Multicalibrated Regression for Downstream Fairness

Ira Globus-Harris[1], Varun Gupta[1], Christopher Jung[2], Michael Kearns[1], Jamie Morgenstern[3], and Aaron Roth[1]

[1]University of Pennsylvania
[2]Stanford University
[3]University of Washington

September 16, 2022

Abstract

We show how to take a regression function $\hat{f}$ that is appropriately "multicalibrated" and efficiently post-process it into an approximately error minimizing classifier satisfying a large variety of fairness constraints. The post-processing requires no labeled data, and only a modest amount of unlabeled data and computation. The computational and sample complexity requirements of computing $\hat{f}$ are comparable to the requirements for solving a single fair learning task optimally, but it can in fact be used to solve *many* different downstream fairness-constrained learning problems efficiently. Our post-processing method easily handles intersecting groups, generalizing prior work on post-processing regression functions to satisfy fairness constraints that only applied to disjoint groups. Our work extends recent work showing that multicalibrated regression functions are "omnipredictors" (i.e. can be post-processed to optimally solve unconstrained ERM problems) to constrained optimization.

1 Introduction

The most common technical framing for fair machine learning is as constrained optimization. The goal is to solve an empirical risk minimization problem over some class of models $\mathcal{H}$, subject to *fairness constraints*. For example, we might ask to find the best performing model $h \in \mathcal{H}$ that equalizes false positive rates, false negative rates, raw error rates, or positive classification rates across some collection of groups $\mathcal{G}$ [Hardt et al., 2016, Dwork et al., 2012]). For each of these notions of fairness, there is a continuum of relaxations to consider: rather than asking that (e.g.) false positive rates be exactly equalized across groups, we could ask that they deviate by not more than 5%, or 10%, or 15%, etc. Because these relaxations trade off with model accuracy (tracing out Pareto frontiers), it is common to explore the entire range of tradeoffs for a particular family of fairness constraints (see e.g. [Agarwal et al., 2018, Kearns et al., 2018]).

Each of these are distinct problems that seemingly require training fresh models on the data. And each of these problems can be computationally expensive to solve: for example, the "reductions" approach of Agarwal et al. [2018] requires solving roughly $\log|\mathcal{G}|/\epsilon^2$ empirical risk minimization problems over $\mathcal{H}$ to produce an ϵ-approximately optimal solution to any one of them, and the computations cannot be reused. Our goal is to understand when we can pre-compute a single regression model $\hat{f}$ which is sufficient to solve *all* of the fair machine learning problems described above, each as only a computationally easy *post-processing* of $\hat{f}$, without sacrificing accuracy.

1.1 Our Results in Context

The idea of *post-processing* a trained model $\hat{f}$ in order to satisfy fairness constraints is not new. For example, Hardt et al. [2016] propose a simple post-processing of a regression function $\hat{f}$ to derive a classifier subject

to false positive or negative rate constraints. However, the conditions under which such post-processing approaches work are not yet well understood. In particular, two important questions about post-proceessing remain: First, *how* should one algorithmically post-process a regression function $\hat{f}$ to obtain a good (and fair) downstream classifier, and *what properties* must $\hat{f}$ satisfy? Prior work [Hardt et al., 2016] handles the case in which the groups $\mathcal{G}$ are disjoint, by finding a different *thresholding* of $\hat{f}$ for each group $g \in \mathcal{G}$. Is there a simple, efficient post-processing that applies in the common case that groups intersect — as is the case e.g. with groups defined by race and gender? Second and similarly, Hardt et al. [2016] and Corbett-Davies et al. [2017] show that this post-processing yields the Bayes Optimal fair classifier if $\hat{f}$ is the true conditional label distribution. Are there weaker conditions on $\hat{f}$ (that can be efficiently satisfied from only a polynomial number of samples) that also lead to guarantees? We answer both of these questions in the affirmative.

Post Processing for Intersecting Groups Suppose we have $k = |\mathcal{G}|$ groups that are intersecting (e.g. divisions of a population by race, gender, income, nationality, etc.) A naive reduction to the post-processing approach of Hardt et al. [2016] would consider all 2^k (now disjoint) intersections of groups, and find a separate thresholding of $\hat{f}(x)$ for each one. We show that even when groups intersect, for a variety of fairness constraints, the optimal post-processing $\hat{h}$ remains a thresholding that depends on only k parameters λ_g, one for each group g. The value at which to threshold $\hat{f}(x)$ now depends only on these k parameters *and the subset of groups that x is contained in.* We give a simple, efficient algorithm to compute these optimal post-processings. The algorithm is efficient in the worst case — i.e. it does not have to call any heuristic "learning oracle" as direct learning approaches do [Agarwal et al., 2018, Kearns et al., 2018], and requires access only to a modest amount of *unlabeled* data from the underlying distribution.

Accuracy Guarantees from Multicalibration As in Hardt et al. [2016] when $\hat{f}$ is the Bayes optimal regression function, for a variety of fairness constraints, our post-processing $\hat{h}$ is the Bayes optimal fair classifier. But in general we cannot hope to learn the Bayes optimal regression function $\hat{f}$ given only a polynomial amount of data and computation. We show that substantially weaker conditions suffice: If $\hat{f}$ is *multicalibrated* with respect to a class of models $\mathcal{H}$, a class of groups $\mathcal{G}$, and a simple class of functions derived from $\mathcal{H}$ and $\mathcal{G}$, then the post-processing $\hat{h}$ of $\hat{f}$ will be as accurate as the best fair model in $\mathcal{H}$ while satisfying all of the fairness constraints defined over $\mathcal{G}$. Learning a multicalibrated predictor with respect to these classes can be done with polynomial sample complexity in an oracle-efficient manner whenever $\mathcal{H}$ and $\mathcal{G}$ have polynomial VC dimension — and so both the sample and computational complexity of computing $\hat{f}$ are comparable to what would be required to directly solve a single instance of a fairness constrained optimization problem over $\mathcal{H}$.

Experimental Evaluation We provide preliminary experimental evaluation of our method on a dataset derived from Pennsylvania Census data provided by the Folktables package [Ding et al., 2021]. The experimental results support the theoretical findings that our method is able to quickly converge to a solution that approximately satisfies the given target constraints.

Taken together, our results contribute to the following conclusion: even when the notion of fairness that is eventually desired in downstream tasks is one that approximately equalizes some notion of statistical error across groups, this is *not* necessarily what should be trained. Aiming instead for group-wise fidelity in the form of *multicalibration* provides the flexibility to deploy an optimal downstream model subject to a variety of fairness constraints without destroying information that would be needed to later relax or tighten those constraints, to remove them or to add more, or to change their type.

1.2 Additional Related Work

There are a number of other papers that study the problem of converting a regression (or "score") function into a classification rule in the context of fair machine learning. For example, Woodworth et al. [2017] shows

that post-proceessing a learned binary classification model to satisfy fairness constraints can be substantially suboptimal even when the hypothesis class under consideration contains the Bayes optimal predictor, which motivates a focus on post-processing regression functions instead. Yang et al. [2020] study the structure of the Bayes optimal fair classifier for several notions of fairness when groups are intersecting, under a continuity assumption on the underlying distribution; they do not consider utility guarantees for post-processing a regression function that does not completely represent the underlying probability distribution. Wei et al. [2021] and Alabdulmohsin and Lucic [2021] give post-processing algorithms that transforms a score function into a regression function that optimizes different measures of accuracy subject to a variety of fairness constraints using a similar primal/dual perspective that we use in this paper. But these papers do not address the two main questions we raise in our work: intersecting groups, and efficiently learnable conditions on the score function that lead to utility guarantees (they assume that in the limit the true conditional label distribution is learnable and given as input to their algorithm)

In proving our accuracy bounds, we draw on a recent line of work on *multicalibration* [Hébert-Johnson et al., 2018, Kim et al., 2019, Jung et al., 2021a, Dwork et al., 2021, Gupta et al., 2022]. In particular, Gopalan et al. [2022] showed that regression functions that are multicalibrated with respect to a class of models $\mathcal{H}$ are *omni-predictors* with respect to $\mathcal{H}$, which means that they can be post-processed to perform as well as the best model in $\mathcal{H}$ with respect to any convex loss function satisfying mild technical conditions. The results in our paper can be viewed as being a *constrained optimization* parallel to Gopalan et al. [2022], which studies *unconstrained* optimization.

Several other papers also use multicalibration of intermediate statistical products to argue for the utility of downstream models. Zhao et al. [2021] consider the problem of calibrating a model to the utility function of a downstream utility maximizing decision maker to preserve the usefulness of the model for the decision-maker. Diana et al. [2021] show that a *proxy-model* for a protected attribute can be useful in enforcing fairness constraints on a downstream model when the real protected attribute is not available if the proxy is appropriately multicalibrated. Burhanpurkar et al. [2021] propose training a multicalibrated predictor on the level sets of a low dimensional learned representation as a means of obtaining Bayes optimality. Kim et al. [2022] show that a predictor that is multicalibrated with respect to a function class can adapt to new domains with covariate shift as well as a model trained using propensity-score reweighting via any propensity score function in the class.

Hu et al. [2022] independently study a similar problem. Our two papers derive a closely related but incomparable set of results. Hu et al. [2022] tackles a more general problem, and studies a richer set of objective functions and constraints (whereas we restrict attention to the classification error objective and fairness motivated constraints). In contrast, in our paper, we are able to take advantage of the additional structure of our problem to derive improved bounds. In particular, we can handle intersecting groups (with running time and sample complexity depending polynomially on the number of groups), whereas Hu et al. [2022] requires taking all of the exponentially many group intersections to recover disjoint groups—which leads to an exponential (in the number of groups) loss in the running time and sample complexity. Similarly, they require more precise multicalibration as more groups are added, whereas we derive results from a multicalibrated predictor with parameter that is independent of the number of groups.

1.3 Limitations

This work explores approaches to fairness that make a jump between complex and ambiguous social ideas of fairness and mathematical guarantees such as equality of false positive rates between groups of individuals. Our work can be applied only when evaluating the membership of an individual to a group is well-defined, and when consideration of group membership is legal[1], and when the training data is representative of the underlying population. There will be contexts in which these assumptions are either false, overly simplistic, or bypass larger questions: e.g. an application might be fair in its performance but still entirely unethical, or

[1]Note that in some contexts such as consumer lending in the United States, direct consideration of membership in protected groups such as race is illegal. However, demographic information can be used when designing and auditing a decision-making process, so long as those characteristics are not part of the real-time lending decisions.

groups may be systematically underrepresented in datasets. In the latter case, the guarantees of our work cannot be interpreted as guarantees relative to the optimal predictor for the true distribution over groups.

It is worth noting that while the assumption that we can define group membership of individuals simplifies the complexities of personal identity, this work does improve on the existing literature on post-processing approaches to fairness in that it allows for *non-disjoint*, or intersectional, group membership. In general, this work (and all work in algorithmic fairness) should not be assumed to "solve" fairness. Instead it should be taken as a tool in a larger system to evaluate and remediate issues of fairness and ethics in machine learning.

2 Preliminaries

We study binary classification problems. Let $\mathcal{X}$ be an arbitrary feature space and $\mathcal{Y} = \{0, 1\}$ be a binary label space. A classification problem is defined by an underlying data distribution $\mathcal{D} \in \Delta(\mathcal{X} \times \mathcal{Y})$. In general we will not have direct access to the data distribution, but rather only to samples drawn i.i.d. from $\mathcal{D}$. We let D denote a dataset of size n, drawn i.i.d. from $\mathcal{D}$: $D \sim \mathcal{D}^n$.

We will study both regression functions $f : \mathcal{X} \to \mathbb{R}$ and classification functions (classifiers) $h : \mathcal{X} \to \{0, 1\}$. In general we will use f and variants ($f^*, \hat{f}$, etc.) when speaking of regression functions and h and variants ($h^*, \hat{h}$, etc.) when speaking of classification functions. Our interest will be in regression functions used to estimate conditoinal label expectations in binary prediction problems, and so the natural range of our regression functions will be (discrete subsets of) $[0, 1]$.

Definition 1 (Bayes Optimal Regression Function)**.** *We let f^* denote the* Bayes Optimal Regression Function *$f^* = \arg\min_f \mathbb{E}_{(x,y)\sim\mathcal{D}}(f(x) - y)^2$ which takes value:*

$$f^*(x) = \mathop{\mathbb{E}}_{(x',y')\sim\mathcal{D}}[y'|x' = x]$$

Remark. *The property of f^* that we are interested in is that it encodes the true conditional label expectations. The fact that it minimizes squared error is not important — f^* would also minimize any other proper loss function.*

Let $\mathcal{D}_\mathcal{X}$ denote the marginal distribution on features induced by projecting $\mathcal{D}$ onto $\mathcal{X}$. Note that we can equivalently sample a pair $(x, y) \sim \mathcal{D}$ by first sampling $x \sim \mathcal{D}_\mathcal{X}$ and then sampling $y = 1$ with probability $f^*(x)$ and $y = 0$ otherwise.

Given a classifier $h : \mathcal{X} \to \mathcal{Y}$, and a data distribution $\mathcal{D}$, we can refer to various notions of error. We will be interested in error rates not just overall, but on subsets of the data that we call *groups* (which we might think of as e.g. demographic groups when the data represents people). We will represent groups by group indicator functions:

Definition 2. *Let $\mathcal{G}$ denote a collection of* groups*, each represented by a group indicator function $g : \mathcal{X} \to \{0, 1\}$. If $g(x) = 1$ we say that x is a* member *of group g. Let I denote the group containing all elements ($I(x) = 1$ for all x). We will always assume that $I \in \mathcal{G}$.*

We allow $\mathcal{G}$ to contain arbitrarily intersecting groups. We can now define error rates over these groups, and a notion of fairness.

Definition 3. *The error of a classifier $h : \mathcal{X} \to \mathcal{Y}$ on a group g as measured over distribution $\mathcal{D}$ is:*

$$err(h, g, \mathcal{D}) = \Pr_{(x,y)\sim\mathcal{D}}[h(x) \neq y|g(x) = 1] = \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[\ell(h(x), y)|g(x) = 1]$$

The false positive rate of a classifier $h : \mathcal{X} \to \mathcal{Y}$ on a group g is:

$$\rho(h, g, \mathcal{D}) = \Pr_{(x,y)\sim\mathcal{D}}[h(x) \neq y|y = 0, g(x) = 1]$$

When h is a randomized classifier, the probabilities are computed over the randomness of h as well. For convenience, we write $err(h) = err(h, I, \mathcal{D})$, $\rho_g(h) \equiv \rho(h, g, \mathcal{D})$, and $\rho(h) \equiv \rho(h, I, \mathcal{D})$.

Definition 4. *We say that classifier $h : \mathcal{X} \to \mathcal{Y}$ satisfies γ-False Positive (FP) Fairness with respect to $\mathcal{D}$ and $\mathcal{G}$ if for all $g \in \mathcal{G}$,*

$$w_g \left|\rho_g(h) - \rho(h)\right| \leq \gamma.$$

where $w_g = \Pr_{(x,y)\sim\mathcal{D}}[g(x) = 1, y = 0]$.

Remark. *In the above definition, we include a multiplicative factor that provides slack in the fairness guarantee for groups with small weight over the distribution. This approximation parameter is necessary, as statistical estimation over small groups is inherently more difficult. By including this factor directly into our fairness constraint rather than incorporating it indirectly into sample complexity guarantees, we are able to elide exposition in our later proofs. An equivalent (up to reparameterization) alternative would be to remove the w_g term in our constraints, but to provide guarantees only for groups for whom w_g is sufficiently large.*

For the sake of brevity and clarity, in the main body of this paper we prove all results in the context of γ-False Positive Fairness. We discuss the modifications necessary to extend the results to other fairness notions in Appendix A.

We will study how to derive classifiers with optimal error properties, subject to fairness-motivated constraints on group-wise error rates from regression functions satisfying *multicalibration* constraints [Hébert-Johnson et al., 2018]. Informally, if $\hat{f}$ is multicalibrated with respect to a class of functions C, then $\hat{f}(x)$ takes values equal to $f^*(x)$ in expectation, even conditional on both the value of $\hat{f}(x)$ and on the value of $c(x)$ for each $c \in C$. We use two variants. The first (multicalibration in expectation) was defined and studied in [Gopalan et al., 2022]:

Definition 5 (Multicalibration in Expectation [Hébert-Johnson et al., 2018, Gopalan et al., 2022])**.** *Fix a distribution $\mathcal{D}$ and let C be a collection of functions $c : \mathcal{X} \to \{0, 1\}$. We say that a predictor $\hat{f} : \mathcal{X} \to R$ where R is some discrete domain $R \subseteq [0, 1]$ is α-approximately multicalibrated with respect to C if for every $c \in C$:*

$$\begin{aligned}
&\sum_{v\in R} \Pr_{x\sim\mathcal{D}_\mathcal{X}}[\hat{f}(x) = v] \cdot \left| \mathop{\mathbb{E}}_{(x,y)} \left[(\hat{f}(x) - f^*(x)) \cdot c(x) \middle| \hat{f}(x) = v \right] \right| \\
&= \sum_{v\in R} \Pr_{x\sim\mathcal{D}_\mathcal{X}}[\hat{f}(x) = v] \cdot \frac{\Pr_{x\sim\mathcal{D}_\mathcal{X}}[c(x) = 1|\hat{f}(x) = v]}{\Pr_{x\sim\mathcal{D}_\mathcal{X}}[c(x) = 1|\hat{f}(x) = v]} \cdot \left| \mathop{\mathbb{E}}_{(x,y)} \left[(\hat{f}(x) - f^*(x)) \cdot c(x) \middle| \hat{f}(x) = v \right] \right| \\
&= \sum_{v\in R} \Pr_{x\sim\mathcal{D}_\mathcal{X}}[c(x) = 1, \hat{f}(x) = v] \cdot \left| \mathop{\mathbb{E}}_{(x,y)} \left[\hat{f}(x) - f^*(x) \middle| \hat{f}(x) = v, c(x) = 1 \right] \right| \\
&= \sum_{v\in R} \Pr_{x\sim\mathcal{D}_\mathcal{X}}[c(x) = 1, \hat{f}(x) = v] \cdot \left| v - \mathop{\mathbb{E}}_{(x,y)} \left[f^*(x) \middle| \hat{f}(x) = v, c(x) = 1 \right] \right| \\
&\leq \alpha
\end{aligned}$$

We will require this notion of multicalibration with respect to the set of groups $\mathcal{G}$ with which we define our fairness constraints, for the classifiers $h \in \mathcal{H}$, and for the intersection of these classes $\mathcal{G} \times \mathcal{H} = \{g(x) \cdot h(x) | g \in \mathcal{G}, h \in \mathcal{H}\}$. We will also need a variant of multicalibration that is tailored to two-argument functions $c : \mathcal{X} \times R \to \{0, 1\}$ in order to argue about the properties of thresholding functions, which take both a value $x \in \mathcal{X}$ and a threshold in a discrete domain $R \subseteq [0, 1]$, and which threshold predictions to $\{0, 1\}$.

In this definition, when we condition on $\hat{f}(x) = v$, we also condition on the second argument of c taking the same value v. We call this *joint*-multicalibration. It is only a modest generalization of multicalibration: we verify in Appendix C that existing algorithms for obtaining multicalibrated predictors easily extend to our definition of joint multicalibration.

Definition 6 (Joint Multicalibration in Expectation)**.** *We say that a predictor $\hat{f} : \mathcal{X} \to R$ where R is some discrete domain $R \subseteq [0, 1]$ is α-approximately jointly multicalibrated with respect to a class of functions $c : \mathcal{X} \times R \to \{0, 1\}$ if for every $c \in C$:*

$$\sum_{v\in R} \Pr[\hat{f}(x) = v, c(x, v) = 1] \cdot \left| \mathop{\mathbb{E}}_{(x,y)} \left[(\hat{f}(x) - f^*(x)) \middle| \hat{f}(x) = v, c(x, v) = 1 \right] \right| \leq \alpha,$$

which is equivalent to

$$\sum_{v \in R} \Pr[\hat{f}(x) = v] \cdot \left| \mathop{\mathbb{E}}_{(x,y)} \left[c(x,v) \cdot (\hat{f}(x) - f^*(x)) \middle| \hat{f}(x) = v \right] \right| \leq \alpha.$$

3 The Structure of an Optimal Post-Processing

In this section, we consider a fairness-constrained optimization problem that seeks to find the (distribution over) model(s) in $\mathcal{H}$ that minimize error subject to a constraint on group-wise false positive rates:

$$\begin{aligned} &\min_{h \in \Delta\mathcal{H}} && \mathrm{err}(h) \\ &\text{s.t. for each } g \in \mathcal{G}: && w_g |\rho_g(h) - \rho(h)| \leq \gamma, \end{aligned} \tag{1}$$

where $w_g, \rho_g(h)$, and $\rho(h)$ are defined as in Definition 3.

It will be useful for us to re-write this optimization problem in terms of the conditional label expectation $f^*(x)$. Since later in the paper we will want to replace $f^*(x)$ with a different regression function f that is easier to learn, we define the linear program generically in terms of an arbitrary regression function f:

Definition 7. *Let $f : \mathcal{X} \to R \subseteq [0,1]$ be some regression function and let $\gamma \in \mathbb{R}_+$. Define $\psi(f, \gamma, \mathcal{H})$ to be the following optimization problem:*

$$\begin{aligned} &\min_{h \in \Delta\mathcal{H}} && \mathop{\mathbb{E}}_{x \sim \mathcal{D}_{\mathcal{X}}} [f(x)\ell(h(x),1) + (1 - f(x))\ell(h(x),0)] \\ &\textit{s.t. for each } g \in \mathcal{G}: && |\mathbb{E}[\ell(h(x),0)g(x)(1-f(x))] - \beta_g \mathbb{E}[\ell(h(x),0)(1-f(x))]| \leq \gamma, \end{aligned}$$

where $\beta_g = \Pr[g(x) = 1 | y = 0]$.

Lemma 1. *Let f^* be the Bayes optimal regression function over $\mathcal{D}$. Then optimization problem $\psi(f^*, \gamma, \mathcal{H})$ is equivalent to the fairness-constrained optimization problem* (1).

The proof is in Appendix B.

We will be interested in the structure and properties of the optimal solution to $\psi(f, \gamma, \mathcal{H})$, which will be elucidated via its Lagrangian. Note that the optimization problem has $2|\mathcal{G}|$ linear inequality constraints. Let $\lambda = \{\lambda_g^{\pm}\}_{g \in \mathcal{G}}$ denote the vector of $2|\mathcal{G}|$ dual variables corresponding to those constraints, and write $\lambda_g = \lambda_g^+ - \lambda_g^-$.

Definition 8 (Lagrangian)**.** *Given any regression function f, we define a Lagrangian of the optimization problem $\psi(f, \gamma, \mathcal{H})$ as $L_f : \mathcal{H} \times \mathbb{R}^{2|\mathcal{G}|} \to \mathbb{R}$:*

$$\begin{aligned} L_f(h, \lambda) = \mathop{\mathbb{E}}_{x \sim \mathcal{D}_{\mathcal{X}}} \Big[& f(x)\ell(h(x),1) + (1 - f(x))\ell(h(x),0) \\ & + \sum_{g \in \mathcal{G}} \lambda_g^+ \big(\ell(h(x),0)g(x)(1-f(x)) - \beta_g \ell(h(x),0)(1-f(x)) - \gamma \big) \\ & + \sum_{g \in \mathcal{G}} \lambda_g^- \big(\beta_g \ell(h(x),0)(1-f(x)) - \ell(h(x),0)g(x)(1-f(x)) - \gamma \big) \Big] \end{aligned}$$

For convenience, given a Bayes optimal regressor f^*, we write $L^* = L_{f^*}$. Similarly, given some other regressor $\hat{f}$, we write $\hat{L} = L_{\hat{f}}$.

Let $\mathcal{H}_A = 2^{\mathcal{X}}$ be the set of all Boolean functions $f : \mathcal{X} \to \{0,1\}$. We will consider solving our optimization problem over this set of functions $\mathcal{H}_A$.

Definition 9 (Optimal post-processed classifier)**.** *We say that a classifier h_f is an optimal post-processing of f if there exists a vector λ^f such that the following primal/dual optimality conditions are simultaneously met:*

$$h_f(x) \in \arg\min_{h \in \mathcal{H}_A} L_f(h, \lambda^f) \quad \lambda^f \in \arg\max_{\lambda \in \mathbb{R}^{2|\mathcal{G}|}} L_f(h_f, \lambda).$$

For convenience, we write

$$h^*(x) = h_{f^*}(x) \quad \textit{and} \quad \lambda^* = \lambda^{f^*}$$
$$\hat{h}(x) = h_{\hat{f}}(x) \quad \textit{and} \quad \hat{\lambda} = \lambda^{\hat{f}}$$

where f^* is the Bayes optimal regressor and $\hat{f}$ is any other regressor. We will write λ^*_g and $\hat{\lambda}_g$ to refer to the dual variable in λ^* and $\hat{\lambda}$ for group g, respectively. We observe that as the optimal solution to the Lagrangian minimax optimization problem, $h^*(x)$ is the Bayes optimal classifier subject to the fairness constraints in 1.

Lemma 2. *The optimal post-processed classifier h of $\psi(f, \gamma, \mathcal{H}_A)$ for some regressor f takes the following form:*

$$h(x) = \begin{cases} 1, & \text{if } f(x) > \frac{1+\sum_{g\in\mathcal{G}} \lambda_g(g(x)-\beta_g)}{2+\sum_{g\in\mathcal{G}} \lambda_g(g(x)-\beta_g)} \text{ and } 2+\sum_{g\in\mathcal{G}} \lambda_g(g(x)-\beta_g) > 0, \\ 0, & \text{if } f(x) < \frac{1+\sum_{g\in\mathcal{G}} \lambda_g(g(x)-\beta_g)}{2+\sum_{g\in\mathcal{G}} \lambda_g(g(x)-\beta_g)} \text{ and } 2+\sum_{g\in\mathcal{G}} \lambda_g(g(x)-\beta_g) > 0, \\ 1, & \text{if } f(x) < \frac{1+\sum_{g\in\mathcal{G}} \lambda_g(g(x)-\beta_g)}{2+\sum_{g\in\mathcal{G}} \lambda_g(g(x)-\beta_g)} \text{ and } 2+\sum_{g\in\mathcal{G}} \lambda_g(g(x)-\beta_g) < 0, \\ 0, & \text{if } f(x) > \frac{1+\sum_{g\in\mathcal{G}} \lambda_g(g(x)-\beta_g)}{2+\sum_{g\in\mathcal{G}} \lambda_g(g(x)-\beta_g)} \text{ and } 2+\sum_{g\in\mathcal{G}} \lambda_g(g(x)-\beta_g) < 0. \end{cases}$$

In the edge case in which $f(x) = \frac{1+\sum_{g\in\mathcal{G}} \lambda_g(g(x)-\beta_g)}{2+\sum_{g\in\mathcal{G}} \lambda_g(g(x)-\beta_g)}$, $h(x)$ could take either value and might be randomized.

The proof is in Appendix B.

3.1 Computing the optimally post-processed classifier

To approximate h given f, we need to compute an approximately optimal solution to the linear program $\psi(f, \gamma, \mathcal{H}_A)$. We can do this by playing a no-regret algorithm over the dual variables λ and best response over the primal variables as defined in Definition 23. We can approximate the losses to the no-regret algorithm from a finite sample of unlabelled data of size scaling logarithmically in the number of constraints and linearly in the number of rounds T of the no regret dynamics, which using standard techniques we can show yields an approximately optimal solution to the original LP. The algorithm is described in Algorithm 1. We state its approximate guarantees then spend the rest of this section formalizing the structure necessary for the result.

Theorem 1. *Let* OPT *be the objective value of the optimal solution to $\psi(f, \gamma, \mathcal{H}_A)$. Then, for any $C \in \mathbb{R}$, after $T = \frac{1}{4} \cdot C^2 \cdot \left(C^2 + 4|\mathcal{G}|\right)^2$ iterations, Algorithm 1 outputs a randomized hypothesis $\bar{h}$ such that $err(\bar{h}) \leq \text{OPT} + \frac{2}{C}$ and $w_g|\rho_g(\bar{h}) - \rho(\bar{h})| \leq \gamma + \frac{1}{C} + \frac{2}{C^2}$.*

In order to prove Theorem 1 (which is proved fully in Appendix B),we first must specify the game formulation of the problem and demonstrate that constraining the dual player still allows for an adequate approximation to the original problem.

Game formulation We pose the optimization of our original linear program as a zero-sum game between a primal (minimization) player who plays over the set of hypotheses and a dual (maximization) player who plays over the set of dual variables. The utility function of the game is the Lagrangian of our linear program as stated in Definition 8. The value of this game is given by

$$\min_{h \in \Delta\mathcal{H}} \max_{\lambda \in \mathbb{R}^{2|\mathcal{G}|}} L_f(h, \lambda).$$

Constraining the linear program In order to compute an approximate minimax solution to this game, we need to constrain the strategy space of the dual player. That is, we need to bound the dual space to a region $\Lambda = \left\{\lambda \in \mathbb{R}^{2\mathcal{G}} \middle| \|\lambda\|_1 \leq C\right\}$. We call this constrained version of the problem the Λ-bounded Lagrangian problem, which has value

$$\min_{h \in \Delta\mathcal{H}} \max_{\lambda : |\lambda|_1 \leq C} L_f(h, \lambda). \tag{2}$$

We can apply the minimax theorem to this bounded game to see:

$$\min_{h \in \Delta\mathcal{H}} \max_{\lambda : |\lambda|_1 \leq C} L_f(h, \lambda) \equiv \max_{\lambda : |\lambda|_1 \leq C} \min_{h \in \Delta\mathcal{H}} L_f(h, \lambda).$$

We will only be able to achieve an approximate solution to the problem, which we define as follows.

Definition 10. *We say that (h, λ) is a v-approximate minimax solution to the Λ-bounded Lagrangian problem L_f if $L_f(h, \lambda) \leq \min_{h' \in \Delta\mathcal{H}} L_f(h', \lambda) + v$ and $L_f(h, \lambda) \geq \max_{\lambda' \in \Lambda} L_f(h, \lambda') - v$.*

An approximate minimax solution to this bounded version of the problem is also an approximate solution to the original problem we described in Equation 1:

Theorem 2. *[Kearns et al. [2018]] Let (h, λ) be a v-approximate minimax solution to the Λ-bounded Lagrangian problem L_f and let* OPT *be the objective value of the optimal solution to $\psi(f, \gamma, \mathcal{H}_A)$. Then, $err(h) \leq OPT + 2v$, and $\forall g \in \mathcal{G}, w_g|\rho_g(h) - \rho(h)| \leq \gamma + (1 + 2v)/C$.*

Approximate equilibrium of the constrained game Now, we can proceed with no-regret play to find an approximate solution to the game. The dual player will play projected gradient descent over their vector λ and the primal player will best respond, as described in Algorithm 1.

Theorem 3. *Algorithm 1 returns an $\epsilon-$approximate equilibrium solution to the zero-sum game defined by Equation 2 after $T = \frac{1}{4\epsilon^2}\left(\frac{1}{\epsilon^2} + 4|\mathcal{G}|\right)^2$ rounds.*

The proof of Theorem 3 is in Appendix B. Combining Theorem 2 and Theorem 3 gives us the proof of Theorem 1, which appears in Appendix B.

3.2 Beginning with a Multicalibrated Regression Function $\hat{f}$

Thus far, we have considered the optimization problem $\psi(f, \gamma, \mathcal{H}_A)$ in the abstract, have characterized its optimal solution h, and have given a simple algorithm to find $\bar{h}$, an approximately optimal solution. When $f = f^*$, $h = h^*$ is the Bayes optimal fair classifier, and $\bar{h}$ is approximately Bayes optimal. But in practice, we will not have access to f^*, but will instead only have some surrogate function, which we will call $\hat{f}(x)$. We will argue that if $\hat{f}$ is appropriately *multicalibrated*, then it is good enough for our purposes. We will compare the approximate solution $\bar{h}$ produced by Algorithm 1 to the optimization problem $\psi(\hat{f}, \gamma, \mathcal{H}_A)$ which has corresponding Lagrangian $\hat{L}(\hat{h}, \hat{\lambda})$, as defined in Definition 8 to the optimal solution (h^*, λ^*) to the optimization problem $\psi(f^*, \gamma, \mathcal{H})$ for some constrained class $\mathcal{H}$, and show conditions under which they are close.

In order to proceed, we first need to determine what our surrogate function ought to be multicalibrated with respect to. In addition to being α-approximately multicalibrated in expectation with respect to $\mathcal{G}$ and $\mathcal{H}$, we will require that $\hat{f}$ be α-approximately multicalibrated with respect to $\mathcal{G} \times \mathcal{H} = \{g(x) \cdot h(x) | g \in \mathcal{G}, h \in \mathcal{H}\}$. Furthermore, we will need to require that $\hat{f}$ be α-approximately *jointly* multicalibrated in expectation with respect to a set of thresholding functions, defined below:

Definition 11 (Set of thresholding functions $\mathcal{B}(C)$)**.** *Let $x_{\mathcal{G}} \in \{0, 1\}^{|\mathcal{G}|}$ denote the group membership indicator vector of some point x. Define the function*

$$d(v) := \frac{2v - 1}{1 - v}.$$

Algorithm 1: Projected Gradient Descent Algorithm

Input: (*D: dataset, $f : \mathcal{X} \to [0,1]$: regression function, $\mathcal{G}$: groups, γ: tolerance on fairness violation, C: bound on dual ($\|\lambda\|_1 \le C$), η: learning rate*)

Initialize dual vector $\lambda^0 = \mathbf{0}$ and set $T = \frac{1}{4} \cdot C^2 \cdot (C^2 + 4|\mathcal{G}|)^2$.

for $t = 1, \ldots, T$ **do**

Primal player updates h_t

$$h_t(x) = \begin{cases} 1, & \text{if } f(x) \ge \frac{1+\sum_{g\in\mathcal{G}} \lambda_g^{t-1}(g(x)-\beta_g)}{2+\sum_{g\in\mathcal{G}} \lambda_g^{t-1}(g(x)-\beta_g)} \text{ and } 2+\sum_{g\in\mathcal{G}} \lambda_g^{t-1}(g(x)-\beta_g) > 0, \\ 0, & \text{if } f(x) < \frac{1+\sum_{g\in\mathcal{G}} \lambda_g^{t-1}(g(x)-\beta_g)}{2+\sum_{g\in\mathcal{G}} \lambda_g^{t-1}(g(x)-\beta_g)} \text{ and } 2+\sum_{g\in\mathcal{G}} \lambda_g^{t-1}(g(x)-\beta_g) > 0, \\ 1, & \text{if } f(x) \le \frac{1+\sum_{g\in\mathcal{G}} \lambda_g^{t-1}(g(x)-\beta_g)}{2+\sum_{g\in\mathcal{G}} \lambda_g^{t-1}(g(x)-\beta_g)} \text{ and } 2+\sum_{g\in\mathcal{G}} \lambda_g^{t-1}(g(x)-\beta_g) < 0, \\ 0, & \text{if } f(x) > \frac{1+\sum_{g\in\mathcal{G}} \lambda_g^{t-1}(g(x)-\beta_g)}{2+\sum_{g\in\mathcal{G}} \lambda_g^{t-1}(g(x)-\beta_g)} \text{ and } 2+\sum_{g\in\mathcal{G}} \lambda_g^{t-1}(g(x)-\beta_g) < 0, \\ 1, & \text{if } 2+\sum_{g\in\mathcal{G}} \lambda_g^{t-1}(g(x)-\beta_g) = 0 \end{cases}$$

Compute

$$\hat{\rho}_g^t = \mathop{\mathbb{E}}_{(x,y)\sim D}[\ell(h_t(x),0)g(x)(1-f(x))] \text{ for all } g \in \mathcal{G},$$
$$\hat{\rho}^t = \mathop{\mathbb{E}}_{(x,y)\sim D}[\beta_g \ell(h_t(x),0)(1-f(x))], \text{ where } \beta_g = \Pr[g(x)=1|y=0]$$

Dual player updates

$$\lambda_g^{t,+} = \max(0, \lambda_g^{t,+} + \eta \cdot (\hat{\rho}_g^t - \hat{\rho}^t - \gamma)),$$
$$\lambda_g^{t,-} = \max(0, \lambda_g^{t,-} + \eta \cdot (\hat{\rho}^t - \hat{\rho}_g^t - \gamma)).$$

Dual player sets $\lambda^t = \sum_{g\in\mathcal{G}} \lambda_g^{t,+} - \lambda_g^{t,-}$.

If $\|\lambda^t\|_1 > C$, set $\lambda^t = \arg\min_{\{\tilde{\lambda}\in\mathbb{R}^{2\mathcal{G}} | \|\tilde{\lambda}\|_1 \le C\}} \|\lambda_t - \tilde{\lambda}\|_2^2$.

end

Output: $\bar{h} := \frac{1}{T}\sum_{t=1}^T \hat{h}_t$, a uniformly random classifier over all rounds' hypotheses.

Then, let for any λ, x, β

$$s_\lambda(x, v) := \mathbb{1}[\langle \lambda, x_{\mathcal{G}} - \beta \rangle \geq d(v)].$$

Define $\mathcal{B}(C) = \{s_\lambda | \lambda \in \Lambda(C), \beta = \beta_{g_1}, \ldots, \beta_{g_{|\mathcal{G}|}}\}$*, where* $\Lambda(C) = \left\{\lambda \in \mathbb{R}^{2\mathcal{G}} \middle| \|\lambda\|_1 \leq C\right\}$*, as defined in Equation 2 and* $\beta_g = \Pr_{(x,y)\sim\mathcal{D}}[g(x) = 1 | y = 0]$*, as defined in Definition 7.*

Remark. *When the groups of interest are disjoint, joint multicalibraiton with respect to this class* $\mathcal{B}(C)$ *is implied by multicalibration with respect to* $\mathcal{G}$*. But when the groups can intersect, this is not an implication, and asking for joint multicalibration with respect to* $\mathcal{B}(C)$ *adds new constraints on* $\hat{f}$*.*

Informally, these functions take an example, and map it to a vector of its group membership, indicating whether a λ-weighting of the example's group membership is larger than some threshold $d(v)$. We will need joint multicalibration with respect to such functions in order to relate the estimated error of $\hat{h}$ to its true error. These thresholding functions $\mathcal{B}(C)$ have a natural relationship to the deterministic thresholded models h_t that we compute at each round of Algorithm 1:

Lemma 3. *Let* h_t *be the response to* λ^{t-1} *described in Algorithm 1 at some round* $t \in [T]$*. Then,*

$$h_t(x) = s_{\lambda^{t-1}}(x, f(x)).$$

The proof is in Appendix B. We verify in Appendix C that a variant of the multicalibration algorithms given in Hébert-Johnson et al. [2018], Gopalan et al. [2022] can guarantee joint multicalibration with respect to $\mathcal{B}(C)$ as well.

With these preliminaries behind us, we can now state our main theorem, which shows that for any class of models $\mathcal{H}$ and class of groups $\mathcal{G}$, given an appropriately multicalibrated $\hat{f}$ (with multicalibration requirements depending on $\mathcal{H}$ and $\mathcal{G}$), the model $\bar{h}$ output by Algorithm 1 achieves an error rate and fairness guarantees comparable to the optimal solution to $\psi(f^*, \gamma, \mathcal{H})$:

Theorem 4. *Set* $C = \sqrt{1/\alpha}$*. Let* $\hat{f}$ *be* α*-approximately multicalibrated in expectation with respect to* $\mathcal{G}$*,* $\mathcal{H}$*, and* $\mathcal{G} \times \mathcal{H}$ *and* α*-approximately jointly multicalibrated in expectation with respect to* $\mathcal{B}(C)$*. Let* $\bar{h}$ *be the result of running Algorithm 1 with input* $\hat{f}$ *and* C*. Then,* $err(\bar{h}) \leq err(h^*) + \alpha(5 + 2\sqrt{1/\alpha}) + 2\sqrt{\alpha}$*, and for all* $g \in \mathcal{G}, w_g \left|\rho_g(\bar{h}) - \rho(\bar{h})\right| \leq w_g \left|\rho_g(h^*) - \rho(h^*)\right| + w_g\alpha$*.*

Proof Sketch: Generalizing notation from the previous sections, let $\text{err}(h) = \mathbb{E}_{x\sim\mathcal{D}_\mathcal{X}}[f^*(x)\ell(h(x), 1) + (1 - f^*(x))\ell(h(x), 0)]$ denote the true error of h on the distribution (i.e. as measured according to the true conditional label distribution f^*), and let $\widehat{\text{err}}(h) = \mathbb{E}_{x\sim\mathcal{D}_\mathcal{X}}[\hat{f}(x)\ell(h(x), 1) + (1 - \hat{f}(x))\ell(h(x), 0)]$ denote the error of h as estimated using the surrogate function $\hat{f}$. At a high level, the proof of Theorem 4 will proceed as follows:

$$\begin{aligned}
\text{err}(h^*) &= L^*(h^*, \lambda^*) && (3)\\
&\geq L^*(h^*, \hat{\lambda}) && (4)\\
&\approx \hat{L}(h^*, \hat{\lambda}) && (5)\\
&\geq \hat{L}(\hat{h}, \hat{\lambda}) && (6)\\
&= \widehat{\text{err}}(\hat{h}) && (7)\\
&\approx \widehat{\text{err}}(\bar{h}) && (8)\\
&\approx \text{err}(\bar{h}). && (9)
\end{aligned}$$

Each of these steps takes a lemma (presented in full in the appendix) to justify, but the logic is at a high level as follows: The equalities on lines 3 and 7 follow from complimentary slackness: at the optimal solution (h^*, λ^*) it must be that for each constraint g either the constraint is exactly tight so that its "violation" term in the Lagrangian evaluates to 0, or its corresponding dual variable $\lambda_g^\pm = 0$. Thus, all terms in the Lagrangian other than the objective evaluate to 0. The inequality in line 4 follows from the dual optimality condition that

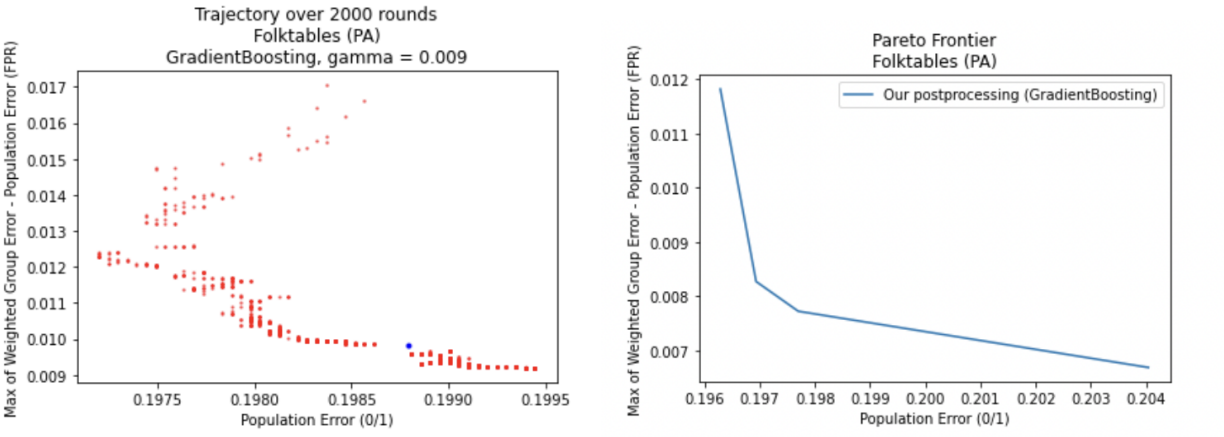

Figure 1: The plot on the left is a trajectory over 2000 iterations of gradient descent of our method post-processing a base model of gradient-boosted regression trees, for a single value of $\gamma = 0.01$. The trajectory starts at the top of the figure and moves downwards with time, and the blue point represents the uniform distribution over the constituent models of the 2000 iterations. The blue point, whose error is 0.1987 and maximum violation of group error - population error is 0.0098, shows our method limits the maximum violation to γ. The plot on the right shows the pareto curve for our method for constraint values ranging between $0.003 \leq \gamma \leq 0.01$, showing that large reductions in false positive rate disparities cost modestly in error.

$\lambda^* \in \arg\max_\lambda L^*(h^*, \lambda)$ and similarly the inequality in line 6 follows from the primal optimality condition that $\hat{h} \in \arg\min_{h \in \mathcal{H}_A} \hat{L}(h, \hat{\lambda})$. Line 8 follows from the fact that $\bar{h}$ is an approximately optimal solution to $\psi(\hat{f}, \gamma, \mathcal{H}_A)$. Steps 5 and 9 follow from our multicalibration guarantees, the former from multicalibration with respect to groups and our hypothesis class, and the latter from joint multicalibration with respect to the set of thresholding functions from Definition 11. The complete proof is found in Appendix B.

4 Experiments

In this section, we evaluate our post-processing algorithm on a dataset derived from Pennsylvania Census data provided by the Folktables package [Ding et al., 2021], which we use under its MIT license. The sensitive attributes we use from the dataset are binarized gender and the re-coded detailed race code (RAC1P) to create two classes of overlapping groups. We run our algorithm on top of a regression function $\hat{f}$ trained using the sklearn gradient-boosted decision trees package — notably it is *not* guaranteed to be multicalibrated in any of the ways our theorems require! Nevertheless our experiments bear out that our post-processing method performs well even on top of off-the-shelf regression methodologies. We expand on our experimental investigation in Appendix E.

The experimental findings we present support the theoretical analysis that the algorithm quickly converges to classifier approximately satisfying our fairness constraints.

We emphasize that our post-processing method is extremely lightweight. As a primal/dual algorithm, it is very similar in structure to the "fair reductions" method of [Agarwal et al., 2018]. However where [Agarwal et al., 2018] needs to solve an ERM problem at every iteration and then evaluate the performance of the resulting trained model, we entirely skip the ERM step and need only evaluate the performance of a thresholded classifier which we have in closed form.

Acknowledgements

This work was supported in part by NSF grants AF-1763307, CCF-2217062, and FAI-2147212 and a grant from the Simons Foundation.

References

Alekh Agarwal, Alina Beygelzimer, Miroslav Dudík, John Langford, and Hanna Wallach. A reductions approach to fair classification. In *International Conference on Machine Learning*, pages 60–69. PMLR, 2018.

Ibrahim M Alabdulmohsin and Mario Lucic. A near-optimal algorithm for debiasing trained machine learning models. *Advances in Neural Information Processing Systems*, 34:8072–8084, 2021.

Raef Bassily, Kobbi Nissim, Adam Smith, Thomas Steinke, Uri Stemmer, and Jonathan Ullman. Algorithmic stability for adaptive data analysis. In *Proceedings of the forty-eighth annual ACM symposium on Theory of Computing*, pages 1046–1059, 2016.

Maya Burhanpurkar, Zhun Deng, Cynthia Dwork, and Linjun Zhang. Scaffolding sets. *arXiv preprint arXiv:2111.03135*, 2021.

Sam Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd acm sigkdd international conference on knowledge discovery and data mining*, pages 797–806, 2017.

Emily Diana, Wesley Gill, Michael Kearns, Krishnaram Kenthapadi, Aaron Roth, and Saeed Sharifi-Malvajerdi. Multiaccurate proxies for downstream fairness. *arXiv preprint arXiv:2107.04423*, 2021.

Frances Ding, Moritz Hardt, John Miller, and Ludwig Schmidt. Retiring adult: New datasets for fair machine learning. *Advances in Neural Information Processing Systems*, 34, 2021.

Dheeru Dua and Casey Graff. UCI machine learning repository, 2017. URL `http://archive.ics.uci.edu/ml`.

Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, pages 214–226, 2012.

Cynthia Dwork, Vitaly Feldman, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Aaron Leon Roth. Preserving statistical validity in adaptive data analysis. In *Proceedings of the forty-seventh annual ACM symposium on Theory of computing*, pages 117–126, 2015.

Cynthia Dwork, Michael P Kim, Omer Reingold, Guy N Rothblum, and Gal Yona. Outcome indistinguishability. In *Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing*, pages 1095–1108, 2021.

Yoav Freund and Robert E. Schapire. Game theory, on-line prediction and boosting. In *Proceedings of the Ninth Annual Conference on Computational Learning Theory*, COLT '96, page 325–332, New York, NY, USA, 1996. Association for Computing Machinery. ISBN 0897918118. doi: 10.1145/238061.238163. URL `https://doi.org/10.1145/238061.238163`.

Parikshit Gopalan, Adam Tauman Kalai, Omer Reingold, Vatsal Sharan, and Udi Wieder. Omnipredictors. In *13th Innovations in Theoretical Computer Science Conference (ITCS 2022)*. Schloss Dagstuhl-Leibniz-Zentrum für Informatik, 2022.

Varun Gupta, Christopher Jung, Georgy Noarov, Mallesh M Pai, and Aaron Roth. Online multivalid learning: Means, moments, and prediction intervals. In *13th Innovations in Theoretical Computer Science Conference (ITCS 2022)*. Schloss Dagstuhl-Leibniz-Zentrum für Informatik, 2022.

Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29, 2016.

Ursula Hébert-Johnson, Michael Kim, Omer Reingold, and Guy Rothblum. Multicalibration: Calibration for the (computationally-identifiable) masses. In *International Conference on Machine Learning*, pages 1939–1948. PMLR, 2018.

Lunjia Hu, Inbal Livni-Navon, Omer Reingold, and Chutong Yang. Omnipredictors for constrained optimization. In *Manuscript*, 2022.

Christopher Jung, Changhwa Lee, Mallesh Pai, Aaron Roth, and Rakesh Vohra. Moment multicalibration for uncertainty estimation. In *Conference on Learning Theory*, pages 2634–2678. PMLR, 2021a.

Christopher Jung, Katrina Ligett, Seth Neel, Aaron Roth, Saeed Sharifi-Malvajerdi, and Moshe Shenfeld. A new analysis of differential privacy's generalization guarantees. In *Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing*, pages 9–9, 2021b.

Michael Kearns, Seth Neel, Aaron Roth, and Zhiwei Steven Wu. Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. In *International Conference on Machine Learning*, pages 2564–2572. PMLR, 2018.

Michael P Kim, Amirata Ghorbani, and James Zou. Multiaccuracy: Black-box post-processing for fairness in classification. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pages 247–254, 2019.

Michael P Kim, Christoph Kern, Shafi Goldwasser, Frauke Kreuter, and Omer Reingold. Universal adaptability: Target-independent inference that competes with propensity scoring. *Proceedings of the National Academy of Sciences*, 119(4), 2022.

Dennis Wei, Karthikeyan Natesan Ramamurthy, and Flavio P Calmon. Optimized score transformation for consistent fair classification. *J. Mach. Learn. Res.*, 22:258–1, 2021.

Blake Woodworth, Suriya Gunasekar, Mesrob I Ohannessian, and Nathan Srebro. Learning non-discriminatory predictors. In *Conference on Learning Theory*, pages 1920–1953. PMLR, 2017.

Forest Yang, Mouhamadou Cisse, and Sanmi Koyejo. Fairness with overlapping groups; a probabilistic perspective. *Advances in neural information processing systems*, 33:4067–4078, 2020.

Shengjia Zhao, Michael Kim, Roshni Sahoo, Tengyu Ma, and Stefano Ermon. Calibrating predictions to decisions: A novel approach to multi-class calibration. *Advances in Neural Information Processing Systems*, 34, 2021.

Martin Zinkevich. Online convex programming and generalized infinitesimal gradient ascent. In *Proceedings of the 20th international conference on machine learning (icml-03)*, pages 928–936, 2003.

A Generalization to other fairness notions

A.1 False Negative (FN) Fairness

Definition 12. *The false negative rate of a classifier $h : \mathcal{X} \to \mathcal{Y}$ on a group g is:*

$$\rho_{FN}(h, g, \mathcal{D}) = \Pr_{(x,y)\sim\mathcal{D}}[h(x) \neq y | y = 1, g(x) = 1]$$

When h is a randomized classifier, the probabilities are computed over the randomness of h as well. $\rho_g^{FN}(h) \equiv \rho_{FN}(h, g, \mathcal{D})$, and $\rho_{FN}(h) \equiv \rho(h, I, \mathcal{D})$.

Definition 13. *We say that classifier $h : \mathcal{X} \to \mathcal{Y}$ satisfies γ-False Negative (FN) Fairness with respect to $\mathcal{D}$ and $\mathcal{G}$ if for all $g \in \mathcal{G}$,*

$$w_g^{FN} \left|\rho_g^{FN}(h) - \rho_{FN}(h)\right| \leq \gamma.$$

where $w_g^{FN} = \Pr_{(x,y)\sim\mathcal{D}}[g(x) = 1, y = 1]$.

We consider the following fairness-constrained optimization problem:

$$\begin{aligned} &\min_{h\in\Delta\mathcal{H}} && \mathrm{err}(h) \\ &\text{s.t. for each } g \in \mathcal{G}: && w_g^{\mathrm{FN}} |\rho_g^{\mathrm{FN}}(h) - \rho_{\mathrm{FN}}(h)| \leq \gamma, \end{aligned} \tag{10}$$

Definition 14. *Let $f : \mathcal{X} \to R \subseteq [0, 1]$ be some regression function and let $\gamma \in \mathbb{R}_+$. Define $\psi_{FN}(f, \gamma, \mathcal{H})$ to be the following optimization problem:*

$$\begin{aligned} &\min_{h\in\Delta\mathcal{H}} && \mathop{\mathbb{E}}_{x\sim\mathcal{D}_\mathcal{X}}[f(x)\ell(h(x), 1) + (1 - f(x))\ell(h(x), 0)] \\ &\textit{s.t. for each } g \in \mathcal{G}: && \left|\mathbb{E}[\ell(h(x), 1)g(x)f(x)] - \beta_g^{FN}\mathbb{E}\left[\ell(h(x), 1)f(x)\right]\right| \leq \gamma, \end{aligned}$$

where $\beta_g^{FN} = \Pr[g(x) = 1 | y = 1]$.

Lemma 4. *Let f^* be the Bayes optimal regression function over $\mathcal{D}$. Then optimization problem $\psi_{FN}(f^*, \gamma, \mathcal{H})$ is equivalent to the fairness-constrained optimization problem 10.*

Proof. Note that the objective function is equivalent to that of Equation 1, and hence proof of the objectives being equivalent is identical to that of Lemma 1. For the constraints, note that

$$\begin{aligned} w_g^{\mathrm{FN}}|\rho_g^{\mathrm{FN}}(h) - \rho_{\mathrm{FN}}(h)| &= \Pr[g(x) = 1, y = 1]\,|\Pr[h(x) = 0 | g(x) = 1, y = 1] - \Pr[h(x) = 0 | y = 1]| \\ &= \Pr[g(x) = 1, y = 1]\left|\frac{\Pr[h(x) = 0, g(x) = 1, y = 1]}{\Pr[g(x) = 1, y = 1]} - \frac{\Pr[h(x) = 0, y = 1]}{\Pr[Y = 1]}\right| \\ &= \left|\Pr[h(x) = 0, g(x) = 1, y = 1] - \frac{\Pr[g(x) = 1, y = 1]\Pr[h(x) = 0, y = 1]}{\Pr[Y = 1]}\right| \\ &= \left|\mathbb{E}[\ell(h(x), 1)g(x)f^*(x)] - \frac{\Pr[g(x) = 1, y = 1]}{\Pr[Y = 1]}\mathbb{E}\left[\ell(h(x), 1)f^*(x)\right]\right| \\ &= |\mathbb{E}[\ell(h(x), 1)g(x)f^*(x)] - \Pr[g(x) = 1 | Y = 0]\mathbb{E}\left[\ell(h(x), 1)f^*(x)\right]| \\ &= \left|\mathbb{E}[\ell(h(x), 1)g(x)f^*(x)] - \beta_g^{\mathrm{FN}}\mathbb{E}\left[\ell(h(x), 1)f^*(x)\right]\right|. \end{aligned}$$

The result follows. □

Definition 15 (Lagrangian)**.** *Given any regression function f, we define a Lagrangian of the optimization problem $\psi_{FN}(f, \gamma, \mathcal{H})$ as $L_f^{FN} : \mathcal{H} \times \mathbb{R}^{2|\mathcal{G}|} \to \mathbb{R}$:*

$$\begin{aligned} L_f^{FN}(h, \lambda) = \mathop{\mathbb{E}}_{x \sim \mathcal{D}_X} \Big[& f(x)\ell(h(x), 1) + (1 - f(x))\ell(h(x), 0) \\ & + \sum_{g \in \mathcal{G}} \lambda_g^+ \big(\ell(h(x), 1)g(x)f(x) - \beta_g \ell(h(x), 1)f(x) - \gamma\big) \\ & + \sum_{g \in \mathcal{G}} \lambda_g^- \big(\beta_g \ell(h(x), 1)f(x) - \ell(h(x), 1)g(x)f(x) - \gamma\big) \Big] \end{aligned}$$

Lemma 5.

$$\begin{aligned} L_f^{FN}(h, \lambda) = \mathbb{E}_{x \sim \mathcal{D}_X} \Bigg[& \ell(h(x), 0) - \gamma \sum_{g \in \mathcal{G}} (\lambda_g^+ + \lambda_g^-) \\ & + f(x) \left(-\ell(h(x), 0) + \ell(h(x), 1) \left(1 + \sum_{g \in \mathcal{G}} \lambda_g (g(x) - \beta_g^{FN}) \right) \right) \Bigg] \end{aligned}$$

Proof. Distributing out like terms in the expression for the Lagrangian in Definition 15 gives us

$$\begin{aligned} L_f(h, \lambda) &= \mathbb{E}_{x \sim \mathcal{D}_X} \Bigg[\ell(h(x), 0) - \gamma \sum_{g \in \mathcal{G}} (\lambda_g^+ + \lambda_g^-) \\ &\quad + f(x) \left(\ell(h(x), 1) - \ell(h(x), 0) + \ell(h(x), 1) \sum_{g \in \mathcal{G}} (\lambda_g^+ (g(x) - \beta_g) + \lambda_g^- (\beta_g - g(x)) \right) \Bigg] \\ &= \mathbb{E}_{x \sim \mathcal{D}_X} \Bigg[\ell(h(x), 0) - \gamma \sum_{g \in \mathcal{G}} (\lambda_g^+ + \lambda_g^-) \\ &\quad + f(x) \left(-\ell(h(x), 0) + \ell(h(x), 1) \left(1 + \sum_{g \in \mathcal{G}} (\lambda_g^+ - \lambda_g^-)(g(x) - \beta_g) \right) \right) \Bigg]. \end{aligned}$$

Recall that $\lambda_g = \lambda_g^+ - \lambda_g^-$, so we are done. □

Lemma 6. *The optimal post-processed classifier h of $\psi(f, \gamma, \mathcal{H}_A)$ for some regressor f takes the following form:*

$$h(x) = \begin{cases} 1, & \text{if } f(x) > \frac{1}{2 + \sum_{g \in \mathcal{G}} \lambda_g (g(x) - \beta_g)} \text{ and } 2 + \sum_{g \in \mathcal{G}} \lambda_g (g(x) - \beta_g) > 0, \\ 0, & \text{if } f(x) < \frac{1}{2 + \sum_{g \in \mathcal{G}} \lambda_g (g(x) - \beta_g)} \text{ and } 2 + \sum_{g \in \mathcal{G}} \lambda_g (g(x) - \beta_g) > 0, \\ 1, & \text{if } f(x) < \frac{1}{2 + \sum_{g \in \mathcal{G}} \lambda_g (g(x) - \beta_g)} \text{ and } 2 + \sum_{g \in \mathcal{G}} \lambda_g (g(x) - \beta_g) < 0, \\ 0, & \text{if } f(x) > \frac{1}{2 + \sum_{g \in \mathcal{G}} \lambda_g (g(x) - \beta_g)} \text{ and } 2 + \sum_{g \in \mathcal{G}} \lambda_g (g(x) - \beta_g) < 0. \end{cases}$$

In the edge case in which $f(x) = \frac{1}{2 + \sum_{g \in \mathcal{G}} \lambda_g (g(x) - \beta_g)}$, $h(x)$ could take either value and might be randomized.

Proof. Note that since we are optimizing over the set of all binary classifiers, h optimizes the Lagrangian objective pointwise for every x. In particular, we have from Lemma 5 that:

$$h(x) = \arg\min_p \left[\ell(p, 0) + f(x) \left(-\ell(p, 0) + \ell(p, 1) \left(1 + \sum_{g \in \mathcal{G}} \lambda_g (g(x) - \beta_g) \right) \right) \right].$$

In order to determine the threshold, we need to check when setting $p = 1$ leads to a value less than setting $p = 0$. In other words, we need to solve for $f(x)$ when

$$1 - f(x) < f(x)\left(1 + \sum_{g\in\mathcal{G}} \lambda_g(g(x) - \beta_g)\right)$$
$$\Rightarrow f(x) > \frac{1}{2 + \sum_{g\in\mathcal{G}} \lambda_g(g(x) - \beta_g)}.$$

Thus,

$$h(x) = \begin{cases} 1, & \text{if } f(x) > \frac{1}{2+\sum_{g\in\mathcal{G}} \lambda_g(g(x)-\beta_g)} \text{ and } 2 + \sum_{g\in\mathcal{G}} \lambda_g(g(x) - \beta_g) > 0, \\ 0, & \text{if } f(x) < \frac{1}{2+\sum_{g\in\mathcal{G}} \lambda_g(g(x)-\beta_g)} \text{ and } 2 + \sum_{g\in\mathcal{G}} \lambda_g(g(x) - \beta_g) > 0, \\ 1, & \text{if } f(x) < \frac{1}{2+\sum_{g\in\mathcal{G}} \lambda_g(g(x)-\beta_g)} \text{ and } 2 + \sum_{g\in\mathcal{G}} \lambda_g(g(x) - \beta_g) < 0, \\ 0, & \text{if } f(x) > \frac{1}{2+\sum_{g\in\mathcal{G}} \lambda_g(g(x)-\beta_g)} \text{ and } 2 + \sum_{g\in\mathcal{G}} \lambda_g(g(x) - \beta_g) < 0. \end{cases}$$

□

From Lemma 6, we can now define a best-response model and use Algorithm 2 to generate an optimally post-processed model that preserves $\gamma-$False Negative fairness. The algorithm's error bounds may be derived using symmetric arguments to sections 3.1 and 3.2, where $\hat{f}$ is required to be α-approximately jointly multicalibrated in expectation with respect to $s_\lambda(x, v) := \mathbb{1}[\langle\lambda, x_\mathcal{G} - \beta\rangle \geq (1 - 2v)/v]$ following the same arguments as used in Lemma 3.

A.2 Error Fairness

Definition 16. *We say that classifier $h : \mathcal{X} \to \mathcal{Y}$ satisfies γ-Error (E) Fairness with respect to $\mathcal{D}$ and $\mathcal{G}$ if for all $g \in \mathcal{G}$,*

$$w_g^E \left|err(h, g, \mathcal{D}) - err(h, \mathcal{D})\right| \leq \gamma,$$

where $w_g^E = \Pr_{(x,y)\sim\mathcal{D}}[g(x) = 1]$.

We consider the following fairness-constrained optimization problem:

$$\begin{aligned} &\min_{h\in\Delta\mathcal{H}} && \text{err}(h) \\ &\text{s.t. for each } g \in \mathcal{G}: && w_g^{\text{E}} \left|\text{err}(h, g, \mathcal{D}) - \text{err}(h, \mathcal{D})\right| \leq \gamma, \end{aligned} \tag{11}$$

Definition 17. *Let $f : \mathcal{X} \to R \subseteq [0, 1]$ be some regression function and let $\gamma \in \mathbb{R}_+$. Define $\psi_E(f, \gamma, \mathcal{H})$ to be the following optimization problem:*

$$\begin{aligned} &\min_{h\in\Delta\mathcal{H}} \mathop{\mathbb{E}}_{x\sim\mathcal{D}_\mathcal{X}} [f(x)\ell(h(x), 1) + (1 - f(x))\ell(h(x), 0)] \\ &\textit{s.t. for each } g \in \mathcal{G}: \\ &|\mathbb{E}[\ell(h(x), 1)g(x)f^*(x) + \ell(h(x), 0)g(x)(1 - f^*(x)) \\ &\quad - w_g^E(\ell(h(x), 1)f^*(x) - w_g^E\ell(h(x), 0)(1 - f^*(x))]| \leq \gamma, \end{aligned}$$

where $w_g^E = \Pr_{(x,y)\sim\mathcal{D}}[g(x) = 1]$ as in the previous definition.

Lemma 7. *Let f^* be the Bayes optimal regression function over $\mathcal{D}$. Then optimization problem $\psi_E(f^*, \gamma, \mathcal{H})$ is equivalent to the fairness-constrained optimization problem 11.*

Algorithm 2: Projected Gradient Descent Algorithm for γ-False Negative Fairness

Input: (D: *dataset,* $f : \mathcal{X} \to [0,1]$*: regression function,* $\mathcal{G}$*: groups,* γ*: tolerance on fairness violation,* C*: bound on dual* ($\|\lambda\|_1 \leq C$)*,* η*: learning rate*)

Initialize dual vector $\lambda^0 = \mathbf{0}$ and set $T = \frac{1}{4} \cdot C^2 \cdot (C^2 + 4|\mathcal{G}|)^2$.

for $t = 1, \ldots, T$ **do**

Primal player updates h_t

$$h_t(x) = \begin{cases} 1, & \text{if } f(x) \geq \frac{1}{2+\sum_{g\in\mathcal{G}} \lambda_g^{t-1}(g(x)-\beta_g)} \text{ and } 2+\sum_{g\in\mathcal{G}} \lambda_g^{t-1}(g(x)-\beta_g) > 0, \\ 0, & \text{if } f(x) < \frac{1}{2+\sum_{g\in\mathcal{G}} \lambda_g^{t-1}(g(x)-\beta_g)} \text{ and } 2+\sum_{g\in\mathcal{G}} \lambda_g^{t-1}(g(x)-\beta_g) > 0, \\ 1, & \text{if } f(x) < \frac{1}{2+\sum_{g\in\mathcal{G}} \lambda_g^{t-1}(g(x)-\beta_g)} \text{ and } 2+\sum_{g\in\mathcal{G}} \lambda_g^{t-1}(g(x)-\beta_g) < 0, \\ 0, & \text{if } f(x) \geq \frac{1}{2+\sum_{g\in\mathcal{G}} \lambda_g^{t-1}(g(x)-\beta_g)} \text{ and } 2+\sum_{g\in\mathcal{G}} \lambda_g^{t-1}(g(x)-\beta_g) < 0, \\ 0 & \text{if } 2+\sum_{g\in\mathcal{G}} \lambda_g^{t-1}(g(x)-\beta_g) = 0 \end{cases}$$

Compute

$$\hat{\rho}_g^t = \mathop{\mathbb{E}}_{(x,y)\sim D}[\ell(h_t(x),1)g(x)f(x)] \text{ for all } g \in \mathcal{G},$$
$$\hat{\rho}^t = \mathop{\mathbb{E}}_{(x,y)\sim D}[\beta_g \ell(h_t(x),1)f(x)], \text{ where } \beta_g = \Pr[g(x)=1|y=0]$$

Dual player updates

$$\lambda_g^{t,+} = \max(0, \lambda_g^{t,+} + \eta \cdot (\hat{\rho}_g^t - \hat{\rho}^t - \gamma)),$$
$$\lambda_g^{t,-} = \max(0, \lambda_g^{t,-} + \eta \cdot (\hat{\rho}^t - \hat{\rho}_g^t - \gamma)).$$

Dual player sets $\lambda^t = \sum_{g\in\mathcal{G}} \lambda_g^{t,+} - \lambda_g^{t,-}$.

If $\|\lambda^t\|_1 > C$, set $\lambda^t = \arg\min_{\{\tilde{\lambda}\in\mathbb{R}^{2\mathcal{G}} | \|\tilde{\lambda}\|_1 \leq C\}} \|\lambda_t - \tilde{\lambda}\|_2^2$.

end

Output: $\bar{h} := \frac{1}{T}\sum_{t=1}^T \hat{h}_t$, a uniformly random classifier over all rounds' hypotheses.

Proof. Note that the objective function is equivalent to that of Equation 1, and hence proof of the objectives being equivalent is identical to that of Lemma 1. For the constraints, note that

$$\begin{aligned}
w_g^{\mathrm{E}}|\mathrm{err}(h,g,\mathcal{D})| - \mathrm{err}(h,\mathcal{D}) &= \Pr[g(x)=1]\Big|\Pr[y=1|g(x)=1]\Pr[h(x)=0|g(x)=1,y=1] \\
&\quad + \Pr[y=0|g(x)=1]\Pr[h(x)=1|g(x)=1,y=0] \\
&\quad - (\Pr[y=1]\Pr[h(x)=0|y=1] + \Pr[y=0]\Pr[h(x)=1|y=0])\Big| \\
&= \Pr[g(x)=1]\Big|\Pr[y=1|g(x)=1]\frac{\Pr[h(x)=0,g(x)=1,y=1]}{\Pr[g(x)=1,y=1]} \\
&\quad + \Pr[y=1|g(x)=1]\frac{\Pr[h(x)=1,g(x)=1,y=0]}{\Pr[g(x)=1,y=0]} \\
&\quad - \Pr[y=1]\frac{\Pr[h(x)=0,y=1]}{\Pr[y=1]} - \Pr[y=0]\frac{\Pr[h(x)=1,y=1]}{\Pr[y=0]}\Big| \\
&= |\mathbb{E}[\ell(h(x),1)g(x)f^*(x) + \ell(h(x),0)g(x)(1-f^*(x)) \\
&\quad - w_g^{\mathrm{E}}(\ell(h(x),1)f^*(x) - w_g^{\mathrm{E}}\ell(h(x),0)(1-f^*(x))|
\end{aligned}$$

□

Definition 18 (Lagrangian)**.** *Given any regression function f, we define a Lagrangian of the optimization problem $\psi_E(f,\gamma,\mathcal{H})$ as $L_f^E : \mathcal{H}\times\mathbb{R}^{2|\mathcal{G}|}\to\mathbb{R}$:*

$$\begin{aligned}
L_f^E(h,\lambda) = \mathbb{E}_{x\sim\mathcal{D}_\mathcal{X}}\Big[&f(x)\ell(h(x),1) + (1-f(x))\ell(h(x),0) \\
&+ \sum_{g\in\mathcal{G}}\lambda_g^+\big(\ell(h(x),1)g(x)f(x) + \ell(h(x),0)g(x)(1-f(x)) \\
&\qquad - w_g^E\ell(h(x),1)f(x) - w_g^E\ell(h(x),0)(1-f(x)) - \gamma\big) \\
&+ \sum_{g\in\mathcal{G}}\lambda_g^-\big(w_g^E\ell(h(x),1)f(x) + w_g^E\ell(h(x),0)(1-f(x)) \\
&\qquad - \ell(h(x),1)g(x)f(x) - \ell(h(x),0)g(x)(1-f(x)) - \gamma\big)\Big].
\end{aligned}$$

Lemma 8.

$$\begin{aligned}
L_f^E(h,\lambda) = \mathbb{E}_{x\sim\mathcal{D}_x}\Bigg[&\ell(h(x),0)\Big(1 + \sum_{g\in\mathcal{G}}\lambda_g(g(x)-w_g^E)\Big) - \gamma\sum_{g\in\mathcal{G}}(\lambda_g^+ + \lambda_g^-) \\
&+ f(x)\Bigg(-\ell(h(x),0)\Big[1 + \sum_{g\in\mathcal{G}}\lambda_g(g(x)-w_g^E)\Big] \\
&\qquad + \ell(h(x),1)\Big[1 + \sum_{g\in\mathcal{G}}\lambda_g(g(x)-w_g^E)\Big]\Bigg)\Bigg]
\end{aligned}$$

Proof. Distribute out like terms as shown previously. □

Lemma 9. *The optimal post-processed classifier h of $\psi(f,\gamma,\mathcal{H}_A)$ for some regressor f takes the following form:*

$$h(x) = \begin{cases} 1, & \text{if } f(x) > \frac{1+\sum_{g\in\mathcal{G}} \lambda_g(g(x)-w_g^E)}{2+2\sum_{g\in\mathcal{G}} \lambda_g(g(x)-w_g^E)} \text{ and } 2+2\sum_{g\in\mathcal{G}} \lambda_g(g(x)-w_g^E) > 0, \\ 0, & \text{if } f(x) < \frac{1+\sum_{g\in\mathcal{G}} \lambda_g(g(x)-w_g^E)}{2+2\sum_{g\in\mathcal{G}} \lambda_g(g(x)-w_g^E)} \text{ and } 2+2\sum_{g\in\mathcal{G}} \lambda_g(g(x)-w_g^E) > 0, \\ 1, & \text{if } f(x) < \frac{1+\sum_{g\in\mathcal{G}} \lambda_g(g(x)-w_g^E)}{2+2\sum_{g\in\mathcal{G}} \lambda_g(g(x)-w_g^E)} \text{ and } 2+2\sum_{g\in\mathcal{G}} \lambda_g(g(x)-w_g^E) < 0, \\ 0, & \text{if } f(x) > \frac{1+\sum_{g\in\mathcal{G}} \lambda_g(g(x)-w_g^E)}{2+2\sum_{g\in\mathcal{G}} \lambda_g(g(x)-w_g^E)} \text{ and } 2+2\sum_{g\in\mathcal{G}} \lambda_g(g(x)-w_g^E) > 0. \end{cases}$$

In the edge case in which $f(x) = \frac{1+\sum_{g\in\mathcal{G}} \lambda_g(g(x)-w_g^E)}{2+2\sum_{g\in\mathcal{G}} \lambda_g(g(x)-w_g^E)}$, $h(x)$ *could take either value and might be randomized.*

Proof. Note that since we are optimizing over the set of all binary classifiers, h optimizes the Lagrangian objective pointwise for every x. In particular, we have from Lemma 8 that:

$$\begin{aligned} h(x) = \arg\min_p \Bigg[& \ell(p,0)\Bigg(1+\sum_{g\in\mathcal{G}} \lambda_g(g(x)-w_g^{\mathrm{E}})\Bigg) \\ & + f(x)\Bigg(-\ell(p,0)\Bigg[1+\sum_{g\in\mathcal{G}} \lambda_g(g(x)-w_g^{\mathrm{E}})\Bigg] + \ell(p,1)\Bigg[1+\sum_{g\in\mathcal{G}} \lambda_g(g(x)-w_g^{\mathrm{E}})\Bigg]\Bigg)\Bigg] \end{aligned}$$

Setting $p=0$ makes the inner portion of the expression evaluate to

$$f(x)\Bigg(1+\sum_{g\in\mathcal{G}} \lambda_g(g(x)-w_g^{\mathrm{E}})\Bigg),$$

and setting $p=1$ makes the inner portion of the expression evaluate to

$$\Bigg(1+\sum_{g\in\mathcal{G}} \lambda_g(g(x)-w_g^{\mathrm{E}})\Bigg) - f(x)\Bigg(1+\sum_{g\in\mathcal{G}} \lambda_g(g(x)-w_g^{\mathrm{E}})\Bigg)$$

In order to find the optimal h, we want to find the threshold at which setting $p=1$ minimizes the expression, and hence:

$$\begin{aligned} \Bigg(1+\sum_{g\in\mathcal{G}} \lambda_g(g(x)-w_g^{\mathrm{E}})\Bigg) - f(x)\Bigg(1+\sum_{g\in\mathcal{G}} \lambda_g(g(x)-w_g^{\mathrm{E}})\Bigg) &< f(x)\Bigg(1+\sum_{g\in\mathcal{G}} \lambda_g(g(x)-w_g^{\mathrm{E}})\Bigg) \\ \frac{1+\sum_{g\in\mathcal{G}} \lambda_g(g(x)-w_g^{\mathrm{E}})}{\Bigg(1+\sum_{g\in\mathcal{G}} \lambda_g(g(x)-w_g^{\mathrm{E}})\Bigg)+\Bigg(1+\sum_{g\in\mathcal{G}} \lambda_g(g(x)-w_g^{\mathrm{E}})\Bigg)} &< f(x) \\ \frac{1+\sum_{g\in\mathcal{G}} \lambda_g(g(x)-w_g^{\mathrm{E}})}{2+2\sum_{g\in\mathcal{G}} \lambda_g(g(x)-w_g^{\mathrm{E}})} &< f(x) \end{aligned}$$

Thus,

$$h(x) = \begin{cases} 1, & \text{if } f(x) > \frac{1+\sum_{g\in\mathcal{G}} \lambda_g(g(x)-w_g^{\mathrm{E}})}{2+2\sum_{g\in\mathcal{G}} \lambda_g(g(x)-w_g^{\mathrm{E}})} \text{ and } 2+2\sum_{g\in\mathcal{G}} \lambda_g(g(x)-w_g^{\mathrm{E}}) > 0, \\ 0, & \text{if } f(x) < \frac{1+\sum_{g\in\mathcal{G}} \lambda_g(g(x)-w_g^{\mathrm{E}})}{2+2\sum_{g\in\mathcal{G}} \lambda_g(g(x)-w_g^{\mathrm{E}})} \text{ and } 2+2\sum_{g\in\mathcal{G}} \lambda_g(g(x)-w_g^{\mathrm{E}}) > 0, \\ 1, & \text{if } f(x) < \frac{1+\sum_{g\in\mathcal{G}} \lambda_g(g(x)-w_g^{\mathrm{E}})}{2+2\sum_{g\in\mathcal{G}} \lambda_g(g(x)-w_g^{\mathrm{E}})} \text{ and } 2+2\sum_{g\in\mathcal{G}} \lambda_g(g(x)-w_g^{\mathrm{E}}) < 0, \\ 0, & \text{if } f(x) > \frac{1+\sum_{g\in\mathcal{G}} \lambda_g(g(x)-w_g^{\mathrm{E}})}{2+2\sum_{g\in\mathcal{G}} \lambda_g(g(x)-w_g^{\mathrm{E}})} \text{ and } 2+2\sum_{g\in\mathcal{G}} \lambda_g(g(x)-w_g^{\mathrm{E}}) > 0. \end{cases}$$

□

From Lemma 9, we can now define a best-response model and use Algorithm 3 to generate an optimally post-processed model that preserves $\gamma-$Error fairness. The algorithm's error bounds may be derived using symmetric arguments to sections 3.1 and 3.2, where $\hat{f}$ is α-multicalibrated in expectation with respect to $\mathcal{G}, \mathcal{H}$, and $\mathcal{G} \times \mathcal{H}$ and is jointly multicalibrated with respect to functions of the form:

$$\mathbb{1}\left[\langle \lambda^{t-1}, x_{\mathcal{G}} - w^{\mathrm{E}} \rangle \geq \frac{2v-1}{1-2v}\right]$$

the proofs from section 3.2 may be modified to get its desired error bounds.

Algorithm 3: Projected Gradient Descent Algorithm for γ-Error Fairness

Input: (D: *dataset,* $f : \mathcal{X} \to [0,1]$*: regression function,* $\mathcal{G}$*: groups,* γ*: tolerance on fairness violation,* C*: bound on dual* ($\|\lambda\|_1 \leq C$)*,* η*: learning rate*)

Initialize dual vector $\lambda^0 = \mathbf{0}$ and set $T = \frac{1}{4} \cdot C^2 \cdot (C^2 + 4|\mathcal{G}|)^2$.

for $t = 1, \ldots, T$ **do**

Primal player updates h_t

$$h_t(x) = \begin{cases} 1, & \text{if } f(x) > \frac{1+\sum_{g\in\mathcal{G}} \lambda_g^{t-1}(g(x)-w_g^{\mathrm{E}})}{2+2\sum_{g\in\mathcal{G}} \lambda_g^{t-1}(g(x)-w_g^{\mathrm{E}})} \text{ and } 2+2\sum_{g\in\mathcal{G}} \lambda_g^{t-1}(g(x)-w_g^{\mathrm{E}}) > 0, \\ 0, & \text{if } f(x) < \frac{1+\sum_{g\in\mathcal{G}} \lambda_g^{t-1}(g(x)-w_g^{\mathrm{E}})}{2+2\sum_{g\in\mathcal{G}} \lambda_g^{t-1}(g(x)-w_g^{\mathrm{E}})} \text{ and } 2+2\sum_{g^{t-1}\in\mathcal{G}} \lambda_g(g(x)-w_g^{\mathrm{E}}) > 0, \\ 1, & \text{if } f(x) < \frac{1+\sum_{g\in\mathcal{G}} \lambda_g^{t-1}(g(x)-w_g^{\mathrm{E}})}{2+2\sum_{g\in\mathcal{G}} \lambda_g^{t-1}(g(x)-w_g^{\mathrm{E}})} \text{ and } 2+2\sum_{g\in\mathcal{G}} \lambda_g^{t-1}(g(x)-w_g^{\mathrm{E}}) < 0, \\ 0, & \text{if } f(x) > \frac{1+\sum_{g\in\mathcal{G}} \lambda_g^{t-1}(g(x)-w_g^{\mathrm{E}})}{2+2\sum_{g\in\mathcal{G}} \lambda_g^{t-1}(g(x)-w_g^{\mathrm{E}})} \text{ and } 2+2\sum_{g\in\mathcal{G}} \lambda_g^{t-1}(g(x)-w_g^{\mathrm{E}}) > 0, \\ 1, & \text{if } 2+2\sum_{g\in\mathcal{G}} \lambda_g^{t-1}(g(x)-w_g^{\mathrm{E}}) = 0. \end{cases}$$

Compute

$$\begin{aligned} \hat{\rho}_g^t &= \mathbb{E}_{(x,y)\sim D}[\ell(h_t(x),1)g(x)f(x) + \ell(h_t(x),0)g(x)(1-f(x)) \\ &\quad - w_g^{\mathrm{E}}(\ell(h_t(x),1)f(x) - w_g^{\mathrm{E}}\ell(h_t(x),0)(1-f(x))] \text{ for all } g \in \mathcal{G}, \\ \hat{\rho}^t &= \mathbb{E}_{(x,y)\sim D}[f(x)\ell(h_t(x),1) + (1-f(x))\ell(h_t(x),0)], \end{aligned}$$

where $w_g^{\mathrm{E}} = \Pr_{(x,y)\sim\mathcal{D}}[g(x) = 1]$.

Dual player updates

$$\begin{aligned} \lambda_g^{t,+} &= \max(0, \lambda_g^{t,+} + \eta \cdot (\hat{\rho}_g^t - \hat{\rho}^t - \gamma)), \\ \lambda_g^{t,-} &= \max(0, \lambda_g^{t,-} + \eta \cdot (\hat{\rho}^t - \hat{\rho}_g^t - \gamma)). \end{aligned}$$

Dual player sets $\lambda^t = \sum_{g\in\mathcal{G}} \lambda_g^{t,+} - \lambda_g^{t,-}$.

If $\|\lambda^t\|_1 > C$, set $\lambda^t = \arg\min_{\{\tilde{\lambda}\in\mathbb{R}^{2\mathcal{G}} | \|\tilde{\lambda}\|_1 \leq C\}} \|\lambda_t - \tilde{\lambda}\|_2^2$.

end

Output: $\bar{h} := \frac{1}{T}\sum_{t=1}^{T} \hat{h}_t$, a uniformly random classifier over all rounds' hypotheses.

A.3 Statistical Parity Fairness

Definition 19. *We say that classifier* $h : \mathcal{X} \to \mathcal{Y}$ *satisfies* γ*-Statistical Parity (SP) Fairness with respect to* $\mathcal{D}$ *and* $\mathcal{G}$ *if for all* $g \in \mathcal{G}$*,*

$$\Pr_{(x,y)\sim\mathcal{D}}[g(x) = 1] \left| \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[h(x)|g(x) = 1] - \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[h(x)] \right| \in \gamma.$$

We consider the following fairness-constrained optimization problem:

$$\begin{aligned} &\min_{h\in\Delta\mathcal{H}} && \mathrm{err}(h) \\ &\text{s.t. for each } g\in\mathcal{G}: && \Pr[g(x)=1]\left|\mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[h(x)|g(x)=1]-\mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[h(x)]\right|\leq\gamma, \end{aligned} \tag{12}$$

Definition 20. *Let $f:\mathcal{X}\to R\subseteq[0,1]$ be some regression function and let $\gamma\in\mathbb{R}_+$. Define $\phi_{SP}(f,\gamma,\mathcal{H})$ to be the following optimization problem:*

$$\begin{aligned} &\min_{h\in\Delta\mathcal{H}} && \mathop{\mathbb{E}}_{x\sim\mathcal{D}_\mathcal{X}}[f(x)\ell(h(x),1)+(1-f(x))\ell(h(x),0)] \\ &\textit{s.t. for each } g\in\mathcal{G}: && \left|\mathop{\mathbb{E}}_{x\sim\mathcal{D}_\mathcal{X}}[h(x)g(x)]-w_g^{SP}\mathop{\mathbb{E}}_{x\sim\mathcal{D}_\mathcal{X}}[h(x)]\right|\leq\gamma \end{aligned}$$

where $w_g^{SP}=\Pr[g(x)=1]$.

Lemma 10. *Let f^* be the Bayes optimal regression function over $\mathcal{D}$. Then optimization problem $\psi_{SP}(f^*,\gamma,\mathcal{H})$ is equivalent to the fairness-constrained optimization problem 12.*

Definition 21 (Lagrangian)**.** *Given any regression function f, we define a Lagrangian of the optimization problem $\psi_{SP}(f,\gamma,\mathcal{H})$ as $L_f^{SP}:\mathcal{H}\times\mathbb{R}^{2|\mathcal{G}|}\to\mathbb{R}$:*

$$\begin{aligned} L_f^{SP}(h,\lambda)=\mathop{\mathbb{E}}_{x\sim\mathcal{D}_\mathcal{X}}\Bigg[&f(x)\ell(h(x),1)+(1-f(x))(\ell(h(x),0) \\ &+\sum_{g\in\mathcal{G}}\lambda_g^+\big(h(x)(g(x)-1)-\gamma\big)+\sum_{g\in\mathcal{G}}\lambda_g^-\big(h(x)(1-g(x))-\gamma\big)\Bigg]. \end{aligned}$$

Lemma 11. *The optimal post-processed classifier h of $\psi_{SP}(f,\gamma,\mathcal{H}_A)$ for some regressor f takes the following form:*

$$h(x)=\begin{cases}1, & \text{if } f(x)>1/2+(1/2)\sum_{g\in\mathcal{G}}\lambda_g(g(x)-1),\\ 0, & \text{if } f(x)<1/2+(1/2)\sum_{g\in\mathcal{G}}\lambda_g(g(x)-1).\end{cases}$$

In the edge case in which $f(x)=1/2+(1/2)\sum_{g\in\mathcal{G}}\lambda_g(g(x)-1)$, $h(x)$ could take either value and might be randomized.

Proof. Note that we can rewrite our Lagrangian from Definition 21 as

$$L_f^{\mathrm{SP}}(h,\lambda)=\mathop{\mathbb{E}}_{x\sim\mathcal{D}_\mathcal{X}}\left[f(x)(\ell(h(x),1)-\ell(h(x),0))+\ell(h(x),0)+h(x)\sum_{g\in\mathcal{G}}\lambda_g(g(x)-1)+\gamma\sum_{g\in\mathcal{G}}(\lambda^+-\lambda^-)\right],$$

and hence our optimal h will be optimal pointwise, i.e.

$$h(x)\arg\min_p\left[f(x)(\ell(p,1)+\ell(p,0))-\ell(p,0)+p\sum_{g\in\mathcal{G}}\lambda_g(g(x)-1)\right]$$

We can then find our threshold by comparing this expression when $p=0$ and $p=1$, i.e.

$$-f(x) + 1 + \sum_{g \in \mathcal{G}} \lambda_g (g(x) - 1) < f(x)$$
$$\frac{1 + \sum_{g \in \mathcal{G}} \lambda_g (g(x) - 1)}{2} < f(x).$$

Hence,

$$h(x) = \begin{cases} 1, & \text{if } f(x) > 1/2 + (1/2) \sum_{g \in \mathcal{G}} \lambda_g (g(x) - 1), \\ 0, & \text{if } f(x) < 1/2 + (1/2) \sum_{g \in \mathcal{G}} \lambda_g (g(x) - 1). \end{cases}$$

□

We can now define a best-response model and use Algorithm 4 to generate an optimally post-processed model that preserves γ-Statistical Parity fairness. Assuming that $\hat{f}$ is α-multicalibrated in expectation with respect to $\mathcal{G}, \mathcal{H}$, and $\mathcal{G} \times \mathcal{H}$ and is jointly multicalibrated with respect to functions of the form $\mathbb{1}[\langle \lambda, x_{\mathcal{G}} - \beta \rangle \geq 2v - 1]$, the proofs from section 3.2 may be modified to get its desired error bounds.

Algorithm 4: Projected Gradient Descent Algorithm for γ-Statistical Parity Fairness

Input: (*D: dataset, $f : \mathcal{X} \to [0,1]$: regression function, $\mathcal{G}$: groups, γ: tolerance on fairness violation, C: bound on dual ($\|\lambda\|_1 \leq C$), η: learning rate*)

Initialize dual vector $\lambda^0 = \mathbf{0}$ and set $T = \frac{1}{4} \cdot C^2 \cdot (C^2 + 4|\mathcal{G}|)^2$.

for $t = 1, \ldots, T$ **do**

Primal player updates h_t

$$h_t(x) = \begin{cases} 1, & \text{if } f(x) \geq 1/2 + (1/2) \sum_{g \in \mathcal{G}} \lambda_g^{t-1} (g(x) - 1), \\ 0, & \text{if } f(x) < 1/2 + (1/2) \sum_{g \in \mathcal{G}} \lambda_g^{t-1} (g(x) - 1). \end{cases}$$

Compute

$$\hat{\rho}_g^t = \left| \mathop{\mathbb{E}}_{x \sim \mathcal{D}_{\mathcal{X}}} [h_t(x) g(x)] - w_g^{\mathrm{SP}} \mathop{\mathbb{E}}_{x \sim \mathcal{D}_{\mathcal{X}}} [h_t(x)] \right| \text{ for all } g \in \mathcal{G},$$
$$\hat{\rho}^t = \mathbb{E}_{(x,y) \sim D} [f(x) \ell(h_t(x), 1) + (1 - f(x)) \ell(h_t(x), 0)],$$

where $w_g^{\mathrm{SP}} = \Pr[g(x) = 1]$.

Dual player updates

$$\lambda_g^{t,+} = \max(0, \lambda_g^{t,+} + \eta \cdot (\hat{\rho}_g^t - \hat{\rho}^t - \gamma)),$$
$$\lambda_g^{t,-} = \max(0, \lambda_g^{t,-} + \eta \cdot (\hat{\rho}^t - \hat{\rho}_g^t - \gamma)).$$

Dual player sets $\lambda^t = \sum_{g \in \mathcal{G}} \lambda_g^{t,+} - \lambda_g^{t,-}$.

If $\|\lambda^t\|_1 > C$, set $\lambda^t = \arg\min_{\{\tilde{\lambda} \in \mathbb{R}^{2\mathcal{G}} | \|\tilde{\lambda}\|_1 \leq C\}} \|\lambda_t - \tilde{\lambda}\|_2^2$.

end

Output: $\bar{h} := \frac{1}{T} \sum_{t=1}^{T} \hat{h}_t$, a uniformly random classifier over all rounds' hypotheses.

A.4 Achieving All Fairness Notions

Ideally, we would like our function to be multicalibrated so that we can achieve any fairness notion downstream. Putting everything together from the previous sections, we can do so.

Definition 22 (Set of thresholding functions $\mathcal{B}(C)$)**.** *Let $x_{\mathcal{G}} \in \{0,1\}^{|\mathcal{G}|}$ denote the group membership indicator vector of some point x, and define the following functions:*

$$\begin{aligned} d^{FP}(v) &:= \frac{2v-1}{1-v}, \\ d^{FN}(v) &:= \frac{1-2v}{v}, \\ d^{E}(v) &:= \frac{2v-1}{1-2v}, \\ d^{SP}(v) &:= 2v-1. \end{aligned}$$

Then, for any λ, x, β, let

$$\begin{aligned} s_\lambda^{FP}(x,v) &:= \mathbb{1}[\langle \lambda, x_{\mathcal{G}} - \beta^{FP} \rangle \geq d^{FP,E}(v)], \\ s_\lambda^{FN}(x,v) &:= \mathbb{1}[\langle \lambda, x_{\mathcal{G}} - \beta^{FN} \rangle \geq d^{FN}(v)], \\ s_\lambda^{E}(x,v) &:= \mathbb{1}[\langle \lambda, \alpha x_{\mathcal{G}} - w^{E} \rangle \geq d^{E}(v)], \\ s_\lambda^{SP}(x,v) &:= \mathbb{1}[\langle \lambda, x_{\mathcal{G}} - 1 \rangle \geq d^{SP}(v)], \end{aligned}$$

where

$$\begin{aligned} \beta^{FP} &= \{\Pr_{(x,y)\sim\mathcal{D}}[g(x)=1|y=0]\}_{g\in\mathcal{G}}, \\ \beta^{FN} &= \{\Pr_{(x,y)\sim\mathcal{D}}[g(x)=1|y=1]\}_{g\in\mathcal{G}}, \\ w^{E} &= \{\Pr_{(x,y)\sim\mathcal{D}}[g(x)=1]\}_{g\in\mathcal{G}}. \end{aligned}$$

Define $\mathcal{B}(C) = \{s_\lambda^{FP} | \lambda \in \Lambda(C)\} \cup \{s_\lambda^{FN} | \lambda \in \Lambda(C)\} \cup \{s_\lambda^{E} | \lambda \in \Lambda(C)\} \cup \{s_\lambda^{SP} | \lambda \in \Lambda(C)\}$, where $\Lambda(C) = \left\{\lambda \in \mathbb{R}^{2\mathcal{G}} \middle| \|\lambda\|_1 \leq C\right\}$, as defined in Equation 2.

Then, if f is multicalibrated with respect to $\mathcal{B}(C)$, any of the projected gradient descent algorithms covered above (Algorithms 1 through 3) may be run to achieve the desired fairness notion.

B Expanded Proofs and Section 3 Discussion

Lemma 1. *Let f^* be the Bayes optimal regression function over $\mathcal{D}$. Then optimization problem $\psi(f^*, \gamma, \mathcal{H})$ is equivalent to the fairness-constrained optimization problem* (1)*.*

Proof. We confirm that the objective and constraints are both equivalent. First the objective:

$$\begin{aligned} err(h) &= \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[\ell(h(x),y)] \\ &= \sum_{(x,y)\in\mathcal{X}\times\mathcal{Y}} \Pr(X=x, Y=y)\,\ell(h(x),y) \\ &= \sum_{x\in\mathcal{X}} \Pr(X=x, Y=0)\,\ell(h(x),0) + \Pr(X=x, Y=1)\,\ell(h(x),1) \\ &= \mathop{\mathbb{E}}_{x\in\mathcal{X}}[(1-f^*(x))\ell(h(x),0) + f^*(x)\ell(h(x),1)] \end{aligned}$$

For the constraints, note that

$$\begin{aligned}
w_g|\rho_g(h) - \rho(h)| &= \Pr[g(x) = 1, y = 0]\,|\Pr[h(x) = 1|g(x) = 1, y = 0] - \Pr[h(x) = 1|y = 0]| \\
&= \Pr[g(x) = 1, y = 0]\left|\frac{\Pr[h(x) = 1, g(x) = 1, y = 0]}{\Pr[g(x) = 1, y = 0]} - \frac{\Pr[h(x) = 1, y = 0]}{\Pr[Y = 0]}\right| \\
&= \left|\Pr[h(x) = 1, g(x) = 1, y = 0] - \frac{\Pr[g(x) = 1, y = 0]\Pr[h(x) = 1, y = 0]}{\Pr[Y = 0]}\right| \\
&= \left|\mathbb{E}[\ell(h(x), 0)g(x)(1 - f^*(x))] - \frac{\Pr[g(x) = 1, y = 0]}{\Pr[Y = 0]}\mathbb{E}\left[\ell(h(x), 0)(1 - f^*(x))\right]\right| \\
&= |\mathbb{E}[\ell(h(x), 0)g(x)(1 - f^*(x))] - \Pr[g(x) = 1|Y = 0]\mathbb{E}\left[\ell(h(x), 0)(1 - f^*(x))\right]| \\
&= |\mathbb{E}[\ell(h(x), 0)g(x)(1 - f^*(x))] - \beta_g\mathbb{E}\left[\ell(h(x), 0)(1 - f^*(x))\right]|\,.
\end{aligned}$$

The result follows. □

Lemma 12.

$$\begin{aligned}
L_f(h, \lambda) = \mathop{\mathbb{E}}_{x\sim\mathcal{D}_\mathcal{X}}\Bigg[&\ell(h(x), 0)\Big(1 + \sum_{g\in\mathcal{G}}\lambda_g(g(x) - \beta_g)\Big) - \gamma\sum_{g\in\mathcal{G}}(\lambda_g^+ + \lambda_g^-) \\
&- f(x)\Big(-\ell(h(x), 1) + \ell(h(x), 0)\big(1 + \sum_{g\in\mathcal{G}}\lambda_g(g(x) - \beta_g)\big)\Big)\Bigg].
\end{aligned}$$

Proof. Distributing out like terms in the expression for the Lagrangian in Definition 8 gives us

$$\begin{aligned}
L_f(h, \lambda) = \mathop{\mathbb{E}}_{x\sim\mathcal{D}_\mathcal{X}}\Bigg[&f(x)\ell(h(x), 1) + (1 - f(x))\ell(h(x), 0) \\
&+ \sum_{g\in\mathcal{G}}\lambda_g^+\big(\ell(h(x), 0)g(x)(1 - f(x)) - \beta_g\ell(h(x), 0)(1 - f(x)) - \gamma\big) \\
&+ \lambda_g^-\big(\beta_g\ell(h(x), 0)(1 - f(x)) - \ell(h(x), 0)g(x)(1 - f(x)) - \gamma\big)\Bigg] \\
= \mathop{\mathbb{E}}_{x\sim\mathcal{D}_\mathcal{X}}\Bigg[&\ell(h(x), 0)\Big(1 + \sum_{g\in\mathcal{G}}\lambda_g^+(g(x) - \beta_g) + \lambda_g^-(\beta_g - g(x))\Big) - \gamma\sum_{g\in\mathcal{G}}(\lambda_g^+ + \lambda_g^-) \\
&- f(x)\Big(-\ell(h(x), 1) + \ell(h(x), 0)\big(1 + \sum_{g\in\mathcal{G}}\lambda_g^+(g(x) - \beta_g) + \sum_{g\in\mathcal{G}}\lambda_g^-(\beta_g - g(x))\big)\Big)\Bigg] \\
= \mathop{\mathbb{E}}_{x\sim\mathcal{D}_\mathcal{X}}\Bigg[&\ell(h(x), 0)\Big(1 + \sum_{g\in\mathcal{G}}(\lambda_g^+ - \lambda_g^-)(g(x) - \beta_g)\Big) - \gamma\sum_{g\in\mathcal{G}}(\lambda_g^+ + \lambda_g^-) \\
&- f(x)\Big(-\ell(h(x), 1) + \ell(h(x), 0)\big(1 + \sum_{g\in\mathcal{G}}(\lambda_g^+ - \lambda_g^-)(g(x) - \beta_g)\big)\Big)\Bigg].
\end{aligned}$$

Recall that $\lambda_g = \lambda_g^+ - \lambda_g^-$, so we are done. □

Lemma 2. *The optimal post-processed classifier h of $\psi(f, \gamma, \mathcal{H}_A)$ for some regressor f takes the following*

form:

$$h(x) = \begin{cases} 1, & \text{if } f(x) > \frac{1+\sum_{g\in\mathcal{G}} \lambda_g(g(x)-\beta_g)}{2+\sum_{g\in\mathcal{G}} \lambda_g(g(x)-\beta_g)} \text{ and } 2+\sum_{g\in\mathcal{G}} \lambda_g(g(x)-\beta_g) > 0, \\ 0, & \text{if } f(x) < \frac{1+\sum_{g\in\mathcal{G}} \lambda_g(g(x)-\beta_g)}{2+\sum_{g\in\mathcal{G}} \lambda_g(g(x)-\beta_g)} \text{ and } 2+\sum_{g\in\mathcal{G}} \lambda_g(g(x)-\beta_g) > 0, \\ 1, & \text{if } f(x) < \frac{1+\sum_{g\in\mathcal{G}} \lambda_g(g(x)-\beta_g)}{2+\sum_{g\in\mathcal{G}} \lambda_g(g(x)-\beta_g)} \text{ and } 2+\sum_{g\in\mathcal{G}} \lambda_g(g(x)-\beta_g) < 0, \\ 0, & \text{if } f(x) > \frac{1+\sum_{g\in\mathcal{G}} \lambda_g(g(x)-\beta_g)}{2+\sum_{g\in\mathcal{G}} \lambda_g(g(x)-\beta_g)} \text{ and } 2+\sum_{g\in\mathcal{G}} \lambda_g(g(x)-\beta_g) < 0. \end{cases}$$

In the edge case in which $f(x) = \frac{1+\sum_{g\in\mathcal{G}} \lambda_g(g(x)-\beta_g)}{2+\sum_{g\in\mathcal{G}} \lambda_g(g(x)-\beta_g)}$, $h(x)$ *could take either value and might be randomized.*

Proof. Note that since we are optimizing over the set of all binary classifiers, h optimizes the Lagrangian objective pointwise for every x. In particular, we have from Lemma 12 that:

$$h(x) = \arg\min_p \left[\ell(p,0)\Big(1+\sum_{g\in\mathcal{G}} \lambda_g(g(x)-\beta_g)\Big) - f(x)\Big(-\ell(p,1)+\ell(p,0)\big(1+\sum_{g\in\mathcal{G}} \lambda_g(g(x)-\beta_g)\big)\Big)\right].$$

Determining the optimal threshold is equivalent to determining when the above expression with $\ell(p,0)=1$ and $\ell(p,1)=0$ is less than $f(x)$, i.e.

$$1+\sum_{g\in\mathcal{G}} \lambda_g(g(x)-\beta_g) - f(x)\big(1+\sum_{g\in\mathcal{G}} \lambda_g(g(x)-\beta_g)\big) < f(x)$$

$$1+\sum_{g\in\mathcal{G}} \lambda_g(g(x)-\beta_g) < f(x)\left(1+\big(1+\sum_{g\in\mathcal{G}} \lambda_g(g(x)-\beta_g)\big)\right).$$

Thus,

$$h(x) = \begin{cases} 1, & \text{if } f(x) > \frac{1+\sum_{g\in\mathcal{G}} \lambda_g(g(x)-\beta_g)}{2+\sum_{g\in\mathcal{G}} \lambda_g(g(x)-\beta_g)} \text{ and } (2+\sum_{g\in\mathcal{G}} \lambda_g(g(x)-\beta_g)) > 0, \\ 0, & \text{if } f(x) < \frac{1,+\sum_{g\in\mathcal{G}} \lambda_g(g(x)-\beta_g)}{2+\sum_{g\in\mathcal{G}} \lambda_g(g(x)-\beta_g)} \text{ and } (2+\sum_{g\in\mathcal{G}} \lambda_g(g(x)-\beta_g)) > 0 \\ 1, & \text{if } f(x) < \frac{1+\sum_{g\in\mathcal{G}} \lambda_g(g(x)-\beta_g)}{2+\sum_{g\in\mathcal{G}} \lambda_g(g(x)-\beta_g)} \text{ and } (2+\sum_{g\in\mathcal{G}} \lambda_g(g(x)-\beta_g)) < 0, \\ 0, & \text{if } f(x) > \frac{1,+\sum_{g\in\mathcal{G}} \lambda_g(g(x)-\beta_g)}{2+\sum_{g\in\mathcal{G}} \lambda_g(g(x)-\beta_g)} \text{ and } (2+\sum_{g\in\mathcal{G}} \lambda_g(g(x)-\beta_g)) < 0 \end{cases}$$

□

In Lemma 2, we can only describe the optimal post-processed classifier for cases where either $f(x)$ is less than or greater than the threshold $\frac{1+\sum_{g\in\mathcal{G}} \lambda_g(g(x)-\beta_g)}{2+\sum_{g\in\mathcal{G}} \lambda_g(g(x)-\beta_g)}$, $h(x)$. In practice, our algorithm will need to update h at round t according to the current dual variables λ in a way that is well-defined for all values of $f(x)$. Hence, we define our best response as follows, where ties between $f(x)$ and the threshold are broken by rounding to 1.

Definition 23 (Best Response Model)**.** *Given regressor* f *and dual variables* λ*, let the best response* h *be defined as*

$$h(x) = \begin{cases} 1, & \text{if } f(x) \geq \frac{1+\sum_{g\in\mathcal{G}} \lambda_g(g(x)-\beta_g)}{2+\sum_{g\in\mathcal{G}} \lambda_g(g(x)-\beta_g)} \text{ and } (2+\sum_{g\in\mathcal{G}} \lambda_g(g(x)-\beta_g)) > 0, \\ 0, & \text{if } f(x) < \frac{1+\sum_{g\in\mathcal{G}} \lambda_g(g(x)-\beta_g)}{2+\sum_{g\in\mathcal{G}} \lambda_g(g(x)-\beta_g)} \text{ and } (2+\sum_{g\in\mathcal{G}} \lambda_g(g(x)-\beta_g)) > 0, \\ 1, & \text{if } f(x) \leq \frac{1+\sum_{g\in\mathcal{G}} \lambda_g(g(x)-\beta_g)}{2+\sum_{g\in\mathcal{G}} \lambda_g(g(x)-\beta_g)} \text{ and } (2+\sum_{g\in\mathcal{G}} \lambda_g(g(x)-\beta_g)) < 0, \\ 0, & \text{if } f(x) > \frac{1+\sum_{g\in\mathcal{G}} \lambda_g(g(x)-\beta_g)}{2+\sum_{g\in\mathcal{G}} \lambda_g(g(x)-\beta_g)} \text{ and } (2+\sum_{g\in\mathcal{G}} \lambda_g(g(x)-\beta_g)) < 0. \end{cases}$$

Lemma 13. *For any regression model* f *and dual variables* λ*, The classifier* h *defined in Definition 23 is a "best response" in the sense that:*

$$h \in \arg\min_{h\in\mathcal{H}_A} L_f(h,\lambda).$$

B.1 Proofs from Section 3.1

Theorem 1. *Let* OPT *be the objective value of the optimal solution to $\psi(f, \gamma, \mathcal{H}_A)$. Then, for any $C \in \mathbb{R}$, after $T = \frac{1}{4} \cdot C^2 \cdot \left(C^2 + 4|\mathcal{G}|\right)^2$ iterations, Algorithm 1 outputs a randomized hypothesis $\bar{h}$ such that $err(\bar{h}) \leq \text{OPT} + \frac{2}{C}$ and $w_g|\rho_g(\bar{h}) - \rho(\bar{h})| \leq \gamma + \frac{1}{C} + \frac{2}{C^2}$.*

Theorem 3. *Algorithm 1 returns an $\epsilon-$approximate equilibrium solution to the zero-sum game defined by Equation 2 after $T = \frac{1}{4\epsilon^2}\left(\frac{1}{\epsilon^2} + 4|\mathcal{G}|\right)^2$ rounds.*

To prove this, we will use the following result from Freund and Shapire.

Theorem 5 (Freund and Schapire [1996])**.** *(Approximately solving a game). If $\lambda_1, \ldots, \lambda_T \in \Delta_\lambda$ is the sequence of distributions over λ played by the dual player and $h_1, \ldots, h_T \in \mathcal{H}$ is the sequence of best-response hypotheses played by the primal player satisfying regret guarantees*

$$\frac{1}{T}\max_{\lambda \in \Lambda} \sum_{t=1}^{T} U(h_t, \lambda) - \frac{1}{T}\sum_{t=1}^{T} \mathop{\mathbb{E}}_{\lambda \sim \lambda_t}[U(h_t, \lambda)] \leq \Delta_1$$

and

$$\frac{1}{T}\sum_{t=1}^{T} \mathop{\mathbb{E}}_{\lambda \sim \lambda_t}[U(h_t, \lambda)] - \frac{1}{T}\min_{h \in \mathcal{H}} \sum_{t=1}^{T} \mathop{\mathbb{E}}_{\lambda \sim \lambda_t}[U(h, \lambda)] \leq \Delta_2$$

then the time-average of the two players' empirical distributions is a $(\Delta_1 + \Delta_2)-$approximate equilibrium.

Proof of Theorem 3. We follow the regret analysis of Zinkevich [2003]. To instantiate their result, we need a bound on the norm of the gradients of the loss function and on the diameter of the feasible set F. First, we see that at each step the gradient of the loss seen by gradient descent is bounded:

$$\|\nabla \ell\|^2 = \sum_{g \in \mathcal{G}} w_g \left(\rho_g - \rho - \gamma\right)^2 + w_g \left(-\rho_g + \rho - \gamma\right)^2 \leq 2|\mathcal{G}|.$$

Second, we see that if we consider the feasible set such that $\|\lambda\| \leq \frac{1}{\epsilon}$, then $\|F\|^2 = \frac{1}{\epsilon^2}$. Thus we have that the regret of the dual player is bounded:

$$\mathcal{R}(T) \leq \frac{\|F\|^2\sqrt{T}}{2} + (\sqrt{T} - \frac{1}{2})\|\nabla \ell\|^2$$
$$\frac{\mathcal{R}(T)}{T} \leq \frac{1}{T}\left(\frac{\frac{1}{\epsilon^2}\sqrt{T}}{2} + (\sqrt{T} - \frac{1}{2})2|\mathcal{G}|\right) \leq \frac{\frac{1}{\epsilon^2} + 4|\mathcal{G}|}{2\sqrt{T}}.$$

After $T = \frac{1}{4\epsilon^2}\left(\frac{1}{\epsilon^2} + 4|\mathcal{G}|\right)^2$ rounds, by Freund and Schapire [1996] the average over empirical distributions of play of the dual and primal players, $\bar{\lambda}$ and $\bar{h}$, respectively, form an $\epsilon-$approximate equilibrium solution to the zero-sum game defined by 2. □

Proof of Theorem 1. Applying Theorems 2 and 3, we have that after T rounds $(\bar{h}, \bar{\lambda})$ is an ϵ-approximate equilibrium to the zero-sum game of 2 and equivalently a minimax solution to the Λ-bounded Lagrangian. Taking $\epsilon = 1/C$, the solution $(\bar{h}, \bar{\lambda})$ is a $\frac{1+2\epsilon}{1/\epsilon} = 1/C + 2/C^2$ approximate solution to the original linear program 1. □

B.2 Proofs from Section 3.2

Lemma 3. *Let h_t be the response to λ^{t-1} described in Algorithm 1 at some round $t \in [T]$. Then,*

$$h_t(x) = s_{\lambda^{t-1}}(x, f(x)).$$

Proof. Recall from Lemma 13 and Algorithm 1 that the best response to λ that the primal player can make is to compute h based on the thresholding of the expression

$$\tau = \frac{1 + \sum_{g \in \mathcal{G}} \lambda_g^{t-1}(g(x) - \beta_g)}{2 + \sum_{g \in \mathcal{G}} \lambda_g^{t-1}(g(x) - \beta_g)}.$$

Setting this threshold to be greater than or equal to some value v, note the following is implied:

$$\begin{aligned}
&\frac{1 + \sum_{g \in \mathcal{G}} \lambda_g^{t-1}(g(x) - \beta_g)}{2 + \sum_{g \in \mathcal{G}} \lambda_g^{t-1}(g(x) - \beta_g)} \geq v, \\
\Rightarrow \sum_{g \in \mathcal{G}} \lambda_g^{t-1}(g(x) - \beta_g) - v \sum_{g \in \mathcal{G}} \lambda_g^{t-1}(g(x) - \beta_g) &\geq 2v - 1, \\
\Rightarrow (1 - v)(\sum_{g \in \mathcal{G}} \lambda_g^{t-1}(g(x) - \beta_g) &\geq 2v - 1, \\
\Rightarrow \langle \lambda^{t-1}, x_{\mathcal{G}} - \beta \rangle = \sum_{g \in \mathcal{G}} \lambda_g^{t-1}(g(x) - \beta_g) &\geq \frac{2v - 1}{1 - v}.
\end{aligned}$$

Thus, taking the indicator of

$$\mathbb{1}[\langle \lambda^{t-1}, x_{\mathcal{G}} - \beta \rangle \geq d(v)]$$

is equivalent to determining if the threshold τ is greater than or equal to some v, and hence by the definition of $s_{\lambda^{t-1}}(x, v)$ in Definition 11 and of the best response h in Definition 23, if v is set to $f(x)$ it follows that

$$h(x) = s_{\lambda^{t-1}}(x, f(x)).$$

□

Theorem 4. *Set $C = \sqrt{1/\alpha}$. Let $\hat{f}$ be α-approximately multicalibrated in expectation with respect to $\mathcal{G}$, $\mathcal{H}$, and $\mathcal{G} \times \mathcal{H}$ and α-approximately jointly multicalibrated in expectation with respect to $\mathcal{B}(C)$. Let $\bar{h}$ be the result of running Algorithm 1 with input $\hat{f}$ and C. Then, $err(\bar{h}) \leq err(h^*) + \alpha(5 + 2\sqrt{1/\alpha}) + 2\sqrt{\alpha}$, and for all $g \in \mathcal{G}, w_g \left|\rho_g(\bar{h}) - \rho(\bar{h})\right| \leq w_g \left|\rho_g(h^*) - \rho(h^*)\right| + w_g\alpha$.*

In order to prove this, we will proceed through the specifics of each line of the proof sketch in the section 3.2 through Lemmas 14 through 20.

Lemma 14 (Equality in Equation 3)**.**

$$err(h^*) = L^*(h^*, \lambda^*)$$

Proof. Consider the optimal solution (h^*, λ^*) to $\psi(f^*, \gamma, \mathcal{H})$, and recall that $\text{err}(h) = \mathbb{E}_{x \sim \mathcal{D}_{\mathcal{X}}}[f^*(x)\ell(h(x), 1) + (1 - f^*(x))\ell(h(x), 0)]$. Since the solution is optimal, it follows from complementary slackness, for each group g one of the following must hold: Either the constraint is exactly tight and so its "violation" term in the Lagrangian evaluates to 0, or its corresponding dual variables $\lambda_g^{\pm} = 0$. Thus, $L_f^*(h^*, \lambda^*)$ simplifies to

$$\begin{aligned}
L_f^*(h^*, \lambda^*) = &\mathop{\mathbb{E}}_{x \sim \mathcal{D}_{\mathcal{X}}} \Big[f(x)\ell(h(x), 1) + (1 - f(x))\ell(h(x), 0) \\
&+ 0 \cdot \sum_{g \in \mathcal{G}} \lambda_g^+ \big(\ell(h(x), 0)g(x)(1 - f(x)) - \beta_g \ell(h(x), 0)(1 - f(x)) - \gamma\big) \\
&+ 0 \cdot \sum_{g \in \mathcal{G}} \lambda_g^- \big(\beta_g \ell(h(x), 0)(1 - f(x)) - \ell(h(x), 0)g(x)(1 - f(x)) - \gamma\big) \Big] \\
= &\mathop{\mathbb{E}}_{x \sim \mathcal{D}_{\mathcal{X}}} \Big[f(x)\ell(h(x), 1) + (1 - f(x))\ell(h(x), 0) \Big] \\
= &\,\text{err}(h^*)
\end{aligned}$$

□

Lemma 15 (Bounding Equation 3 by Equation 4)**.**

$$L^*(h^*, \lambda^*) \geq L^*(h^*, \hat{\lambda}).$$

Proof. This follows from the dual optimality condition that $\lambda^* \in \arg\max_\lambda L^*(h^*, \lambda)$. □

Lemma 16 (Bounding Equation 4 by Equation 5)**.** *Fix any λ. If $\hat{f}$ is α-multicalibrated with respect to $\mathcal{G}, \mathcal{H}$, and $\mathcal{G} \times \mathcal{H} = \{g(x) \cdot h(x) | g \in \mathcal{G}, h \in \mathcal{H}\}$, then then we have*

$$\left|\hat{L}(h^*, \lambda) - L^*(h^*, \lambda)\right| \leq \alpha(3 + 2\|\lambda\|_1).$$

Proof. Observe that we can write:

$$\hat{L}(h, \lambda) = L_1(h, \lambda) - \gamma \sum_{g \in \mathcal{G}} (\lambda_g^+ + \lambda_g^-) - \hat{L}_2(h, \lambda),$$

where

$$L_1(h, \lambda) = \mathop{\mathbb{E}}_{x \sim \mathcal{D}_\mathcal{X}} \left[\ell(h(x), 0) \Big(1 + \sum_{g \in \mathcal{G}} \lambda_g (g(x) - \beta_g) \Big) \right],$$

$$\hat{L}_2(h, \lambda) = \mathop{\mathbb{E}}_{x \sim \mathcal{D}_\mathcal{X}} \left[\hat{f}(x) \Big(-\ell(h(x), 1) + \ell(h(x), 0) \big(1 + \sum_{g \in \mathcal{G}} \lambda_g (g(x) - \beta_g) \big) \Big) \right].$$

Similarly, we can write:

$$L^*(h, \lambda) = L_1(h, \lambda) - \gamma \sum_{g \in \mathcal{G}} (\lambda_g^+ + \lambda_g^-) - L_2^*(h, \lambda),$$

where

$$L_2^*(h, \lambda) = \mathop{\mathbb{E}}_{x \sim \mathcal{D}_\mathcal{X}} \left[f^*(x) \Big(-\ell(h(x), 1) + \ell(h(x), 0) \big(1 + \sum_{g \in \mathcal{G}} \lambda_g (g(x) - \beta_g) \big) \Big) \right].$$

Observe that the L_1 term does not depend on $\hat{f}$ or f^* and so is common between $\hat{L}$ and L^*. We can bound $\hat{L}_2$ as follows:

$$\begin{aligned}
\hat{L}_2(h^*, \lambda) &= \mathop{\mathbb{E}}_{x \sim \mathcal{D}_\mathcal{X}} \left[\hat{f}(x) \Big(-\ell(h^*(x), 1) + \ell(h^*(x), 0) \big(1 + \sum_{g \in \mathcal{G}} \lambda_g (g(x) - \beta_g) \big) \Big) \right] \\
&= \mathop{\mathbb{E}}_{x \sim \mathcal{D}_\mathcal{X}} \left[\hat{f}(x) \Big(-(1 - h^*(x)) + h^*(x) \big(1 + \sum_{g \in \mathcal{G}} \lambda_g (g(x) - \beta_g) \big) \Big) \right] \\
&= \sum_{v \in R} \Pr[\hat{f}(x) = v] \mathop{\mathbb{E}}_{x \sim \mathcal{D}_x} \left[\hat{f}(x) \Big(-(1 - h^*(x)) + h^*(x) \big(1 + \sum_{g \in \mathcal{G}} \lambda_g (g(x) - \beta_g) \big) \Big) \middle| \hat{f}(x) = v \right] \\
&\leq \sum_{v \in R} \Pr[\hat{f}(x) = v] \mathop{\mathbb{E}}_{x \sim \mathcal{D}_x} \left[f^*(x) \Big(-(1 - h^*(x)) + h^*(x) \big(1 + \sum_{g \in \mathcal{G}} \lambda_g (g(x) - \beta_g) \big) \Big) \middle| \hat{f}(x) = v \right] \\
&\quad + \alpha \left(3 + \sum_{g \in \mathcal{G}} \lambda_g (1 + \beta_g) \right) \\
&\leq L_2^*(h^*, \lambda) + \alpha \left(3 + 2\|\lambda\|_1 \right),
\end{aligned}$$

where the first inequality follows from the fact that $h^* \in \mathcal{H}$ and $\hat{f}$ is multicalibrated with respect to $\mathcal{G}, \mathcal{H}$, and $\mathcal{G} \times \mathcal{H}$, which we verify below:

$$\begin{aligned}
&\sum_{v \in R} \Pr[\hat{f}(x) = v] \mathop{\mathbb{E}}_{x \sim \mathcal{D}_x} \left[\left(f^*(x) - \hat{f}(x) \right) \cdot \Big(-(1 - h^*(x)) + h^*(x)\big(1 + \sum_{g \in \mathcal{G}} \lambda_g (g(x) - \beta_g)\big) \Big) \middle| \hat{f}(x) = v \right] \\
&= \sum_{v \in R} \Pr[\hat{f}(x) = v] \left[\left(f^*(x) - \hat{f}(x) \right) \cdot \Big(-1 + 2h^*(x) + h^*(x) \sum_{g \in \mathcal{G}} \lambda_g (g(x) - \beta_g) \Big) \middle| \hat{f}(x) = v \right] \\
&= -\sum_{v \in R} \Pr[\hat{f}(x) = v] \mathop{\mathbb{E}}_{x \sim \mathcal{D}_x} \left[\hat{f}^*(x) - \hat{f}(x) \middle| \hat{f}(x) = v \right] \\
&\quad + 2 \sum_{v \in R} \Pr[\hat{f}(x) = v] \mathop{\mathbb{E}}_{x \sim \mathcal{D}_x} \left[(f^*(x) - \hat{f}(x)) h^*(x) \middle| \hat{f}(x) = v \right] \\
&\quad + \sum_{v \in R} \Pr[\hat{f}(x) = v] \sum_{g \in \mathcal{G}} \lambda_g \mathop{\mathbb{E}}_{x \sim \mathcal{D}_x} \left[(f^*(x) - \hat{f}(x)) h^*(x) g(x) \middle| \hat{f}(x) = v \right] \\
&\quad - \sum_{v \in R} \Pr[\hat{f}(x) = v] \sum_{g \in \mathcal{G}} \lambda_g \beta_g \mathop{\mathbb{E}}_{x \sim \mathcal{D}_x} \left[(f^*(x) - \hat{f}(x)) h^*(x) \middle| \hat{f}(x) = v \right] \\
&\leq 3\alpha + \sum_{g \in \mathcal{G}} \lambda_g (1 + \beta_g) \alpha \\
&\leq 3\alpha + \alpha \sum_{g \in \mathcal{G}} \lambda_g (1 + \max_{g' \in \mathcal{G}} \beta_{g'}) \\
&\leq 3\alpha + \alpha \sum_{g \in \mathcal{G}} \lambda_g (1 + 1) \\
&\leq 3\alpha + 2\|\lambda\|_1 \alpha
\end{aligned}$$

Similarly, we can show that $L^*(h^*, \lambda) - \hat{L}(h^*, \lambda) \leq \alpha\,(3 + 2\|\lambda\|_1)$. Putting everything together, we get that:

$$\left| \hat{L}(h^*, \lambda) - L^*(h^*, \lambda) \right| \leq \alpha(3 + 2\|\lambda\|_1).$$

This concludes the proof. □

Lemma 17 (Bounding Equation 5 by Equation 6)**.**

$$\hat{L}(h^*, \hat{\lambda}) \geq \hat{L}(\hat{h}, \hat{\lambda})$$

Proof. This follows from the primal optimality condition that $\hat{h} \in \arg\min_{h \in \mathcal{H}_A} \hat{L}(h, \hat{\lambda})$ and that $\mathcal{H} \subseteq \mathcal{H}_A$. □

Lemma 18 (Equality of Equation 6 and Equation 7)**.**

$$\hat{L}(\hat{h}, \hat{\lambda}) = \widehat{err}(\hat{h})$$

Proof. This follows the same complimentary slackness argument as the proof of Lemma 14. □

Lemma 19 (Bound of Equation 7 by Equation 8)**.** *Consider $\bar{h}$ output by algorithm 1 after $T = \frac{1}{4} \cdot C^2 \cdot \left(C^2 + 4|\mathcal{G}|\right)^2$ rounds. Then,*

$$\widehat{err}(\hat{h}) + 2/C \geq \widehat{err}(\bar{h})$$

Proof. This follows directly from Theorem 1. □

Lemma 20 (Bound of Equation 8 by Equation 9)**.** *Let $\hat{f}$ be α-approximately jointly multicalibrated with respect to $\mathcal{B}(C)$. Then,*

$$\left| \widehat{err}(\bar{h}) - err(\bar{h}) \right| \leq 2\alpha.$$

Proof. Since $\bar{h}$ is a randomized model that mixes uniformly over model $\hat{h}_t$ for $t \in [T]$, it suffices to show that for every $t \in [T]$,

$$\left|\widehat{\mathrm{err}}(\hat{h}_t) - \mathrm{err}(\hat{h}_t)\right| \leq 2\alpha.$$

We can compute:

$$\begin{aligned}
\widehat{\mathrm{err}}(\hat{h}_t) &= \mathop{\mathbb{E}}_{x\sim\mathcal{D}_\mathcal{X}}\left[\hat{f}(x)\ell(\hat{h}_t, 1) + (1-\hat{f}(x))\ell(\hat{h}_t(x), 0)\right], \\
&= \sum_{v\in R} \Pr[\hat{f}(x) = v, s_{\lambda^{t-1}}(x, v) = 0] \mathop{\mathbb{E}}_{x\sim\mathcal{D}_\mathcal{X}}[\hat{f}(x)\ell(\hat{h}_t(x), 1) + (1-\hat{f}(x))\ell(\hat{h}_t(x), 0) | \hat{f}(x) = v, s_{\lambda^{t-1}}(x, v) = 0], \\
&+ \sum_{v\in R} \Pr[\hat{f}(x) = v, s_{\lambda^{t-1}}(x, v) = 1] \mathop{\mathbb{E}}_{x\sim\mathcal{D}_\mathcal{X}}[\hat{f}(x)\ell(\hat{h}_t(x), 1) + (1-\hat{f}(x))\ell(\hat{h}_t(x), 0) | \hat{f}(x) = v, s_{\lambda^{t-1}}(x, v) = 1].
\end{aligned}$$

By Lemma 3, $\hat{h}_t(x) = s_{\lambda^{t-1}}(x, \hat{f}(x))$, and so in particular conditioning on $\hat{f}(x) = v$ and $s_{\lambda^{t-1}}(x, v)$ fixes the value of $\hat{h}_t(x)$. So, we can rewrite the above as

$$\begin{aligned}
\widehat{\mathrm{err}}(\hat{h}_t) &= \sum_{v\in R} \Pr[\hat{f}(x) = v, s_{\lambda^{t-1}}(x, v) = 0] \mathop{\mathbb{E}}_{x\sim\mathcal{D}_\mathcal{X}}[\hat{f}(x) | \hat{f}(x) = v, s_{\lambda^{t-1}}(x, v) = 0] \\
&+ \sum_{v\in R} \Pr[\hat{f}(x) = v, s_{\lambda^{t-1}}(x, v) = 1] \mathop{\mathbb{E}}_{x\sim\mathcal{D}_\mathcal{X}}[1-\hat{f}(x) | \hat{f}(x) = v, s_{\lambda^{t-1}}(x, v) = 1] \\
&\leq \sum_{v\in R} \Pr[\hat{f}(x) = v, s_{\lambda^{t-1}}(x, v) = 0] \mathop{\mathbb{E}}_{x\sim\mathcal{D}_\mathcal{X}}[f^*(x) | \hat{f}(x) = v, s_{\lambda^{t-1}}(x, v) = 0] + \alpha \\
&+ \sum_{v\in R} \Pr[\hat{f}(x) = v, s_{\lambda^{t-1}}(x, v) = 1] \mathop{\mathbb{E}}_{x\sim\mathcal{D}_\mathcal{X}}[1-f^*(x) | \hat{f}(x) = v, s_{\lambda^{t-1}}(x, v) = 1] + \alpha \\
&= \mathop{\mathbb{E}}_{x\sim D_\mathcal{X}}[f^*(x)\ell(\hat{h}_t(x), 1) + (1-f^*(x))\ell(\hat{h}_t(x), 0)] + 2\alpha \\
&= \mathrm{err}(\hat{h}_t) + 2\alpha,
\end{aligned}$$

where the inequality comes from our α-approximate joint multicalibration guarantee. The same argument yields the opposite direction, so we are done. □

We now have the tools to prove our main theorem.

Proof of Theorem 4. Applying lemmas 14 through 20 gives us

$$\begin{aligned}
\mathrm{err}(h^*) &= L^*(h^*, \lambda^*) \quad \text{(Lemma 14)} && (13)\\
&\geq L^*(h^*, \hat{\lambda}) \quad \text{(Lemma 15)} && (14)\\
&\geq \hat{L}(h^*, \hat{\lambda}) - \alpha(3 + 2\|\lambda\|_1) \quad \text{(Lemma 16)} && (15)\\
&\geq \hat{L}(\hat{h}, \hat{\lambda}) - \alpha(3 + 2\|\lambda\|_1) \quad \text{(Lemma 17)}, && (16)
\end{aligned}$$

and

$$\begin{aligned}
\hat{L}(\hat{h}, \hat{\lambda}) &= \widehat{\mathrm{err}}(\hat{h}) \quad \text{(Lemma 18)} && (17)\\
&\geq \widehat{\mathrm{err}}(\bar{h}) - 2/C \quad \text{(Lemma 19)} && (18)\\
&\geq \mathrm{err}(\bar{h}) - 2/C - 2\alpha \quad \text{(Lemma 20)}. && (19)
\end{aligned}$$

Putting this all together gives us

$$\begin{aligned}\mathrm{err}(h^*) &\geq \mathrm{err}(\bar{h}) - \alpha(3 + 2\|\lambda\|_1) - 2/C - 2\alpha \\ &= \mathrm{err}(\bar{h}) - \alpha(5 + 2\|\lambda\|_1) - 2/C \\ &\geq \mathrm{err}(\bar{h}) - \alpha(5 + 2C) - 2/C\end{aligned}$$

We want to set C to minimize this discrepancy. Noting that the derivative of $\alpha(5 + 2C) + 2/C$ with respect to C is $2\alpha - 2/C^2$, we get a minimization at $C = \sqrt{1/\alpha}$.

Setting C as such gives the desired bound:

$$\mathrm{err}(h^*) \geq \mathrm{err}(\bar{h}) - \alpha(5 + 2\sqrt{1/\alpha}) - 2\sqrt{\alpha}.$$

Following a similar analysis as Lemma 20, we can bound the fairness constraints on $\bar{h}$ by bounding them for the model $\hat{h}_t$ found at every round $t \in [T]$ of algorithm 1.

$$\begin{aligned}&\hat{\rho}_g(\hat{h}_t) - \hat{\rho}(\hat{h}_t) = \mathop{\mathbb{E}}_{x\sim\mathcal{D}_x}[(1 - \hat{f}(x))\ell(\hat{h}_t(x), 0)g(x)] - \mathop{\mathbb{E}}_{x\sim\mathcal{D}_x}[(1 - \hat{f}(x))\ell(\hat{h}_t(x), 0)] \\ &= \sum_{v\in R} \Pr[\hat{f}(x) = v, s_{\lambda^{t-1}}(x, v) = 0] \mathop{\mathbb{E}}_{x\sim\mathcal{D}_x}[(1 - \hat{f}(x))\ell(\hat{h}_t(x), 0) \cdot (g(x) - 1)|\hat{f}(x) = v, s_{\lambda^{t-1}}(x, v) = 0] \\ &\quad + \Pr[\hat{f}(x) = v, s_{\lambda^{t-1}}(x, v) = 1] \mathop{\mathbb{E}}_{x\sim\mathcal{D}_x}[(1 - \hat{f}(x))\ell(\hat{h}_t(x), 0) \cdot (g(x) - 1)|\hat{f}(x) = v, s_{\lambda^{t-1}}(x, v) = 1] \\ &= \sum_{v\in R} \Pr[\hat{f}(x) = v, s_{\lambda^{t-1},\geq}(x, v) = 1] \mathop{\mathbb{E}}_{x\sim\mathcal{D}_x}[(1 - \hat{f}(x))\ell(\hat{h}_t(x), 0) \cdot (g(x) - 1)|\hat{f}(x) = v, s_{\lambda^{t-1}}(x, v) = 1] \\ &\leq \sum_{v\in R} \Pr[\hat{f}(x) = v, s_{\lambda^{t-1}}(x, v) = 1] \mathop{\mathbb{E}}_{x\sim\mathcal{D}_x}[(1 - f^*(x))\ell(\hat{h}_t(x), 0) \cdot (g(x) - 1)|\hat{f}(x) = v, s_{\lambda^{t-1}}(x, v) = 1] + \alpha \\ &= \mathop{\mathbb{E}}_{x\in\mathcal{D}_x}[(1 - f^*(x))\ell(h_t(x), 0) \cdot (g(x) - 1)] + \alpha \\ &= \rho_g(h_t) - \rho(h_t) + \alpha.\end{aligned}$$

Here, the inequality comes from our multicalibration guarantees. We can repeat the same argument in the opposite direction, and get that

$$w_g \left|\rho_g(h^*) - \rho(h^*)\right| \geq w_g \left|\rho_g(\bar{h}) - \rho(\bar{h})\right| - w_g\alpha.$$

□

C Achieving Joint Multicalibration

In this section we give an algorithm that can take as input any model $f : \mathcal{X} \to [0, 1]$ and transform it into a new model $\hat{f} : \mathcal{X} \to R$ such that $\hat{f}$ achieves multicalibration in expectation with respect to a class of functions $\mathcal{C}_1 \subset \{0, 1\}^{\mathcal{X}}$ and simultaneously, joint multicalibration in expectation with respect to a class of functions $\mathcal{C}_2 \subset \{0, 1\}^{\mathcal{X}\times R}$ where $R = \{0, \frac{1}{m}, \frac{2}{m}, \ldots, 1\}$ for some $m > 0$. Our algorithm can be viewed as a variant of the original multicalibration algorithm of Hébert-Johnson et al. [2018] (our variant achieves the stronger guarantee of calibration in expectation, first defined in Gopalan et al. [2022]), or a simplification of the split-and-marge algorithm of Gopalan et al. [2022], which replaces the "merge" operation with simple per-update rounding.

First we observe that without loss of generality, we can focus on achieving joint multicalibration for a single class of functions. To see this, note that given $\mathcal{C}_1 \subset \{0, 1\}^{\mathcal{X}}$, we can transform it into an identical class of two argument functions that simply ignore their second argument:

$$\mathcal{C}_2' = \{c \text{ where } c(x, v) = c_1(x) \text{ for every } c_1 \in \mathcal{C}_1\}.$$

Note that if $\hat{f}$ is α-approximately joint-multicalibrated with respect to $\mathcal{C}_2'$, then it is α-approximately multicalibrated with respect to $\mathcal{C}_1$ and vice versa. In other words, in order to be simultaniously multicalibrated with respect to $\mathcal{C}_1$ *and* joint-multicalibrated with respect to $\mathcal{C}_2$, it is sufficient (actually equivalent) to be joint-multicalibrated with respect to $\mathcal{C}_2 \cup \mathcal{C}_2'$. Therefore, we focus on enforcing joint-multicalibration with respect to arbitrary $\mathcal{C} \subset \{0,1\}^{\mathcal{X}\times[0,1]}$.

Before we describe the algorithm, we define the round operation. Write $[\frac{1}{m}] = \{0, \frac{1}{m}, \frac{2}{m}, \dots, 1\}$ for any $m > 0$. We let $f' = \mathit{Round}(f, m)$ to denote the function that simply rounds the output of f to the nearest grid point of $[1/m]$. Similarly, we write $\mathit{Round}(v, m) = \arg\min_{v' \in [1/m]} |v' - v|$ to denote the grid point of $[\frac{1}{m}]$ closest to v.

Algorithm 5: Multicalibration algorithm

Input: $(\alpha, f, \mathcal{C})$
$m = \frac{1}{\alpha}$
$f_0 = \mathit{Round}(f, m)$
$t = 0$
while *there exists a* $c \in \mathcal{C}$ *such that:*

$$\sum_{v \in R} \Pr_{x \sim \mathcal{D}_{\mathcal{X}}}[f_t(x) = v, c(x, v) = 1] \left(v - \Pr_{(x,y)\sim\mathcal{D}}[y | f_t(x) = v, c(x, v) = 1] \right)^2 \geq \alpha$$

do
Let

$$(v_t, c_t) = \arg\max_{v \in R, c \in \mathcal{C}} \Pr_{x \sim \mathcal{D}_{\mathcal{X}}}[f_t(x) = v, c(x, v) = 1] \cdot \left(v - \Pr_{(x,y)\sim\mathcal{D}}[y | f_t(x) = v, c(x, v) = 1] \right)^2$$

$S_t = \{x \in \mathcal{X} : f_t(x) = v, c_t(x, v_t) = 1\}$
$\tilde{v}_t = \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[y | x \in S_t]$
$v_t' = \mathit{Round}(\tilde{v}_t, m)$

Let

$$f_{t+1}(x) = \begin{cases} v_t' & \text{if } x \in S_t \\ f_t(x) & \text{otherwise.} \end{cases}$$

$t = t + 1$
end

Theorem 6. *The output of Algorithm 5* $f_T : \mathcal{X} \to \{0, \alpha, 2\alpha, \dots, 1\}$ *is* $\sqrt{\alpha}$*-approximately jointly multicalibrated with respect to* $\mathcal{C}$ *where* $T \leq \frac{4}{\alpha^2}$.

Proof. By definition, the output of the algorithm f_T is such that

$$\sum_{v \in R} \Pr_{x \sim \mathcal{D}_{\mathcal{X}}}[f_t(x) = v, c(x, v) = 1] \left(v - \Pr_{(x,y)\sim\mathcal{D}}[y | f_t(x) = v, c(x, v) = 1] \right)^2 < \alpha$$

for every $c \in \mathcal{C}$, meaning it is satisfies $\sqrt{\alpha}$-joint calibration:

$$\begin{aligned}
&\sum_{v \in R} \Pr_{x \sim \mathcal{D}_\mathcal{X}}[f_t(x) = v, c(x,v) = 1] \left| v - \Pr_{(x,y)\sim\mathcal{D}}[y | f_t(x) = v, c(x,v) = 1] \right| \\
&\leq \sqrt{\sum_{v \in R} \Pr_{x \sim \mathcal{D}_\mathcal{X}}[f_t(x) = v, c(x,v) = 1] \left(v - \Pr_{(x,y)\sim\mathcal{D}}[y | f_t(x) = v, c(x,v) = 1] \right)^2} \\
&< \sqrt{\alpha}.
\end{aligned}$$

So it suffices to show that the algorithm halts in less than $T \leq \frac{4}{\alpha^2}$ rounds. Define

$$B(f) = \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[(y - f(x))^2].$$

We use B as a potential function and show that we decrease it in each round in the following lemma.

Lemma 21. *For every $t < T$, $B(f_{t+1}) - B(f_t) \leq -\frac{\alpha^2}{4}$*

Proof. Define $\tilde{f}_t$ such that

$$\tilde{f}_t(x) = \begin{cases} \tilde{v}_t & \text{if } x \in B_t \\ f_t(x) & \text{otherwise.} \end{cases}$$

$$B(f_{t+1}) - B(f_t) = \underbrace{\left(B(f_{t+1}) - B(\tilde{f}_t)\right)}_{(*)} + \underbrace{\left(B(\tilde{f}_t) - B(f_t)\right)}_{(**)}$$

Bounding (*):

$$\begin{aligned}
B(f_{t+1}) - B(\tilde{f}_t) &= \Pr_{x \sim \mathcal{D}_\mathcal{X}}[x \in S_t] \cdot \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[(y - f_{t+1}(x))^2 - (y - \tilde{f}_t(x))^2 | x \in S_t] \\
&= \Pr_{x \sim \mathcal{D}_\mathcal{X}}[x \in S_t] \cdot \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[((y - \tilde{f}_t(x)) + (\tilde{v}_t - v'_t))^2 - (y - \tilde{f}_t(x))^2 | x \in S_t] \\
&= \Pr_{x \sim \mathcal{D}_\mathcal{X}}[x \in S_t] \cdot \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[2(y - \tilde{v}_t)(\tilde{v}_t - v'_t) + (\tilde{v}_t - v'_t)^2 | x \in S_t] \\
&\leq \Pr_{x \sim \mathcal{D}_\mathcal{X}}[x \in S_t] \cdot \frac{1}{4m^2}
\end{aligned}$$

where the last inequality follows from the fact that $\tilde{v}_t = \mathbb{E}_{(x,y)\sim\mathcal{D}}[y | x \in S_t]$ and $|\tilde{v}_t - v'_t| \leq \frac{1}{2m}$.

Bounding (**): Because in round t,

$$\sum_{v \in R} \Pr_{x \sim \mathcal{D}_\mathcal{X}}[f_t(x) = v, c(x,v) = 1] \left(v - \Pr_{(x,y)\sim\mathcal{D}}[y | f_t(x) = v, c(x,v) = 1] \right)^2 \geq \alpha,$$

we must have

$$\Pr_{x \sim \mathcal{D}_\mathcal{X}}[x \in S_t] (v_t - \tilde{v}_t)^2 = \Pr_{x \sim \mathcal{D}_\mathcal{X}}[x \in S_t] \left(v_t - \Pr_{(x,y)\sim\mathcal{D}}[y | x \in S_t] \right)^2 \geq \frac{\alpha}{m+1}.$$

Now,we show that

$$\begin{aligned}
B(\tilde{f}_t) - B(f_{t+1}) &= \Pr_{x\sim\mathcal{D}_\mathcal{X}}[x \in S_t] \cdot \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[(y - \tilde{f}_t(x))^2 - (y - f_t(x))^2 | x \in S_t] \\
&= \Pr_{x\sim\mathcal{D}_\mathcal{X}}[x \in S_t] \cdot \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[(y - \tilde{f}_t(x))^2 - ((y - \tilde{f}_t(x)) + (\tilde{v}_t - v_t))^2 | x \in S_t] \\
&= \Pr_{x\sim\mathcal{D}_\mathcal{X}}[x \in S_t] \cdot \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[-2(y - \tilde{v}_t)(\tilde{v}_t - v_t) - (\tilde{v}_t - v_t))^2 | x \in S_t] \\
&\leq \frac{-\alpha}{m+1}
\end{aligned}$$

where the last inequality follows from the fact that $\mathbb{E}_{(x,y)}[y|x \in S_t] = \tilde{v}_t$.

Combining them together, we get

$$\begin{aligned}
B(f_{t+1}) - B(f_t) &\leq \frac{1}{4m^2} - \frac{\alpha}{m+1} \\
&= \frac{\alpha^2}{4} - \frac{\alpha^2}{\alpha+1} \\
&\geq \frac{\alpha^2}{4} - \frac{\alpha^2}{2} \\
&= -\frac{\alpha^2}{4}.
\end{aligned}$$

□

Iterating Lemma 21 over T rounds, we have

$$B(f_T) \leq B(f_0) - T\frac{\alpha^2}{4}.$$

Also, because $B(f) \in [0,1]$ for any f, it must be that $T \leq \frac{4}{\alpha^2}$. □

D Out of Sample Guarantees

In the body of the paper, we assumed that we had direct access to distributional quantities — in particular, we needed to evaluate expectations over the feature distribution. In this section, we show that it is possible to estimate these quantities from modest amounts of unlabeled data sampled from the underlying distribution, and that the guarantees of our algorithm carry over to the underlying distribution. In particular, our algorithm results in a solution to the linear program that approximately satisfies its constraints on the underlying distribution, and achieves objective value that is approximately optimal within its comparison class. The strategy we take is to analyze a slightly modified algorithm (Algorithm 6), which at every stage, uses a fresh sample of data to evaluate the necessary expectations empirically. In particular, it uses a *new* sample at every iteration, and so has sample complexity that scales linearly with the number of iterations. Using techniques from adaptive data analysis Dwork et al. [2015], Bassily et al. [2016], Jung et al. [2021b] similar to how they are used by Hébert-Johnson et al. [2018] to prove sample complexity bounds, we could reduce our linear dependence on T in our sample complexity bound by a quadratic factor by reusing data across rounds, but we settle for the conceptually simpler bound here.

Theorem 7. *Fix any distribution $\mathcal{D}$, hypothesis class $\mathcal{H}$, class of group indicators $\mathcal{G}$, dual bound C, and $\epsilon, \delta > 0$. After T rounds, with probability $1-\delta$, Algorithm 6 outputs a randomized hypothesis $\bar{h}$ such that $err(\bar{h}) \leq OPT + \frac{2}{C} + 8\epsilon$ and $\omega_g|\rho_g(\bar{h}) - \rho(\bar{h})| \leq \gamma + \frac{1}{C} + \frac{2}{C^2} + \frac{8\epsilon}{C}$, where OPT is the objective value of the optimal solution of $\psi(f, \gamma, \mathcal{H}_A)$. It makes use of $m = O\left(T\frac{\log(\frac{2T|G|}{\delta})}{2\epsilon^2}\right)$ samples of unlabeled data drawn i.i.d. from $\mathcal{D}_\mathcal{X}$. Here T is as specified in the algorithm: $T = \frac{1}{4} \cdot C^2 \cdot (C^2 + 4|\mathcal{G}|)^2$.*

Lemma 22. *Fix any distribution $\mathcal{D}$, hypothesis class $\mathcal{H}$, and class of group indicators $\mathcal{G}$. In a single round t of Algorithm 6 with $S_t \sim \mathcal{D}^m$ for $m = O(\frac{\log(\frac{2|G|}{\delta})}{2\epsilon^2})$, Algorithm 6 returns a hypothesis h_t that with probability $1-\delta$ satisfies for all $g \in G$:*

$$|err(h_t, g, \mathcal{D}) - err(h_t, g, S_t)| \leq \epsilon$$
$$|\rho(h_t, g, \mathcal{D}) - \rho(h_t, g, S_t)| \leq \epsilon.$$

Algorithm 6: Projected Gradient Descent Algorithm

Input: ($\mathcal{D}$: *data distribution,* $f : \mathcal{X} \to [0,1]$*: regression function,* $\mathcal{G}$*: groups,* γ*: tolerance on fairness violation,* C*: bound on dual* ($\|\lambda\|_1 \leq C$)*,* η*: learning rate,* $m = \frac{\log(\frac{2|G|}{\delta})}{2\epsilon^2}$*: batch size of fresh data for each round of gradient descent,* ϵ*: per round estimation error,* δ*: failure probability*)

Initialize dual vector $\lambda^0 = \mathbf{0}$ and set $T = \frac{1}{4} \cdot C^2 \cdot (C^2 + 4|\mathcal{G}|)^2$.

for $t = 1, \ldots, T$ **do**

Primal player updates h_t

$$h_t(x) = \begin{cases} 1, & \text{if } f(x) \geq \frac{1+\sum_{g\in\mathcal{G}} \lambda_g^{t-1}(g(x)-\beta_g)}{2+\sum_{g\in\mathcal{G}} \lambda_g^{t-1}(g(x)-\beta_g)} \text{ and } (2+\sum_{g\in\mathcal{G}} \lambda_g^{t-1}(g(x)-\beta_g)) > 0, \\ 0, & \text{if } f(x) < \frac{1+\sum_{g\in\mathcal{G}} \lambda_g^{t-1}(g(x)-\beta_g)}{2+\sum_{g\in\mathcal{G}} \lambda_g^{t-1}(g(x)-\beta_g)} \text{ and } (2+\sum_{g\in\mathcal{G}} \lambda_g^{t-1}(g(x)-\beta_g)) > 0, \\ 1, & \text{if } f(x) \leq \frac{1+\sum_{g\in\mathcal{G}} \lambda_g^{t-1}(g(x)-\beta_g)}{2+\sum_{g\in\mathcal{G}} \lambda_g^{t-1}(g(x)-\beta_g)} \text{ and } (2+\sum_{g\in\mathcal{G}} \lambda_g^{t-1}(g(x)-\beta_g)) < 0, \\ 0, & \text{if } f(x) > \frac{1+\sum_{g\in\mathcal{G}} \lambda_g^{t-1}(g(x)-\beta_g)}{2+\sum_{g\in\mathcal{G}} \lambda_g^{t-1}(g(x)-\beta_g)} \text{ and } (2+\sum_{g\in\mathcal{G}} \lambda_g^{t-1}(g(x)-\beta_g)) < 0. \end{cases}$$

Sample S_t i.i.d. from $\mathcal{D}^m$ Compute

$$\hat{\rho}_g^t = \mathop{\mathbb{E}}_{(x,y)\sim S_t}[\ell(h_t(x), 0) g(x)(1 - f(x))] \text{ for all } g \in \mathcal{G},$$
$$\hat{\rho}^t = \mathop{\mathbb{E}}_{(x,y)\sim S_t}[\beta_g \ell(h_t(x), 0)(1 - f(x))], \text{ where } \beta_g = \Pr[g(x) = 1 | y = 0]$$

Dual player updates

$$\lambda_g^{t,+} = \max(0, \lambda_g^{t,+} + \eta \cdot (\hat{\rho}_g^t - \hat{\rho}^t - \gamma)),$$
$$\lambda_g^{t,-} = \max(0, \lambda_g^{t,-} + \eta \cdot (\hat{\rho}^t - \hat{\rho}_g^t - \gamma)).$$

Dual player sets $\lambda^t = \sum_{g\in\mathcal{G}} \lambda_g^{t,+} - \lambda_g^{t,-}$.

If $\|\lambda^t\|_1 > C$, set $\lambda^t = C \cdot \frac{\lambda^t}{\|\lambda^t\|_1}$.

end

Output: $\bar{h} := \frac{1}{T}\sum_{t=1}^{T} \hat{h}_t$, a uniformly random classifier over all rounds' hypotheses.

Theorem 8 (Chernoff-Hoeffding Bound)**.** *Let $X_1, X_2, \ldots, X_m$ be i.i.d. random variables with $a \leq X_i \leq b$ and $\mathbb{E}[X_i] = \mu$ for all i. Then, for any $\alpha > 0$,*

$$\Pr\left(\left|\frac{\sum_i X_i}{m} - \mu\right| > \alpha\right) \leq 2\exp\left(\frac{-2\alpha^2 m}{(b-a)^2}\right).$$

Proof of Lemma 22. This claim follows by applying a Chernoff-Hoeffding bound with $m \geq \frac{\ln(\frac{2|G|}{\delta})}{2\epsilon^2}$ □

Proof Sketch of Theorem 7. Taking $m > \frac{\log(\frac{2T|G|}{\delta})}{2\epsilon^2}$, we have that in a single round t of our algorithm we are able to estimate the true distributional classification and fairness constraint errors up to an additive error of ϵ

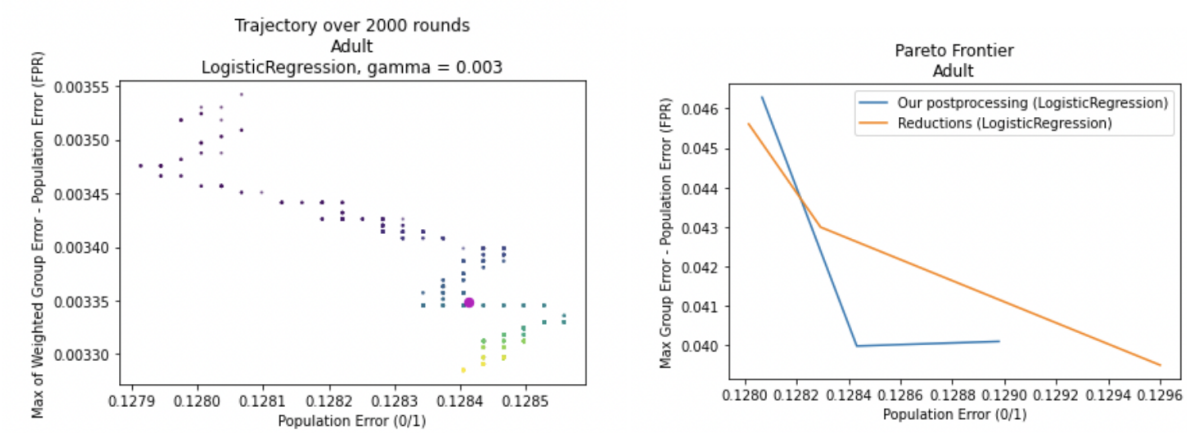

Figure 2: The plot on the left is a trajectory over 2000 iterations of gradient descent of our method post-processing a base logistic regression model, for a single value of $\gamma = 0.003$. The trajectory starts at the top of the figure and moves downwards with time, and the purple point represents the uniform distribution over the constituent models of the 2000 iterations. The plot on the right shows the Pareto curve for our method (blue) and for the fair reductions [Agarwal et al., 2018] for constraint values ranging between $0.0025 \leq \gamma \leq 0.00355$.

with probability $1 - \delta/T$ — and hence with probability $1 - \delta$, we estimate these quantities up to additive error ϵ uniformly over all T rounds. We can then make one small modification to the analysis of Algorithm 1. First observe that since the primal player's best response does not depend on any estimation of a distributional quantity based on the sample S_t, their regret is still zero, as it is in the analysis of Algorithm 1. The dual player, on the other hand, is given loss vectors that deviate from the versions that would have been computed on the underlying distribution by at most 2ϵ in ℓ_∞ norm, and hence experience additional regret (to the true distributional quantities) larger than in the analysis of Algorithm 1 by up to an additional additive 4ϵ. Consequently, the equilibrium solution $(\bar{h}, \bar{\lambda})$ from Algorithm 6 is an $4\epsilon + 1/C$ approximate equilibrium to the zero-sum game of 2 which then, applying Theorem 2, yields a $\frac{2}{C} + 8\epsilon$ approximate solution to the objective of the original linear program. □

E Expanded Experimental Discussion

We provide additional experimental evaluation on the UCI Adult dataset [Dua and Graff, 2017]. The sensitive attributes we use are binary gender and race, categorized as White, Black, Asian and Pacific Islander, American Indian or Eskimo, and Other. Note that race and gender are intersecting attributes. In these experiments our algorithm is post-processing a standard sklearn logistic regression model, notably, as in our previous results, not guaranteed to be multicalibrated in the ways our theory requires. We also provide a comparison to the popular in-processing "fair reductions" method [Agarwal et al., 2018]. We note that our algorithm performs competitively, even Pareto-dominating certain points on the reductions Pareto frontier. However, the reductions method is also able generate points corresponding to constraint violations that our method is not able to access — this does not violate our theoretical findings, since we are not starting with a multicalibrated regression function. Our method requires solving a single logistic regression problem over the dataset (to compute the regression model $\hat{f}$ that we post-process), whereas the method of Agarwal et al. [2018] requires solving a regression problem at every iteration.