3$ significantly. However, the generalization bound over the largest group, 1, is not significantly improved when taken over $\hat{\mathcal{H}}_1$.

4 EXPERIMENTS

We illustrate the utility of our results with some experiments. Our approach is to construct an example dataset where we can demonstrate a clear benefit (to minority groups) to pooled training, and then show how our refined generalization bounds are in fact sharper than standard Rademacher bounds. We do this by assuming that the individual distributions of the groups are similar enough that, for underrepresented minority groups, pooled training reduces generalization error.

Our experiments are based on a binary logistic regression task with 3 groups. Suppose the unit ℓ_∞ ball domain, i.e., $\mathcal{X} = [-1, 1]^{15}$, binary label space $\mathcal{Y} = \pm 1$, and parameter space $\boldsymbol{B} \doteq \left\{\boldsymbol{\beta} \in \mathbb{R}^{15} \,\middle|\, \|\boldsymbol{\beta}\|_1 \leq 15\right\}$. For each group i, we generate samples $(\boldsymbol{x}, y)$ with $\boldsymbol{x} \sim \mathrm{Unif}(\mathcal{X})$, $\mathbb{P}(y = 1) = \mathrm{logistic}(\boldsymbol{x} \cdot \boldsymbol{\beta}_i + \xi)$, with noise $\xi \sim \mathcal{N}(0, 0.1)$, for $\mathrm{logistic}(u) = \frac{1}{1+\exp(-u)}$.

We assume groupwise data generating parameters and a constant proportional composition of the full training sample as in table 2. Notably, the data generating model for groups 1 and 3 are very similar, but there is always much more data available for group 1.

In figure 3, we plot the average test risk of each group i over 7 independent runs for malfare-minimizing models $\hat{h}$ or for risk-minimizing models $\hat{h}_i$ as a function of total training sample size, where test risk is computed from a held-out test set with 20,000 samples for each group. We observe that pooled models almost always have lower per-group test risks than the separately-trained models $\hat{h}_i$ on the minority groups (2 and 3), which we attribute to the regularizing effect of pooled training overcoming the small discrepancies between the data generating parameters of each group (see table 2). While the above describes small-sample behavior, for sufficient sample sizes, per-group models should dominate shared models, and we do observe this for group 3 with the maximum sample size of $32768 \cdot 0.05 \approx 1638$.

We then compute the bounds derived from Monte-Carlo Rademacher averages (with $\delta = 0.1$) over samples $\boldsymbol{z}_i$ over both $\mathcal{H}$ and $\hat{\mathcal{H}}_i$ for each i (figure 4). Since the bounds derived from Rademacher averages over $\mathcal{H}$ essentially function as bounds on the generalization error of the separately trained models, the fact that the bound over $\hat{\mathcal{H}}_i$ is tighter correctly suggests that pooled training is better for the minority groups in this scenario, especially when using egalitarian training.

Table 2: Data generating parameters for logistic regression experiments.

	Data proportion	True parameters
Group 1	75%	$\beta_i = 0.3$
Group 2	20%	$\beta_i = 0.1$
Group 3	5%	$\beta_i = 0.2$

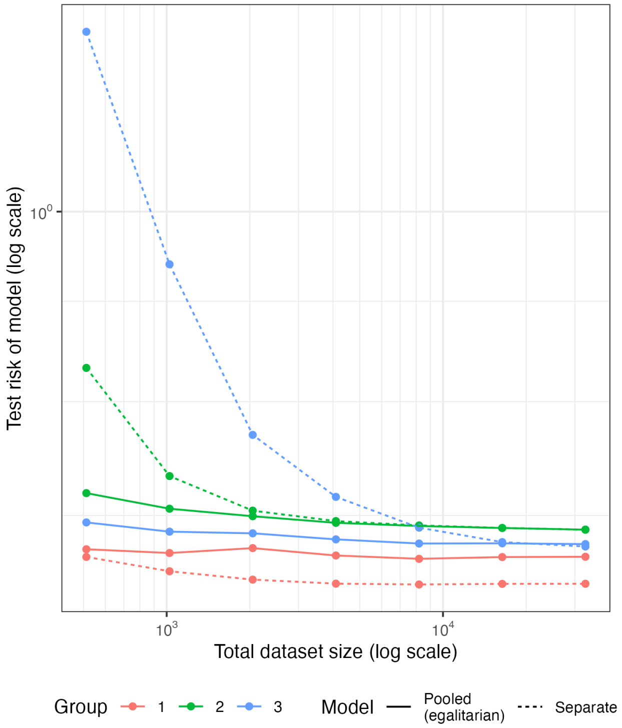

Figure 3: Average test risk of pooled and separately trained models on three groups (see table 2).

In the utilitarian case, we see that initially $\hat{\mathcal{H}}_i$ bounds match $\mathcal{H}$ bounds, but for sufficiently large sample sizes, they diverge. In the egalitarian case, $\hat{\mathcal{H}}_i$ bounds are always better than $\mathcal{H}$ bounds, and they appear to decay at an asymptotically greater rate (slope on the log-log plot), reaching an order of magnitude improvement in the case of the largest group (group 1). This suggests that our bounds characterize generalization error substantially more sharply than the naïve method.

5 CONCLUSION

We show that fair learning, like multitask learning, has a *regularizing effect*, reducing overfitting to each group as compared to per-group models trained solely on their data. Concretely, we show that, from the perspective of each group, fair-learning (empirical malfare minimization) effectively occurs over some *restricted hypothesis class*, and we the bound generalization error of each group's risk in terms of their Rademacher averages over these restricted classes. This technique yields refined

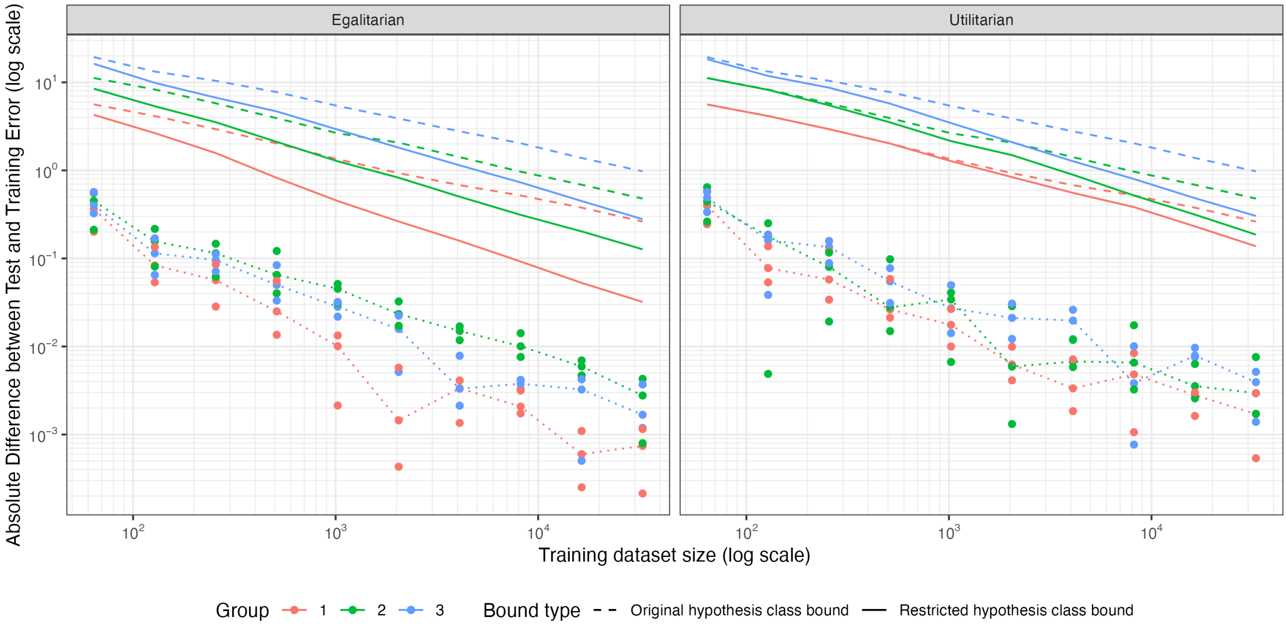

Figure 4: Generalization error bounds derived from original hypothesis class $\mathcal{H}$ and restricted hypothesis classes $\hat{\mathcal{H}}_i$, compared with shared model $\hat{h}$ train-test gap over 7 independent runs, with quartiles and median trend lines.

generalization bounds, not just for the overall learning, task, but also for the risk *of each individual group.*

Such bounds are of particular importance in learning settings where minority groups often suffer poor model performance (Mehrabi et al., 2021), such as medical ML (Obermeyer et al., 2019) and facial recognition (Buolamwini and Gebru, 2018; Cavazos et al., 2020). Moreover, in critical systems, having provable guarantees on the generalization error *of each task*, rather than just the overall generalization error, can greatly improve reliability and user trust. This is also valuable in multi-task learning settings, where task fairness and task-specific bounds are of interest, e.g., in distributionally-robust LLMs (Oren et al., 2019).

While the contributions of this paper are theoretical, our setting is practically motivated. Understanding the generalization error of each group allows modelers to make better-informed decisions, particularly regarding minority groups. Generalization bounds for a group-specific model $\hat{h}_i$ and a shared model $\hat{h}$ can be used to bound risk for group i, which can be used for *model selection* (i.e., group i can select between $\hat{h}$ and $\hat{h}_i$ with confidence). It is known that, given infinite data, individual models are always preferable, and the degree of suboptimality of a shared model can be bounded using transfer learning techniques; however, for data-hungry models, in particular with sparse data for minority groups, a better understanding of the interplay between generalization error and the negative impacts of majority group data on minority group performance are vital.

We also envision more sophisticated applications of our bounds. For example, if some smaller groups are more similar to minority group i than a majority group, a shared model $\hat{h}$ optimizing, say, *utilitarian malfare*, may perform poorly for group i, but perhaps a better-performing $\hat{h}'$ would arise from optimizing a *more egalitarian* malfare function (i.e., higher p power-mean), or one that emphasizes similar groups (through the weights vector $\boldsymbol{w}$). Group-fair learning methods can be combined with other aspects of model selection, such as feature and hyperparameter selection, where the bias-variance tradeoff plays a significant role. Our bounds indicate that we can provably learn a more complex shared model without overfitting, and our analysis enables rigorous model selection guarantees , both for individual group risks and for malfare objectives. We are hopeful that future work explores these model-search questions and other applications of our methods.

Acknowledgments

This research was made possible in part by the generous support of the Ford Foundation and the MacArthur Foundation. Cyrus Cousins also wishes to acknowledge the Center for Data Science at the University of Massachusetts Amherst, where part of this work was conducted under a postdoctoral fellowship. We would also like to thank Kweku Kwegyir-Aggrey for his insights and feedback on early versions of this work.

Bibliography

Abernethy, J. D., Awasthi, P., Kleindessner, M., Morgenstern, J., Russell, C., and Zhang, J. (2022). Active sampling for min-max fairness. In *International Conference on Machine Learning*, volume 162.

Agrawal, A., Verschueren, R., Diamond, S., and Boyd, S. (2018). A rewriting system for convex optimization problems. *Journal of Control and Decision*, 5(1):42–60.

Barman, S., Bhaskar, U., Krishna, A., and Sundaram, R. G. (2020). Tight approximation algorithms for p-mean welfare under subadditive valuations. *Leibniz International Proceedings in Informatics, LIPIcs*, 173.

Bartlett, P. L., Bousquet, O., and Mendelson, S. (2005). Local Rademacher complexities. *The Annals of Statistics*, 33(4):1497–1537.

Bartlett, P. L. and Mendelson, S. (2002). Rademacher and Gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3(Nov):463–482.

Bossert, W. and Kamaga, K. (2020). An axiomatization of the mixed utilitarian-maximin social welfare orderings. *Economic Theory*, 69(2):451–473.

Boyd, S. P. and Vandenberghe, L. (2004). *Convex optimization*. Cambridge university press.

Buolamwini, J. and Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*, pages 77–91. PMLR.

Caruana, R. (1997). Multitask learning. *Machine learning*, 28:41–75.

Cavazos, J. G., Phillips, P. J., Castillo, C. D., and O'Toole, A. J. (2020). Accuracy comparison across face recognition algorithms: Where are we on measuring race bias? *IEEE Transactions on Biometrics, Behavior, and Identity Science*.

Chen, I. Y., Johansson, F. D., and Sontag, D. (2018). Why is my classifier discriminatory? In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, pages 3543–3554, Red Hook, NY, USA. Curran Associates Inc.

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

Cousins, C. (2021). An axiomatic theory of provably-fair welfare-centric machine learning. In *Advances in Neural Information Processing Systems*.

Cousins, C. (2022). Uncertainty and the social planner's problem: Why sample complexity matters. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*.

Cousins, C. (2023a). Algorithms and analysis for optimizing robust objectives in fair machine learning. In *Columbia Workshop on Fairness in Operations and AI*. Columbia University.

Cousins, C. (2023b). Revisiting fair-PAC learning and the axioms of cardinal welfare. In *Artificial Intelligence and Statistics (AISTATS)*.

Cousins, C., Asadi, K., and Littman, M. L. (2022). Fair E^3: Efficient welfare-centric fair reinforcement learning. In *5th Multidisciplinary Conference on Reinforcement Learning and Decision Making (RLDM)*.

Cousins, C. and Riondato, M. (2020). Sharp uniform convergence bounds through empirical centralization. *Advances in Neural Information Processing Systems*, 33.

Cousins, C., Viswanathan, V., and Zick, Y. (2023a). Dividing good and better items among agents with submodular valuations. In *International Conference on Web and Internet Economics*. Springer.

Cousins, C., Viswanathan, V., and Zick, Y. (2023b). The good, the bad and the submodular: Fairly allocating mixed manna under order-neutral submodular preferences. In *International Conference on Web and Internet Economics*. Springer.

Dalton, H. (1920). The measurement of the inequality of incomes. *The Economic Journal*, 30(119):348–361.

Debreu, G. (1959). Topological methods in cardinal utility theory. *Cowles Foundation Discussion Papers*, 76.

Deschamps, R. and Gevers, L. (1978). Leximin and utilitarian rules: a joint characterization. *Journal of Economic Theory*, 17(2):143–163.

Diamond, S. and Boyd, S. (2016). CVXPY: A Python-embedded modeling language for convex optimization. *Journal of Machine Learning Research*, 17(83):1–5.

Diana, E., Gill, W., Kearns, M., Kenthapadi, K., and Roth, A. (2021). Minimax group fairness: Algorithms and experiments. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pages 66–76.

Domahidi, A., Chu, E., and Boyd, S. (2013). ECOS: An SOCP solver for embedded systems. In *European Control Conference (ECC)*, pages 3071–3076.

Dong, E. and Cousins, C. (2022). Decentering imputation: Fair learning at the margins of demographics. In *Queer in AI Workshop @ ICML*.

Dwork, C., Immorlica, N., Kalai, A. T., and Leiserson, M. (2018). Decoupled classifiers for group-fair and efficient machine learning. In *Conference on fairness, accountability and transparency*, pages 119–133. PMLR.

Gajdos, T. and Weymark, J. A. (2005). Multidimensional generalized Gini indices. *Economic Theory*, 26(3):471–496.

Gorman, W. M. (1968). The structure of utility functions. *The Review of Economic Studies*, 35(4):367–390.

Hoeffding, W. (1963). Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30.

Hu, W., Niu, G., Sato, I., and Sugiyama, M. (2018). Does distributionally robust supervised learning give robust classifiers? In *International Conference on Machine Learning*, pages 2029–2037. PMLR.

Lahoti, P., Beutel, A., Chen, J., Lee, K., Prost, F., Thain, N., Wang, X., and Chi, E. (2020). Fairness without demographics through adversarially reweighted learning. *Advances in neural information processing systems*, 33:728–740.

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

McDiarmid, C. (1989). On the method of bounded differences. *Surveys in combinatorics*, 141(1):148–188.

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., and Galstyan, A. (2021). A survey on bias and fairness in machine learning. *ACM computing surveys (CSUR)*, 54(6):1–35.

Mitzenmacher, M. and Upfal, E. (2017). *Probability and computing: Randomization and probabilistic techniques in algorithms and data analysis*. Cambridge University Press, second edition.

Obermeyer, Z., Powers, B., Vogeli, C., and Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464):447–453.

O'Donoghue, B., Chu, E., Parikh, N., and Boyd, S. (2016). Conic optimization via operator splitting and homogeneous self-dual embedding. *Journal of Optimization Theory and Applications*, 169(3):1042–1068.

Oren, Y., Sagawa, S., Hashimoto, T. B., and Liang, P. (2019). Distributionally robust language modeling. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4227–4237.

Pigou, A. C. (1912). *Wealth and welfare*. Macmillan and Company, limited.

Sagawa, S., Koh, P. W., Hashimoto, T. B., and Liang, P. (2019). Distributionally robust neural networks. In *International Conference on Learning Representations*.

Schneider, M. and Kim, B.-C. (2020). The utilitarian-maximin social welfare function and anomalies in social choice. *Southern Economic Journal*, 87(2):629–646.

Shalev-Shwartz, S. and Ben-David, S. (2014). *Understanding machine learning: From theory to algorithms*. Cambridge University Press.

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

Thomas, P. S., da Silva, B. C., Barto, A. G., Giguere, S., Brun, Y., and Brunskill, E. (2019). Preventing undesirable behavior of intelligent machines. *Science*, 366(6468):999–1004.

Viswanathan, V. and Zick, Y. (2023). A general framework for fair allocation under matroid rank valuations. In *Proceedings of the 24th ACM Conference on Economics and Computation*, pages 1129–1152.

Wang, H., Hsu, H., Diaz, M., and Calmon, F. P. (2021). To split or not to split: The impact of disparate treatment in classification. *IEEE Transactions on Information Theory*, 67(10):6733–6757.

Weymark, J. A. (1981). Generalized Gini inequality indices. *Mathematical Social Sciences*, 1(4):409–430.

Zhang, C., Tao, D., Hu, T., and Liu, B. (2020). Generalization bounds of multitask learning from perspective of vector-valued function learning. *IEEE Transactions on Neural Networks and Learning Systems*, 32(5):1906–1919.

Zhang, Y. and Yang, Q. (2018). An overview of multi-task learning. *National Science Review*, 5(1):30–43.

Zhang, Y. and Yang, Q. (2021). A survey on multi-task learning. *IEEE Transactions on Knowledge and Data Engineering*, 34(12):5586–5609.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Not Applicable]
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Yes]
2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes]
   (b) Complete proofs of all theoretical results. [Yes]
   (c) Clear explanations of any assumptions. [Yes]
3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes]
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes]
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Not Applicable]
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Yes]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [Not Applicable]
   (b) The license information of the assets, if applicable. [Not Applicable]
   (c) New assets either in the supplemental material or as a URL, if applicable. [Not Applicable]
   (d) Information about consent from data providers/curators. [Not Applicable]
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]
5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. [Not Applicable]
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

A Proofs

We now show theorem 3.

Theorem 3 (Theoretical Group-Regularized Malfare Bounds)**.** *Suppose a monotonic* malfare function $\mathrm{M}(\cdot) : \mathbb{R}^g \to \mathbb{R}$, hypothesis class $\mathcal{H} \subseteq \mathcal{X} \to \mathcal{Y}'$, loss function $\ell : \mathcal{Y}' \times \mathcal{Y} \to \mathbb{R}$, *per-group* distributions $\mathcal{D}_{1:g}$ *over* $\mathcal{X} \times \mathcal{Y}$, *and per-group* samples $\boldsymbol{z}_{1:g}$, *with* $\boldsymbol{z}_j \sim \mathcal{D}_j^{\boldsymbol{m}_j}$ *for each group* j*. Fix any group index* i*, and take* $\mathcal{H}_i^*$ *defined as in* (7)*. The following then hold.*

1. *With probability at least* $1 - 2\delta$ *over choice of* $\boldsymbol{z}_i$*, it holds that* $\hat{h} \in \mathcal{H}_i^*$*.*
2. *With probability at least* $1 - 4\delta$ *over choice of* $\boldsymbol{z}_i$*,*

$$\left| \mathrm{R}(\hat{h}, \mathcal{D}_i) - \hat{\mathrm{R}}(\hat{h}, \boldsymbol{z}_i) \right| \le 2\boldsymbol{\mathfrak{R}}_{\boldsymbol{m}_i}(\ell \circ \mathcal{H}_i^*, \mathcal{D}_i) + \boldsymbol{\varepsilon}_i \enspace .$$

Proof. We begin by proving part 1 and then we prove part 2 as a consequence.

We now show part 1. Recall that $\boldsymbol{\eta}_i \doteq 2\boldsymbol{\mathfrak{R}}_{\boldsymbol{m}_i}(\ell \circ \mathcal{H}, \mathcal{D}_i) + \boldsymbol{\varepsilon}_i$. With probability at least $1 - 2\delta$, for all $h \in \mathcal{H}$, it holds that

$$\left| \mathrm{R}(h, \mathcal{D}_i) - \hat{\mathrm{R}}(h, \boldsymbol{z}_i) \right| \le \boldsymbol{\eta}_i \enspace .$$

This is a textbook application of McDiarmid's bounded difference inequality, using twice the Rademacher average to bound the expected supremum deviation, i.e., the upper and lower tails of (3).

We could use this directly to show a weaker version of the result, however to show the stated form, we need only one tail of the above, which is used to bound generalization error of the (unknown) $\hat{h}$, and also one tail of the simple Hoeffding's inequality tail bound (2).

Now, suppose some arbitrary but fixed h' that realizes the infimum of (7), i.e.,

$$h' \in \operatorname*{argmin}_{h' \in \mathcal{H}} \mathrm{M}\left(j \mapsto \begin{cases} j \neq i & \hat{\mathrm{R}}(h', \boldsymbol{z}_j) \\ j = i & \mathrm{R}(h', \mathcal{D}_i) + \boldsymbol{\varepsilon}_i \end{cases} \right)$$

(technically, h' may be in $\mathcal{H}$ or a limit of a sequence of functions in $\mathcal{H}$). Recalling $\boldsymbol{\varepsilon}_i \doteq r\sqrt{\frac{\ln \frac{1}{\delta}}{2\boldsymbol{m}_i}}$, we obtain by Hoeffding's inequality that, with probability at least $1 - 2\delta$, it holds

$$\left| \mathrm{R}(h', \mathcal{D}_i) - \hat{\mathrm{R}}(h', \boldsymbol{z}_i) \right| \le \boldsymbol{\varepsilon}_i \enspace .$$

Therefore, when these bounds hold, we have

$$\begin{aligned}
\mathrm{M}\left(j \mapsto \begin{cases} j \neq i & \hat{\mathrm{R}}(\hat{h}, \boldsymbol{z}_j) \\ j = i & \mathrm{R}(\hat{h}, \mathcal{D}_i) - \boldsymbol{\eta}_i \end{cases} \right) &\le \mathrm{M}\Big(j \mapsto \hat{\mathrm{R}}(\hat{h}, \boldsymbol{z}_j) \Big) && \begin{array}{r} \text{W.h.p.: } \mathrm{R}(\hat{h}, \mathcal{D}_i) - \boldsymbol{\eta}_i \le \hat{\mathrm{R}}(\hat{h}, \boldsymbol{z}_i) \\ \text{Monotonicity of } \mathrm{M}(\cdot) \end{array} \\
&= \inf_{h' \in \mathcal{H}} \mathrm{M}\Big(j \mapsto \hat{\mathrm{R}}(h', \boldsymbol{z}_j) \Big) && \text{By Definition} \\
&\le \inf_{h' \in \mathcal{H}} \mathrm{M}\left(j \mapsto \begin{cases} j \neq i & \hat{\mathrm{R}}(h', \boldsymbol{z}_j) \\ j = i & \mathrm{R}(h', \mathcal{D}_i) + \boldsymbol{\varepsilon}_i \end{cases} \right) \enspace . && \begin{array}{r} \text{W.h.p.: } \hat{\mathrm{R}}(h', \boldsymbol{z}_i) \le \mathrm{R}(h', \mathcal{D}_i) + \boldsymbol{\varepsilon}_i \\ \text{Monotonicity of } \mathrm{M}(\cdot) \end{array}
\end{aligned}$$

We may thus conclude with probability at least $1 - 4\delta$ that $\hat{h} \in \mathcal{H}_i^*$ (by definition). However, observe that both the McDiarmid (Rademacher) and Hoeffding bounds required only one tail each, and thus a more careful analysis yields the guarantee with probability at least $1 - 2\delta$.

We now show part 2. By part 1, we have that $\hat{h} \in \mathcal{H}_i^*$ with probability at least $1 - 2\delta$. Then we apply the standard 2-tailed Rademacher bound with McDiarmid's inequality over the restricted class $\mathcal{H}_i^*$, i.e., we have

$$\mathop{\mathbb{P}}_{\boldsymbol{z}_i \sim \mathcal{D}_i^{\boldsymbol{m}_i}} \left(\sup_{h \in \mathcal{H}_i^*} \left| \mathrm{R}(h, \mathcal{D}_i) - \hat{\mathrm{R}}(h, \boldsymbol{z}_i) \right| \le \boldsymbol{\eta}_i \right) \le 1 - 2\delta \enspace .$$

The union bound then yields the desideratum. □

We now show theorem 4.

Theorem 4 (Empirical Group-Regularized Malfare Bounds)**.** *Suppose as in theorem 3. The following then hold for* $\hat{\mathcal{H}}_i$ *defined as in* (8)*.*

1. *With probability at least $1 - 4\delta$ over choice of $\boldsymbol{z}_i$, it holds that $\hat{h} \in \mathcal{H}_i^* \subseteq \hat{\mathcal{H}}_i$.*
2. *With probability at least $1 - 6\delta$, it holds that*

$$\left|\mathrm{R}(\hat{h}, \mathcal{D}_i) - \hat{\mathrm{R}}(\hat{h}, \boldsymbol{z}_i)\right| \leq 2\hat{\boldsymbol{\mathfrak{R}}}_{\boldsymbol{m}_i}(\ell \circ \hat{\mathcal{H}}_i, \boldsymbol{z}_i) + 2\boldsymbol{\varepsilon}_i \enspace .$$

Proof. We begin by proving part 1, and we then show part 2 as a consequence.

We now show part 1. First, we apply part 1 of theorem 3 (2 tails). We will then argue that

$$\mathop{\mathbb{P}}_{\boldsymbol{z}_i \sim \mathcal{D}_i^{\boldsymbol{m}_i}}\left(\mathcal{H}_i^* \subseteq \hat{\mathcal{H}}_i\right) \geq 1 - 2\delta \enspace ,$$

which holds for similar reasons (a 1-tail Rademacher bound for $\mathcal{H}$, and a 1-tail Hoeffding bound for $\hat{h}$, both the opposite tails bounded in part 1 of theorem 3). The result then follows via union bound.

In particular, recall (7)

$$\mathcal{H}_i^* \doteq \left\{ h \in \mathcal{H} \,\middle|\, \mathrm{M}\left(j \mapsto \begin{cases} j \neq i & \hat{\mathrm{R}}(h, \boldsymbol{z}_j) \\ j = i & \mathrm{R}(h, \mathcal{D}_i) - \boldsymbol{\eta}_i \end{cases}\right) \leq \inf_{h' \in \mathcal{H}} \mathrm{M}\left(j \mapsto \begin{cases} j \neq i & \hat{\mathrm{R}}(h', \boldsymbol{z}_j) \\ j = i & \mathrm{R}(h', \mathcal{D}_i) + \boldsymbol{\varepsilon}_i \end{cases}\right) \right\} \enspace ,$$

and also (8)

$$\hat{\mathcal{H}}_i \doteq \left\{ h \in \hat{\mathcal{H}} \,\middle|\, \mathrm{M}\left(j \mapsto \begin{cases} j \neq i & \hat{\mathrm{R}}(h, \boldsymbol{z}_j) \\ j = i & \hat{\mathrm{R}}(h, \boldsymbol{z}_i) - 2\hat{\boldsymbol{\eta}}_i \end{cases}\right) \leq \inf_{h' \in \mathcal{H}} \mathrm{M}\left(j \mapsto \begin{cases} j \neq i & \hat{\mathrm{R}}(h', \boldsymbol{z}_j) \\ j = i & \hat{\mathrm{R}}(h', \boldsymbol{z}_i) + 2\boldsymbol{\varepsilon}_i \end{cases}\right) \right\} \enspace .$$

Now, observe that by McDiarmid's inequality, by essentially the same argument as in (3), it holds that

$$\mathop{\mathbb{P}}_{\boldsymbol{z}_i \sim \mathcal{D}_i^{\boldsymbol{m}_i}}\left(\sup_{h \in \mathcal{H}} \hat{\mathrm{R}}(h, \boldsymbol{z}_i) - \mathrm{R}(h, \mathcal{D}_i) + \boldsymbol{\eta}_i > 2\hat{\boldsymbol{\eta}}_i\right) = \mathop{\mathbb{P}}_{\boldsymbol{z}_i \sim \mathcal{D}_i^{\boldsymbol{m}_i}}\left(\sup_{h \in \mathcal{H}} \hat{\mathrm{R}}(h, \boldsymbol{z}_i) - \mathrm{R}(h, \mathcal{D}_i) + 2\boldsymbol{\mathfrak{R}}_{\boldsymbol{m}_i}(\ell \circ \mathcal{H}, \mathcal{D}_i) > 4\hat{\boldsymbol{\mathfrak{R}}}_{\boldsymbol{m}_i}(\ell \circ \mathcal{H}, \boldsymbol{z}_i) + 3\boldsymbol{\varepsilon}_i\right) < \delta \enspace .$$

We thus have that, with probability at least $1 - \delta$, for all $h \in \mathcal{H}_i^*$,

$$\mathrm{M}\left(j \mapsto \begin{cases} j \neq i & \hat{\mathrm{R}}(h, \boldsymbol{z}_j) \\ j = i & \hat{\mathrm{R}}(h, \boldsymbol{z}_i) - 2\hat{\boldsymbol{\eta}}_i \end{cases}\right) \leq \mathrm{M}\left(j \mapsto \begin{cases} j \neq i & \hat{\mathrm{R}}(h, \boldsymbol{z}_j) \\ j = i & \mathrm{R}(h, \mathcal{D}_i) - \boldsymbol{\eta}_i \end{cases}\right) \enspace ,$$

and similarly, with probability at least $1 - \delta$ by the Hoeffding bound (2) on h', we have

$$\inf_{h' \in \mathcal{H}} \mathrm{M}\left(j \mapsto \begin{cases} j \neq i & \hat{\mathrm{R}}(h', \boldsymbol{z}_j) \\ j = i & \mathrm{R}(h', \mathcal{D}_i) + \boldsymbol{\varepsilon}_i \end{cases}\right) \leq \inf_{h' \in \mathcal{H}} \mathrm{M}\left(j \mapsto \begin{cases} j \neq i & \hat{\mathrm{R}}(h', \boldsymbol{z}_j) \\ j = i & \hat{\mathrm{R}}(h', \boldsymbol{z}_i) + 2\boldsymbol{\varepsilon}_i \end{cases}\right) \enspace ,$$

where both steps apply monotonicity of $\mathrm{M}(\cdot)$.

From this, we may conclude that, with probability at least $1 - 2\delta$, for each $h \in \mathcal{H}$, if the constraint in (7) is satisfied, then the constraint (8) is satisfied, thus $\mathcal{H}_i^* \subseteq \hat{\mathcal{H}}_i$. The union bound over all tail bounds above then yields part 1.

We now show part 2. This result essentially follows the structure of part 2 of theorem 3. However, we now start with part 1 above, which allows us to conclude that $\hat{h} \in \hat{\mathcal{H}}_i$ with probability at least $1 - 4\delta$, and then apply the standard empirical Rademacher bounds, i.e., 2 tails of (3) (we require only the upper and lower bounds to the supremum deviation, not the bound on the Rademacher average itself), to $\hat{\mathcal{H}}_i$ (rather than to $\mathcal{H}_i^*$). Taking the union bound over all events then yields the desideratum. □

We now show corollary 5.

Corollary 5 (Empirical Malfare Generalization Bounds)**.** *Suppose as in theorem 4. Suppose also that there exists some $\lambda > 0$ and norm $\|\cdot\|_{\mathrm{M}}$ such that $\mathrm{M}(\cdot)$ is λ-$\|\cdot\|_{\mathrm{M}}$ Lipschitz continuous, i.e., $\forall \mathcal{S}, \mathcal{S}'$: $\mathrm{M}(\mathcal{S} + \mathcal{S}') \leq \mathrm{M}(\mathcal{S}) + \lambda\|\mathcal{S}'\|_{\mathrm{M}}$. We then have:*

1. *With probability at least $1 - 5g\delta$, the true malfare of $\hat{h}$ is bounded by*

$$\begin{aligned} &\mathrm{M}(j \mapsto \mathrm{R}(\hat{h}, \mathcal{D}_j)) \\ &\quad \leq \mathrm{M}\left(j \mapsto \hat{\mathrm{R}}(\hat{h}, \boldsymbol{z}_j) + 2\hat{\boldsymbol{\mathfrak{R}}}_{\boldsymbol{m}_j}(\hat{\mathcal{H}}_j, \boldsymbol{z}_j) + 2\boldsymbol{\varepsilon}_j\right) \\ &\quad \leq \mathrm{M}\left(j \mapsto \hat{\mathrm{R}}(\hat{h}, \boldsymbol{z}_j)\right) + \lambda\left\|j \mapsto 2\hat{\boldsymbol{\mathfrak{R}}}_{\boldsymbol{m}_j}(\hat{\mathcal{H}}_j, \boldsymbol{z}_j) + 2\boldsymbol{\varepsilon}_j\right\|_{\mathrm{M}} \enspace . \end{aligned}$$

2. With probability at least $1 - 6g\delta$, we bound the suboptimality of $\hat{h}$ as

$$\begin{aligned}
&\mathrm{M}\left(j \mapsto \mathrm{R}(\hat{h}, \mathcal{D}_j)\right) \\
&\qquad\qquad \leq \mathrm{M}\left(j \mapsto \mathrm{R}(h^*, \mathcal{D}_j) + 2\hat{\mathfrak{R}}_{\boldsymbol{m}_j}(\hat{\mathcal{H}}_j, \boldsymbol{z}_j) + 3\boldsymbol{\varepsilon}_j\right) \\
&\quad \implies \quad \left|\mathrm{M}\left(j \mapsto \mathrm{R}(h^*, \mathcal{D}_j)\right) - \mathrm{M}\left(j \mapsto \mathrm{R}(\hat{h}, \mathcal{D}_j)\right)\right| \\
&\qquad\qquad \leq \lambda \left\| j \mapsto 2\hat{\mathfrak{R}}_{\boldsymbol{m}_j}(\hat{\mathcal{H}}_j, \boldsymbol{z}_j) + 3\boldsymbol{\varepsilon}_j \right\|_{\mathrm{M}} \quad .
\end{aligned}$$

Proof. For both results, we apply part 2 of theorem 4 to each group i, which by union bound gives a result with probability at least $1 - 6g\delta$. However, careful accounting reveals that we only require one tail of the final Rademacher bound of theorem 4 part 2, i.e., we require $\mathrm{R}(\hat{h}, \mathcal{D}_j) \leq \hat{\mathrm{R}}(\hat{h}, \boldsymbol{z}_j) + 2\hat{\mathfrak{R}}_{\boldsymbol{m}_j}(\hat{\mathcal{H}}_j, \boldsymbol{z}_j) + 2\boldsymbol{\varepsilon}_j$, *but not* $\hat{\mathrm{R}}(\hat{h}, \boldsymbol{z}_j) \leq \mathrm{R}(\hat{h}, \mathcal{D}_j) + 2\hat{\mathfrak{R}}_{\boldsymbol{m}_j}(\hat{\mathcal{H}}_j, \boldsymbol{z}_j) + 2\boldsymbol{\varepsilon}_j$, thus we begin with tail bounds that hold with probability at least $1 - 5g\delta$.

We now show part 1. Subject to all tail bounds holding, we have for all $j \in 1, \dots, g$ that $\mathrm{R}(\hat{h}, \mathcal{D}_j) \leq \hat{\mathrm{R}}(\hat{h}, \boldsymbol{z}_j) + 2\hat{\mathfrak{R}}_{\boldsymbol{m}_j}(\hat{\mathcal{H}}_j, \boldsymbol{z}_j) + 2\boldsymbol{\varepsilon}_j$, thus by monotonicity of $\mathrm{M}(\cdot)$, we have

$$\mathrm{M}\left(j \mapsto \mathrm{R}(\hat{h}, \mathcal{D}_j)\right) \leq \mathrm{M}\left(j \mapsto \hat{\mathrm{R}}(\hat{h}, \boldsymbol{z}_j) + 2\hat{\mathfrak{R}}_{\boldsymbol{m}_j}(\hat{\mathcal{H}}_j, \boldsymbol{z}_j) + 2\boldsymbol{\varepsilon}_j\right) \quad .$$

Applying the Lipschitz property then yields the final portion of part 1.

We now show part 2. First, observe that $\mathrm{M}\left(j \mapsto \mathrm{R}(h^*, \mathcal{D}_j)\right) \leq \mathrm{M}\left(j \mapsto \mathrm{R}(\hat{h}, \mathcal{D}_j)\right)$ by definition. For the remaining inequality, we introduce one new tail bound for each group j, in particular, a 1-tail Hoeffding bound of

$$\mathop{\mathbb{P}}_{\boldsymbol{z}_j \sim \mathcal{D}_j^{\boldsymbol{m}_j}}\left(\mathrm{R}(h^*, \mathcal{D}_j) \leq \hat{\mathrm{R}}(h^*, \boldsymbol{z}_j) + \boldsymbol{\varepsilon}_j\right) \geq 1 - \delta \quad .$$

This seems familiar, but it is not quite the same as the 2-tail Hoeffding bound on each $\mathrm{R}(h', \mathcal{D}_j)$ used by theorems 3 and 4, thus this tail bound must be counted separately. Now, we substitute into the result of part 1, again applying monotonicity, to get

$$\mathrm{M}\left(j \mapsto \mathrm{R}(\hat{h}, \mathcal{D}_j)\right) \leq \mathrm{M}\left(j \mapsto \mathrm{R}(h^*, \mathcal{D}_j) + 2\hat{\mathfrak{R}}_{\boldsymbol{m}_j}(\hat{\mathcal{H}}_j, \boldsymbol{z}_j) + 3\boldsymbol{\varepsilon}_j\right) \quad .$$

Applying the Lipschitz property then yields the final portion of part 2. By union bound, we may conclude the result with probability at least $1 - 6g\delta$. □

We now show lemma 6.

Lemma 6 (Convex Optimization for Monte-Carlo Rademacher Averages)**.** *Suppose the* parameter space $\boldsymbol{B}$ *of* $\mathcal{H}$ *is a convex set,* loss $\ell(h_{\boldsymbol{\beta}}(\boldsymbol{x}), y)$ *is convex in* $\boldsymbol{\beta} \in \boldsymbol{B}$ *for all* $\boldsymbol{x} \in \mathcal{X}$*,* $y \in \mathcal{Y}$*, and* malfare $\mathrm{M}(\cdot) : \mathbb{R}^g \to \mathbb{R}$ *is quasiconvex and monotonically increasing in each argument. Then the parameter spaces of* $\hat{\mathcal{H}}_i$ *and* $\mathcal{H}_i^*$ *are convex sets.*

Moreover, if $g \circ \mathcal{H}$ *is an affine function family, then* $\hat{\mathfrak{R}}_m^n(g \circ \hat{\mathcal{H}}_i, \boldsymbol{z}_i; \boldsymbol{\sigma})$ *reduces to maximizing a linear function over a convex set. Similarly, if we strengthen the quasiconvexity assumption on* $\mathrm{M}(\cdot)$ *to* convexity*, then EMM reduces to minimizing a convex objective over the convex set* $\boldsymbol{B}$*.*

Proof. We first show that the restricted parameter spaces of $\mathcal{H}_i^*$ and $\hat{\mathcal{H}}_i$ are convex sets.

The crux of this result is to show that $\mathrm{M}\big(j \mapsto f_j(\boldsymbol{\beta})\big)$ is quasiconvex, where $f_j(x)$ represents $\hat{\mathrm{R}}(h_{\boldsymbol{\beta}}, \boldsymbol{z}_j) - \boldsymbol{c}_j$ or $\mathrm{R}(h_{\boldsymbol{\beta}}, \mathcal{D}_j) - \boldsymbol{c}_j$ for some constant $\boldsymbol{c} \in \mathbb{R}^g$. This indeed holds, so long as $f_j(\boldsymbol{\beta})$ is quasiconvex. First note that convexity of *loss* immediately implies convexity of (empirical) *risk*. Now, by standard compositional rules, since we assume $\mathrm{M}(\cdot)$ to be quasiconvex and monotonic, we conclude that $\mathrm{M}(j \mapsto f_j(\boldsymbol{\beta}))$ is quasiconvex in $\boldsymbol{\beta} \in \boldsymbol{B}$.

Now, converting $\mathcal{H}$ to $\boldsymbol{B}$, observe that the parameter spaces associated with both $\mathcal{H}_i^*$ in (7)

$$\left\{\boldsymbol{\beta} \in \boldsymbol{B} \,\middle|\, \mathrm{M}\left(j \mapsto \begin{cases} j \neq i & \hat{\mathrm{R}}(h_{\boldsymbol{\beta}}, \boldsymbol{z}_j) \\ j = i & \mathrm{R}(h_{\boldsymbol{\beta}}, \mathcal{D}_i) - \boldsymbol{\eta}_i \end{cases}\right) \leq \inf_{h' \in \mathcal{H}} \mathrm{M}\left(j \mapsto \begin{cases} j \neq i & \hat{\mathrm{R}}(h', \boldsymbol{z}_j) \\ j = i & \mathrm{R}(h', \mathcal{D}_i) + \boldsymbol{\varepsilon}_i \end{cases}\right)\right\} \enspace ,$$

and also $\hat{\mathcal{H}}_i$ in (8)

$$\left\{ \boldsymbol{\beta} \in \boldsymbol{B} \,\middle|\, \mathrm{M}\left(j \mapsto \begin{cases} j \neq i & \hat{\mathrm{R}}(h_{\boldsymbol{\beta}}, \boldsymbol{z}_j) \\ j = i & \hat{\mathrm{R}}(h_{\boldsymbol{\beta}}, \boldsymbol{z}_i) - 2\hat{\boldsymbol{\eta}}_i \end{cases} \right) \leq \inf_{h' \in \mathcal{H}} \mathrm{M}\left(j \mapsto \begin{cases} j \neq i & \hat{\mathrm{R}}(h', \boldsymbol{z}_j) \\ j = i & \hat{\mathrm{R}}(h', \boldsymbol{z}_i) + 2\boldsymbol{\varepsilon}_i \end{cases} \right) \right\} \ ,$$

are subsets of the convex set $\boldsymbol{B}$. In particular, the RHS of the condition is *constant* in $\boldsymbol{\beta}$, and as above, the LHS is quasiconvex in $\boldsymbol{\beta}$, thus both restricted parameter spaces are convex sets.

Now, note that once we determine the parameter space to be convex, Monte-Carlo Rademacher averages can be efficiently computed via standard convex optimization techniques, e.g., first-order methods to maximize a linear objective on a convex set. Key to this observation is that we assumed $g \circ \mathcal{H}$ to be an affine function family, thus even after multiplying terms by ±1 in the Monte-Carlo Rademacher average (11), the objective of the supremum remains convex.

Finally, observe that if $\mathrm{M}(\cdot)$ is convex and monotonically increasing, then the EMM objective is also convex. This follows from standard compositional rules, see discussion following Boyd and Vandenberghe (2004) equation (3.15). EMM then reduces to minimizing a convex function on a convex set. □

B Implementation details

All code used to generate the results in this paper are available upon request. The computation of each supremum in (11), i.e., $\hat{\boldsymbol{\mathfrak{R}}}^n_m(g \circ \mathcal{H}, \boldsymbol{z}_i; \boldsymbol{\sigma})$ and $\hat{\boldsymbol{\mathfrak{R}}}^n_m(g \circ \hat{\mathcal{H}}_i, \boldsymbol{z}_i; \boldsymbol{\sigma})$ optimize linear functions of β. However, since the restricted hypothesis constraints are convex functions (see lemma 6) over the parameter space, we need to use solvers that can handle nonlinear convex constraints. For this reason, we use either the ECOS (Domahidi et al., 2013) or SCS (O'Donoghue et al., 2016) algorithms available in CVXPY (Diamond and Boyd, 2016; Agrawal et al., 2018). These algorithms are also able to compute the upper bound of the restricted hypothesis class constraint itself, which minimizes an objective which is dependent on the loss ℓ.