Reimagining Meaningful Model Multiplicity

Anonymous Authors[1]

Abstract

Predictive model multiplicity is when multiple models in a hypothesis class $\mathcal{H}$ disagree on their predictions while having similar overall error to the risk minimizer h^*. This has substantial implications for algorithmic fairness: if such multiplicity exists, then there may be a model competitive with the risk minimizer of the target hypothesis class which does substantially better in terms of target fairness metrics; some argue that searching for such a less discriminatory model is a legal duty. Standard approaches formalize this via search over a *Rashomon* set $\mathcal{R}(\mathcal{H})$ of models, but this search is computationally expensive and does not easily provide generalization guarantees. We propose a reframing to "meaningful" model multiplicity that avoids both obstacles. Rather than asking how to find the most fair model in $\mathcal{R}(\mathcal{H})$, we ask whether any $h \in \mathcal{H}$ can outperform h^* on a target group g. We provide efficient ensembling techniques over models which witness such multiplicity which simultaneously will outperform h^* on all affected groups and with generalization guarantees. We further show that multiaccuracy with respect to $\mathcal{G} \times \mathcal{H}$—efficiently achievable via a polynomial number of oracle calls to $\mathcal{H}$—entirely precludes meaningful multiplicity. Finally, for constraint-based fairness metrics for classification tasks, we show that building a multicalibrated predictor for the label probability $\mathbb{E}[y|x]$ and then postprocessing it to satisfy the fairness guarantees will Pareto dominate any approach based on searching the Rashomon set of classifiers.

1. Introduction

As machine learning is adopted in consequential settings, the fairness of such models in terms of their treatment of different demographic groups is an important consideration. One observation that has been central to this discussion is the phenomenon of *model multiplicity*: even within a set hypothesis class $\mathcal{H}$, there may exist many models competitive with the risk minimizer h^* in terms of overall accuracy but which differs substantially in terms of predictions on specific subpopulations. This multiplicity may be observable via perturbations of the training data or pipeline, or via a full examination of the optimization landscape. The legal significance of this multiplicity and the associated implications in terms of arbitrariness of decision-making have been articulated in a variety of recent scholarship [E.g. (Black et al., 2022; 2024; Laufer et al., 2025; Dai et al., 2025)]: practitioners may bear a duty to search for less discriminatory alternatives within their model class. This is typically operationalized as a search over the "Rashomon" set of models that are ϵ- close to the risk minimizer of some hypothesis class $\mathcal{H}$ of models.

Despite its intuitive appeal, search among this Rashomon set faces fundamental obstacles that limit its practical utility. The first is computational: identifying the most fair model within $\mathcal{R}(\mathcal{H})$ is in general computationally intractable, requiring either explicit enumeration of an exponentially large model space or solving constrained, non-convex optimization problems (Laufer et al., 2025; Dai et al., 2025). The second obstacle is statistical: many existing methods assume direct access to the Bayes optimal predictor or do not easily generalize outside of the sample.

In this paper, we propose a reframing of model multiplicity that resolves both obstacles while yielding stronger fairness guarantees. The key conceptual shift is that while the existence of the Rashomon set is a witness to multiplicity, we do not need to search over this set. Rather, we can instead learn—and *boost*—over the full space of models in hypothesis class $\mathcal{H}$, and we can do so efficiently. We first show this while considering notions of fairness which are aligned with accuracy, in the sense that a model is considered "more fair" than another on some subgroup if it is more accurate than another on that subgroup. In other words, for this form of fairness, **we provide constructive methods to outperform any model in the Rashomon set in terms of both fairness and accuracy, while bypassing the need to construct the Rashomon set itself in the first place.**

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

A reader interested in other constraint-based notions of fairness, such as equalizing error rates or parity of positive labels across subpopulations might be skeptical of such claims or consider them specific to the above "accuracy-aligned" notions of fairness. In the second part of the paper, we leverage tools from the multicalibration literature to show that in fact, the above techniques are generally advantageous, even if the goal is a form of fairness that is in tension with accuracy. Specifically, we can efficiently post-process multicalibrated predictors to find fair classifiers—and the classifier we find will be competitive with the best achievable fairness-accuracy tradeoff, without ever searching the Rashomon set of classifiers. Thus, **we show that even when fairness conflicts with accuracy, a two-step approach: first eliminating meaningful model multiplicity, then postprocessing for fairness, is preferable to directly searching the Rashomon set of classifiers.**

1.1. Technical Overview

We consider supervised machine learning of models mapping features $\mathcal{X}$ to outcomes $\mathcal{Y}$, where samples have been drawn from a distribution $\mathcal{D} \in \Delta(\mathcal{X} \times \mathcal{Y})$. In general we will be interested in cases where the true underlying label $y \in \{0, 1\}$ is binary. However, we will consider two different prediction mechanisms to learn this label. In Sections 3–6, we consider learning *real-valued* predictors $f : \mathcal{X} \to [0, 1]$ to estimate the conditional label probability $\mathbb{E}[y|x]$. In Section 7, we shift to classifiers $c : \mathcal{X} \to \{0, 1\}$—which can be generated, for instance, by thresholding the real-valued predictions of f—and show that even practitioners whose ultimate goal is classification are best served by the regression-first approach developed here.

In a perfect world, we would have access to the *Bayes optimal* predictor, which for every datapoint could perfectly estimate $\mathbb{E}[y|x]$. However, in practice due to the constraints of the finite data regime, we can only learn models which approximate Bayes optimality. In particular, we consider real-valued predictors $h \in \mathcal{H}$ mapping from $\mathcal{X}$ to $[0, 1]$. We then discuss "model multiplicity-aware" mechanisms to *boost* over $\mathcal{H}$, generating a model outside of $\mathcal{H}$ which improves over any model in $\mathcal{H}$ in terms of both fairness and overall accuracy, while still admitting good generalization guarantees when trained in-sample.

How should one pick a model within $\mathcal{H}$? The standard approach is empirical risk minimization: choosing $h^* \in \mathcal{H}$ to minimize squared error loss $\ell(h(x), y) = (h(x) - y)^2$, with the guarantee that in-sample estimates *generalize well* to the underlying distribution $\mathcal{D}$ given a large enough sample $D \sim \mathcal{D}^n$. This framing implicitly treats h^* as the uniquely correct model—but in practice, there may be many models in $\mathcal{H}$ that are competitive with h^* on overall error while differing substantially in their predictions on specific subgroups. Whether and when such *model multiplicity* is meaningful, and what to do about it, is the central question of this paper, as formalized next.

Meaningful Model Multiplicity. We are interested in model multiplicity within the class of models $\mathcal{C}$ or $\mathcal{H}$. In other words, we consider the question of how to handle what happens when there are multiple models in $\mathcal{C}$ or $\mathcal{H}$ which are competitive with the loss-minimizer c^* or h^*. We will call this the *ε-Rashomon set* of $\mathcal{H}$:

$$\mathcal{R}(\mathcal{H}) = \{h \in \mathcal{H} : \mathbb{E}[\ell(h(x), y)] \leq \mathbb{E}[\ell(h^*(x), y)] + \varepsilon\}.$$

One object of interest is how models in this set might perform on different subgroups. We will envision this class of groups, $\mathcal{G}$, as identifiable from the features $\mathcal{X}$: in other words, for any subgroup there will be a deterministic binary function $g : \mathcal{X} \to \{0, 1\}$ such that $g(x)$ is 1 for members of that subgroup (which we will call "group g", somewhat abusing notation), and otherwise is 0.

The question of model multiplicity for group fairness on a group g is whether or not there exists a model h contained within the Rashomon set $\mathcal{R}(\mathcal{H})$ which is *better*, from the perspective of those in group g, than the risk minimizer h^*. What "better" means depends on what one means by a fair model: is it one which makes less errors on the group, is it one which classifies members of that group as 1 at similar rates as other groups, is it one which has some notion of similar error rates (overall, or in terms of false positives or negatives) to other groups, or so on. In the context of our real-valued prediction problem of estimating conditional label probabilities—a setting where the group-constrained notions of selection rates or false positives, for instance, does not make sense—in the first half of the paper take the narrow view of fairness as minimizing groupwise error rates. In other words, we will be interested in models $h_g \in \mathcal{R}(\mathcal{H})$ such that $\mathbb{E}[\ell(h_g(x), y)|g(x) = 1] < \mathbb{E}[\ell(h^*(x), y)|g(x) = 1]$. Later, in section 7, we will show that such a view will then, somewhat surprisingly, also allow us to also efficiently handle other fairness metrics which *are* in tension with accuracy, via an efficient postprocessing of the predictor which has no "meaningful" multiplicity with respect to the hypothesis class it is learning over.

What we mean by *meaningful* multiplicity is the following: h_g is not merely competitive overall, but strictly better for group g specifically. This is a stronger condition than Rashomon set membership alone, which places no constraint on group-level performance—a model drawn arbitrarily from the Rashomon set may perform no better, or even worse, on the target subgroup than h^*.

The mechanisms we develop for finding (and ensembling) such models, discussed in Sections 3 through 7, additionally enjoy generalization guarantees, a property that existing

approaches to searching the Rashomon set do not in general provide. As we show in section 3, the presence of meaningful multiplicity in the Rashomon set of a class $\mathcal{H}$ reveals that its loss minimizer h^* is provably suboptimal with respect to the Bayes optimal classifier. Furthermore, there is signal in $\mathcal{H}$, witnessed by the presence of the multiplicitous h_g, that h^* is failing to capture on group g. A natural response is to ask whether we can exploit that signal to build something better.

Resolving Meaningful Multiplicity via Ensembling. Given a collection of models $h_1, \ldots, h_k$ displaying meaningful multiplicity on groups $\mathcal{G}$ with respect to h^*, how should we combine them into a predictor $f : \mathcal{X} \to \mathcal{Y}$? We ask two things of the ensemble. First, it should do no harm: none of $h_1, \ldots, h_k$ should display meaningful multiplicity with respect to f, meaning f matches or outperforms each h_i on its target group. Second, it should outperform h^* overall, and ideally each of $h_1, \ldots, h_k$ as well. Finally, the method must generalize well, even when k is large.

In Section 5 we provide two ensembling techniques satisfying these desiderata. The first constructs a carefully structured decision list over h^* and $h_1, \ldots, h_k$ that performs well on all groups in $\mathcal{G}$ simultaneously, even when groups overlap. To guarantee generalization even when k is large, we leverage tools from adaptive data analysis—which allow statistical queries about different groups to be answered from shared data without the usual penalties for multiple comparisons—to guarantee generalization out of sample. The second method shows that when $|\mathcal{G}|$ is not too large, the best linear combination of h^* and $h_1, \ldots, h_k$ also satisfies the desiderata.

Moving Beyond the Rashomon Set. Now that we know how to combine models which display such multiplicity, how should we find them to begin with? In general, searching over the space of all models in the Rashomon set of $\mathcal{H}$ is a computationally hard problem. Luckily for us, we show that we need not consider this search at all. Crucially, because our goal is not to select a single model from $\mathcal{H}$ but to *ensemble* over models in $\mathcal{H}$—producing a predictor outside $\mathcal{H}$ entirely—individual components need not be globally competitive with h^*, only locally informative on the target subgroup. Hence, we can search within the entirety of $\mathcal{H}$ for such a signal.

Multiaccuracy Precludes Meaningful Multiplicity. Searching all of $\mathcal{H}$ might seem harder than searching $\mathcal{R}(\mathcal{H})$, but in Section 6 we show that the opposite is true: we can efficiently construct an ensemble f that performs at least as well as any model in $\mathcal{H}$ on every group in $\mathcal{G}$, while guaranteeing generalization. The key is *multiaccuracy*: a model f that is multiaccurate with respect to $\mathcal{G} \times \mathcal{H}$—meaning no $h \in \mathcal{H}$ can identify a subgroup in $\mathcal{G}$ on which f has systematic prediction error—is guaranteed to have no competitive outperforming model in $\mathcal{H}$, and thus no model in $\mathcal{R}(\mathcal{H})$, for any group in $\mathcal{G}$. We show that such a multiaccurate predictor can be efficiently constructed with only a polynomial number of squared error oracle calls to $\mathcal{H}$.

Optimal Fair Classification via Multicalibrated Postprocessing. A predictor of conditional label probabilities will ultimately be thresholded to produce a binary classifier. How should we reason about fairness and multiplicity in this classification setting? The methods proposed thus far might seem inapplicable in some such settings, where fairness—characterized as equalizing error rates across groups—is in direct tension with overall accuracy, unlike in our group error minimization formulation above.

In Section 7, we argue that the right approach is to first generate a predictor f that minimizes error on all groups in $\mathcal{G}$ simultaneously, and then *postprocess* f to satisfy the desired constraint-based fairness metric. This approach weakly dominates any classifier derived by searching the Rashomon set of classifiers in $\mathcal{H}$: specifically, postprocessing f approximates the best possible fair postprocessing of the Bayes optimal predictor, and hence weakly dominates any point on the accuracy-fairness Pareto front achievable by thresholding models in $\mathcal{H}$.

This result requires a stronger condition on f than the multiaccuracy used in the preceding section. When thresholding f to produce a classifier, we are effectively conditioning on f's predictions; we therefore need systematic prediction error to be absent not just marginally across $\mathcal{G}$, but also within the level sets of f. This is precisely *multicalibration*. If f is multicalibrated with respect to $\mathcal{H}$, $\mathcal{G}$, and a carefully chosen class of thresholding functions—so that no $h \in \mathcal{H}$ can identify systematic prediction error on any group in $\mathcal{G}$, even after thresholding and conditioning on f's predictions—then postprocessing f yields the Pareto dominance result above.

1.2. Related Work

Search for Fair Models and the Rashomon Set. Discussions of model multiplicity and the Rashomon set extend at least as far back as (Breiman, 2001). There is a wide range of works discussing multiplicity as it pertains to algorithmic fairness and the legal obligation to search for less discriminatory models and the statistical and computational burdens implied by such a search [e.g. (Black et al., 2022; 2024; Laufer et al., 2025; Hays et al., 2025; Xin et al., 2022; Rudin et al., 2024)]. In contrast to these works, which with the exception of (Hays et al., 2025) emphasize finite-sample guarantees, we adopt a statistical learning perspective: rather than reasoning about relabelings or explicitly enumerated

candidate models, we directly optimize over the hypothesis class and obtain distributional guarantees through boosting procedures. This allows us to avoid the explicit search of large Rashomon sets while constructively identifying predictors with improved performance.

Model Reconciliation. Our work is conceptually close to the model reconciliation framework introduced by (Roth et al., 2023) and extended empirically in (Behzad et al., 2025). This work leverages disagreement between models to systematically update them as a mechanism to remediate multiplicity. Our work shares this intuition, but extends it and differs fundamentally in its algorithmic perspective: while their reconciliation procedure operates over an explicitly generated collection of competing models drawn from the Rashomon set and iterating pairwise to resolve disagreement among them, we directly boost over the underlying hypothesis class, without ever explicitly constructing the Rashomon set. From this perspective, pairwise disagreement is merely one witness of residual structure in the hypothesis class rather than the primary learning object.

Multiaccuracy, Multicalibration, and Agreement. The techniques which underpin our results are from the fairness literature on multiaccuracy and multicalibration: boosting techniques which build predictors whose errors are uncorrelated on computationally identifiable subgroups (Hébert-Johnson et al., 2018; Kim et al., 2019). In particular, we rely on results from (Globus-Harris et al., 2022; 2023a) and (Globus-Harris et al., 2023b) in our boosting procedures. The way in which we leverage disagreement is also inspired by a line of work on learning via *agreement* protocols, though here we are strictly in a single-party setting (Aumann, 1976; Aaronson, 2005; Collina et al., 2025; 2026; Kearns et al., 2026).

2. Preliminaries

We consider supervised machine learning of models f mapping features $\mathcal{X}$ to outcomes $\mathcal{Y}$. We will in general consider the case of regression, where outcomes $\mathcal{Y}$ are real valued and bounded in $[0, 1]$ (although all results scale to regression problems where the label range is bounded), where models will be evaluated with respect to the expected value L of a pointwise loss function $\ell : \mathcal{Y} \times \mathcal{Y} \to [0, 1]$, often their squared error. We will be interested in this loss overall as well as with respect to specific groups $g \in \mathcal{G}$ identifiable from features $\mathcal{X}$. Our "gold standard" regression model will be the Bayes Optimal model, which we will denote f^*. We will consider the risk minimizing model h^* within some hypothesis class $\mathcal{H}$ and the Rashomon set $\mathcal{R}(\mathcal{H})$ of models ϵ-close to this model. For brevity, formalizations of these definitions are deferred to Appendix A; in general all proofs in the forthcoming sections are deferred to Appendices B to G.

3. Moving Beyond the Rashomon Set

There are several challenges to standard approaches to model multiplicity: the Rashomon set of close-to-optimal models is computationally hard to search exactly, particularly in ways that generalize out of sample. Furthermore, once enumerated, the set may be large enough that statistically valid approaches to model choice among these models can be prohibitive. Here, we propose an alternative method: consideration of models which witness meaningful improvement in accuracy for subpopulations of interest.

Definition 3.1 (Meaningful Multiplicity)**.** *Fix a hypothesis class $\mathcal{H}$, benchmark error-minimizing model model $h^* \in \mathcal{H}$ such that $h^* = \arg\min_{h \in \mathcal{H}} \mathbb{E}[\ell(h(x), y)]$, and a group indicator $g : \mathcal{X} \to \{0, 1\}$. We say that $\mathcal{H}$ exhibits* meaningful multiplicity *on g at level $\varepsilon > 0$ if there exists $h \in \mathcal{H}$ such that*

$$\mathbb{E}_g\big[\ell(h, y)\big] \leq \mathbb{E}_g\big[\ell(h^*(x), y)\big] - \varepsilon.$$

We will also equivalently say that h witnesses multiplicity with respect to h^* on group g.

While this notion of multiplicity as written is not explicitly with respect to disagreement in predictions and is instead defined in terms of witnessing a difference in group-specific loss, it directly implies meaningful disagreement among the models.

Lemma 3.2 (Meaningful Multiplicity implies Disagreement)**.** *Let $h^* = \arg\min_{h \in \mathcal{H}} \mathbb{E}\big[(h(x) - y)^2\big]$ be the squared-error minimizing model in $\mathcal{H}$. Suppose there exists a group g with density μ_g and a model $h_g \in \mathcal{H}$ such that*

$$\mathbb{E}_g\big[(h_g(x) - y)^2\big] \leq \mathbb{E}_g\big[(h^*(x) - y)^2\big] - \varepsilon.$$

Then the expected squared disagreement between h_g and h^* on g is at least $\mathbb{E}_g\big[(h_g(x) - h^*(x))^2\big] \geq \frac{\varepsilon^2}{4}$, and overall $\mathbb{E}\big[(h_g(x) - h^*(x))^2\big] \geq \mu_g \cdot \frac{\varepsilon^2}{4}$.

However, a substantial *difference* in this formulation of multiplicity compared to previous work is that models which admit this form of multiplicity *may not belong to the Rashomon set*.

Lemma 3.3. *A model h which witnesses multiplicity with respect to h^* on group g with density $\mu_g < 1/2$ may not belong to the ϵ-Rashomon set of h^* for any $\varepsilon \in (0, 1)$ such that $\varepsilon < 1 - 2\mu_g$.*

This may seem to be a troubling claim—surely expanding the space of models we are searching over would lead to worse efficiency and generalization! However, we will show in sections 6 and 7 that this is not the case, as we will be able to use an efficient number of calls to a squared error oracle to search over this space.

4. Model Multiplicity Implies Models are Far From Optimal

Before we build these constructions for searching over this space of models, it will be useful to build some intuition for what it means to have meaningful model multiplicity. Namely, if there is disagreement between the overall risk minimizer h^* and any other model which, for individual groups, does better, then the risk minimizer h^* must be missing some relevant signal for those subgroups that the other models are witnessing.

Lemma 4.1 ((Case of disjoint groups)**.** *Let h^* be the error-minimizing model in convex hypothesis class $\mathcal{H}$ on distribution $\mathcal{D}$ and let f^* be the Bayes optimal model. Let $g_1, \ldots, g_k$ be disjoint groups, where g_i has probability mass μ_i. Let $h_{g_i} \in \arg\min_{h \in \mathcal{H}} \mathbb{E}_{g_i}[\ell(h(x), y)]$ be the error-minimizing model for group g_i, and let $\delta_i = \mathbb{E}_{g_i}[(h^*(x) - h_{g_i}(x))^2]$ be the expected squared disagreement of h^* and h_{g_i} on group g_i. Then*

$$\mathbb{E}[(h^*(x) - y)^2] - \mathbb{E}[(f^*(x) - y)^2] \geq \sum_{i \in [k]} \mu_i \delta_i.$$

Lemma 4.2 ((Overlapping groups))**.** *Let h^* be the error-minimizing model in convex hypothesis class $\mathcal{H}$ on distribution $\mathcal{D}$ and let f^* be the Bayes optimal model. Let $g_1, \ldots, g_k$ be a collection of groups such that $m = \max_{x \in \mathcal{X}} |\mathcal{S}(x)|$ is the largest number of groups any point in $\mathcal{X}$ belongs to and where g_i has probability mass μ_i. Let $h_{g_i} \in \arg\min_{h \in \mathcal{H}} \mathbb{E}_{g_i}[\ell(h(x), y)]$ be the error-minimizing model for group g_i, and let $\delta_i = \mathbb{E}_{g_i}[(h^*(x) - h_{g_i}(x))^2]$ be the expected squared disagreement of h^* and h_{g_i} on group g_i. Then*

$$\mathbb{E}[(h^*(x) - y)^2] - \mathbb{E}[(f^*(x) - y)^2] \geq \frac{1}{m} \sum_{i \in [k]} \mu_i \delta_i.$$

The proofs of the above theorems 4.1 and 4.2, deferred to Appendix C, are constructive: they build a model which is an ensemble over the models $h^*, h_{g_1}, \ldots, h_{g_k}$ and which is provably closer to the Bayes optimal model than h^*.

5. Resolving Model Multiplicity via Ensembling

The constructions used for the bounds above Section 4 demonstrate that witnessing multiplicity provides natural mechanisms to *improve the predictor*, by boosting over the collection of disagreeing models. However, the constructions in the proof of Lemma 4.1 is overly restrictive since it assumes disjoint groups, and the construction in the proof of 4.2 requires enumeration of all intersections of groups, which will be exponential. Here, we describe two alternative and efficient approaches to ensembling multiplicitous models. Here, we will still assume that we have been handed a collection of "multiplicitous" models; in subsequent sections we will remove this assumption. As a note, these methodologies can be considered as complementary to the reconciliation approach recommended by (Roth et al., 2023) and (Behzad et al., 2025), but with less operations required. The proofs and algorithmic details are deferred to Appendices D and E.

5.1. Ensembling via Decision Lists

The first approach generates a decision list over models $h^*, h_{g_1}, \ldots, h_{g_k}$. Note that naively performing this form of update is insufficient to guarantee that f cannot witness meaningful model multiplicity if the groups overlap. For instance, imagine that group g_1 is fully contained within group g_2, and h_{g_1} has better squared error on g_1 than g_2. Then, if g_2 appears in the decision list after g_1, the datapoints in g_1 will be evaluated according to g_2 rather than g_1, and hence h_{g_1} will still be a witness to model multiplicity with respect to f on g_1. Hence, order matters. In (Globus-Harris et al., 2022), the authors provide an efficient algorithm in the context of "bias bounties", which guarantees that group errors always decrease as the decision list is expanded. In the context of model multiplicity, this then can be applied as a method to boost over the models $h_{g_1}, \ldots, h_{g_k}$ such that the final ensemble f will not display meaningful model multiplicity with respect to any of the models $h^*, h_{g_1}, \ldots, h_{g_k}$.

Theorem 5.1 (Mitigating Multiplicity with Bias Bounties)**.** *Let $(g_1, h_{g_1}), \ldots, (g_k, h_{g_k})$ be a collection of group indicators g_i and models h_{g_i}. Let f be the model output by Algorithm 3 with $(g_1, h_1), \ldots, (g_k, h_k)$ as input tuples, initial model f_0, and error tolerance γ. Then $h_{g_1}, \ldots, h_{g_k}$ will not witness meaningful model multiplicity on $g_1, \ldots, g_k$ on the updated model f.*

5.2. Ensembling via Linear Combinations

Another approach would be to take the best linear combination over products of models and groups. Due to the first-order conditions of ordinary least squares, the final model's features will be orthogonal to the models' residuals in expectation, and hence all multiplicity will be removed by the final ensemble.

Theorem 5.2. *Let $(h_{g_1}, \ldots, h_{g_k})$ be a collection of models h_{g_i} which witness meaningful multiplicity with respect to h^* on groups $g_1, \ldots, g_k$ at level ϵ. Let f be the best linear combination of models weighted by group membership on*

distribution $\mathcal{D}$:

$$f(x) = \beta_0 h^*(x) + \sum_{i=1}^{k} \beta_{i_1} h_{g_i}(x) g_i(x) + \beta_{i_2} h^*(x) g_i(x), \tag{1}$$

where

$$\beta = \arg\min_{\beta' \in \mathbb{R}^{2k+1}} \mathbb{E}_{\mathcal{D}} \left[\left(y - \beta'_0 h^*(x) - \sum_{i=1}^{k} \beta'_{i_1} h_{g_i}(x) g_i(x) - \beta'_{i_2} h^*(x) g_i(x) \right)^2 \right].$$

Then, for all $i \in [k]$, h_{g_i} does not witness multiplicity with respect to f on g_i. Moreover, if $\mathbb{E}_{g_i}[(h_{g_i}(x) - f(x))^2] > 0$, then f strictly outperforms h_{g_i} on g_i.

6. Resolving Model Multiplicity via Multiaccuracy

There are a few challenges in the above approaches. Firstly, they require that the alternative models $h_{g_1}, \ldots, h_{g_k}$ be already generated, which may be costly in terms of sample complexity. Secondly, it requires an enumeration of all group identifiers $g \in \mathcal{G}$. In practice, one might wish to achieve good performance on many complex and intersecting groups g, such that their full enumeration is prohibitive. Here, we observe that two relatively weak conditions: multiaccuracy and self-orthogonality, which may be achieved without explicit enumeration of groups, suffice. In other words, a predictor which is multiaccurate with respect to a collection of groups $\mathcal{G}$ and hypotheses $\mathcal{H}$ will not be able to find a hypothesis in $\mathcal{H}$ which can improve with respect to f on group g. Thus, if a model in $\mathcal{H}$ is "multiplicitous" with respect to f, its disagreements with f are not *useful* in that they do not improve model performance compared to f.

Definition 6.1 (Multiaccuracy)**.** *Fix a distribution $\mathcal{D} \in \Delta\mathcal{Z}$ and a model $f : \mathcal{X} \to [0, 1]$ that maps onto a countable subset of its range. Let $\mathcal{H}$ be an arbitrary collection of real valued functions $h : \mathcal{X} \to \mathbb{R}$. We say that f is α-approximately multiaccurate with respect to $\mathcal{D}$ and $\mathcal{H}$ if for every $h \in \mathcal{H}$:*

$$\mathbb{E}_{(x,y)\sim\mathcal{D}}[h(x)(y - f(x))] \leq \alpha.$$

Definition 6.2 (Self-Orthogonality of f with respect to $\mathcal{G}$)**.** *We will say that a model f is η self-orthogonal on a collection of subgroups $\mathcal{G}$ if*

$$|\mathbb{E}[g(x)f(x)(f(x) - y)]| \leq \eta.$$

Theorem 6.3. *Let $\mathcal{G}$ be any (potentially uncountable) collection of group indicator functions and let $\mathcal{H}$ be a hypothesis class mapping to real-valued predictions $\mathcal{Y}$. Let $f : \mathcal{X} \to \mathcal{Y}$ be a η-multiaccurate with respect to the product class $\{g \cdot h : g \in \mathcal{G}$ and $h \in \mathcal{H}\}$ and η-self-orthogonal with respect to $\mathcal{G}$. Then there will be no model $h \in \mathcal{H}$ which witnesses model multiplicity with respect to $g \in \mathcal{G}$ with density μ_g at a level $\varepsilon > 4\eta/\mu_g$.*

How would one build such a multiaccurate predictor? One way would be to apply the same ensembling technique as in Section 5.2, by fitting a linear predictor to a collection of models.

Theorem 6.4. *Let $(h_{g_1}, \ldots, h_{g_k})$ be a collection of models h_{g_i} which witness meaningful multiplicity with respect to h^* on groups $\mathcal{G} = g_1, \ldots, g_k$ at level ϵ. Let f be the best linear combination of models weighted by group membership on distribution $\mathcal{D}$, as in Equation 1 in Theorem 5.2. Then, f is multi-accurate with respect to $\mathcal{G}$.*

However, this will, for instance, require calling at least one squared error oracle call per group in $\mathcal{G}$ on the distribution constrained to that group, which may be prohibitive as $\mathcal{G}$ increases in complexity. Thankfully, we have alternative methods to building multiaccurate and orthogonal predictors, which do not have such expensive training procedures, such as those from (Kim et al., 2019) or (Globus-Harris et al., 2023b).

7. Meaningful Multiplicity for Constraint-Based Fairness Notions

We will now consider what we can say about fairness metrics for classification problems. Until now, we have considered regression models h mapping from features $\mathcal{X}$ to real values $\mathcal{Y}$. Now, we will consider what we can say about constraint-based fairness notions for classifiers $c \in \mathcal{C}$, where $c : \mathcal{X} \to \{0, 1\}$. In this setting, we are interested in notions of fairness which may be in direct tension with model accuracy, such as equalizing false-negative rates or false-positive rates across groups.

Formally, we will consider group fairness metrics which guarantee that some fairness metric ρ on the classifier c when evaluated on any group g is close to that of the metric evaluated on the overall population.

Definition 7.1 (Group-Constrained Fairness Metrics)**.** *We say that a classifier $c : \mathcal{X} \to \{0, 1\}$ satisfies γ-Fairness for metric $\phi : \mathcal{C} \to [0, 1]$ with respect to fairness notion $\bullet \in \{FN, FP, E, SP\}$ with respect to distribution $\mathcal{D}$ and groups $\mathcal{G}$ if for all $g \in \mathcal{G}$,*

$$\phi_g^\bullet(c) = w_g^\bullet |\rho_g^\bullet(c) - \rho^\bullet(c)| \leq \gamma,$$

where $\rho_g^\bullet(c)$ is some fairness notion $\bullet$ evaluated on group g, $\rho(c)$ is the group-fairness notion $\bullet$ evaluated on the overall distribution, and w_g is a slack parameter for the group weight. We will write $\phi^\bullet(c) = \max_{g\in\mathcal{G}} \phi_g^\bullet(c)$.

Remark 7.2. *The methods presented here are agnostic to which particular metric is considered among equalizing false negative rate ("FN"), equalizing false positive rates ("FP"), equalizing error rates ("E"), and statistical parity ("SP"). For completeness, the formal definitions of all are included in Appendix G in addition to the more general form presented here. In the remainder of the main body of the paper, we will simply use ϕ to refer to an arbitrary fairness metric, and omit the $\bullet$ superscript when referring to the subterms, $w_g^\bullet, \rho_g^\bullet$, and $\rho^\bullet$, with the understanding that the techniques may be instantiated for all four metrics and/or some combination thereof.*

The fairness metrics defined in 7.1 may be in direct tension with minimizing model error, which in the classification setting we will take to be zero-one loss.

Definition 7.3 (Zero-one loss)**.**

$$\begin{aligned} err(c) &= \mathbb{E}_{(x,y)\sim\mathcal{D}}\left[\ell_{01}(c(x), y)\right] \\ &= \mathbb{E}_{(x,y)\sim\mathcal{D}}\left[\mathbb{1}[c(x) \neq y]\right]. \end{aligned}$$

The goal, then, is to find a model which lies on the *Pareto front* of the fairness-accuracy tradeoff. Mapping from the original language of model multiplicity from, e.g., (Black et al., 2022), the observation there is that this Pareto front may not require substantial tradeoffs after all: that there may be, in fact, quite fair models which are not substantially dominated in terms of accuracy compared to the error-optimal model.

Definition 7.4. *Consider a collection of classifiers $\mathcal{M}$. We say that $c \in \mathcal{M}$ belongs to the fairness-accuracy Pareto front $\mathcal{P}(\mathcal{M})$ if there is no other $c' \in \mathcal{M}$ such that $err(c') \leq err(c)$ and $\mu(c') \leq \mu(c)$ with at least one inequality strict.*

How might one actually go about finding a model that lies on this Pareto front within some hypothesis class of classifiers $\mathcal{C}$ on a distribution $\mathcal{D}$? Observe that one method would be to solve the following constrained optimization problem over $\mathcal{C}$ for every value of $\gamma \in [0, 1]$:

$$\begin{aligned} \min_{c\in\mathcal{C}} \quad & \text{err}(c), \\ \text{s.t. } \forall g \in \mathcal{G} \quad & \phi(c) \leq \gamma. \end{aligned}$$

Or, if we can instead *randomize* over classifiers, as we will subsequently, we could get a stronger collection of models by running this optimization over the simplex $\Delta\mathcal{C}$.

Definition 7.5 (Optimal Group-Constrained γ-Fair Model)**.**

$$\begin{aligned} \min_{c\in\Delta\mathcal{C}} \quad & err(c), \\ \textit{s.t. } \forall g \in \mathcal{G} \quad & \phi(c) = w_g \left|\rho_g(c) - \rho(c)\right| \leq \gamma \end{aligned}$$

where $w_g, \rho_g(c)$, and $\rho(c)$ are defined as in Definition 7.1.

The key observation we will rely on in the subsequent results from (Globus-Harris et al., 2023a) is that if we knew what the true label probabilities $f^*(x) = \mathbb{E}[y|x]$ were, then for any fairness notion we can rewrite the above constrained optimization problem in terms of this. For instance,

$$\begin{aligned} \text{err}(c) &= \mathbb{E}_{\mathcal{D}}\left[\mathbb{1}[c(x) \neq y]\right] \\ &= \mathbb{E}_{\mathcal{D}}[f^*(x)\ell_{01}(c(x), 1) + (1 - f^*(x))\ell_{01}(c(x), 0)], \end{aligned}$$

and similarly we can write $\phi(x)$ as an expectation over products written in terms of the zero-one loss ℓ_{01}, group indicator function g, and some scalar weighting that will be over the group g: see Lemmas G.3, G.6, G.8, and G.10 for details.

Remark 7.6. *We will write c^* to mean the optimal fairness-constrained policy which uses the Bayes optimal predictor f^* to compute the conditional label distribution and then finds the solution to the optimization problem defined in Definition 7.5 using f^*.*

Note that if we had some arbitrary predictor of label probabilities f, we could consider the optimal postprocessing over $\mathcal{H}$ if we take f to be the true label probabilities.

Definition 7.7. *Given predictor $f : \mathcal{X} \to [0, 1]$, let c_f be the solution to the optimization in Definition 7.5 with f substituted for f^*, i.e., using $f(x)$ as a proxy for $\mathbb{E}[y|x]$ in evaluating $err(c)$ and $\phi(c)$. The explicit substitutions needed for each fairness notion $\bullet$ are given in Appendix G.*

The problem here, of course, is that any f we naively train to reduce squared error over some hypothesis class will not in fact be a sufficiently good predictor of individual label probabilities to output a solution to this optimization problem that is competitive with c^*, the optimal postprocessing of our Bayes predictor f^*. However, in (Globus-Harris et al., 2023a), the authors show that a suitable *multicalibrated* predictor suffices as a proxy for the Bayes optimal predictor in terms of the downstream fairness objective. The formal definitions and conditions of multicalibration needed here are formalized in Appendix G in Definitions G.11, G.12, and G.13, and the theorem statement is as follows:

Theorem 7.8 ((Globus-Harris et al., 2023a))**.** *Let f be α-approximately multicalibrated in expectation with respect to $\mathcal{G}, \mathcal{C}$, and $\mathcal{G} \times \mathcal{C} = \{g(x) \cdot c(x) | g \in \mathcal{G}, c \in \mathcal{C}\}$, and be α-approximately jointly multicalibrated in expectation with respect to $\mathcal{B}(C)$. Let $\bar{c}$ be the optimal postprocessing for fairness metric $\bullet$ evaluated according to the projected gradient descent algorithm in (Globus-Harris et al., 2022).*[1].

[1]Corresponding to, in that work, Algorithm 1 for FP fairness, and the corresponding reformulations in algorithms 2 through 4 for other fairness notions included in Appendix G.

Then,

$$err(\bar{c}) \leq err(c^*) + \alpha(5 + 2\sqrt{1/\alpha}) + 2\sqrt{\alpha},$$
$$\phi(\bar{c}) \leq \phi(c^*) + w_g\alpha.$$

In essence, because the real-valued predictor f, which is approximating the conditional label distribution $\mathbb{E}[y|x]$ is multicalibrated with respect to thresholdings that are group-dependent, this gives us a way to efficiently find mixtures of classifiers in $\mathcal{C}$ which lie on the fairness-accuracy Pareto front.

What does this mean for multiplicity? Note that any desirable multiplicitous model, for instance, one in the Rashomon set of $\mathcal{H}$ which still had good fairness guarantees, would lie on the Pareto front of $\mathcal{C}$. The challenge previously was that finding such a model is hard, in terms of searching the Rashomon set. Even postprocessing promising models, prior to the work of (Globus-Harris et al., 2023a), was challenging when groups overlapped. Here, however, we now have a method to efficiently a predictor which will be competitive with any classifier we might have found via directly searching the Rashomon set, by first learning a multicalibrated predictor f and then postprocessing it to learn a competitive classifier via the projected gradient descent process described in (Globus-Harris et al., 2023a).

Corollary 7.9. *Let f satisfy the multicalibration conditions of Theorem 7.8, and for any $\gamma \in [0, 1]$, let $\bar{c}_\gamma$ denote the output of the optimal postprocessing procedure in 7.8 applied to f with fairness parameter γ. Then, $\bar{c}_\gamma$ lies on the Pareto front $\mathcal{P}(\Delta C)$, up to approximation $\alpha(5 + 2\sqrt{1/\alpha} + 2\sqrt{\alpha}$ for error and $w_g\alpha$ for the group fairness metric. In particular, it will be competitive with any model in the Rashamon set $\mathcal{R}(\mathcal{C})$.*

Proof. This follows directly from Theorem 7.8 and the observation that $\mathcal{C} \subset \Delta\mathcal{C}$. □

References

Aaronson, S. The complexity of agreement. In *Proceedings of the thirty-seventh annual ACM symposium on Theory of computing*, pp. 634–643, 2005.

Aumann, R. J. Agreeing to Disagree. *The Annals of Statistics*, 4(6):1236 – 1239, 1976. doi: 10.1214/aos/1176343654. URL https://doi.org/10.1214/aos/1176343654.

Behzad, T., Casacuberta, S., Diana, E. R., and Tolbert, A. W. Reconciling predictive multiplicity in practice. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*, pp. 3350–3369, 2025.

Black, E., Raghavan, M., and Barocas, S. Model multiplicity: Opportunities, concerns, and solutions. In *Proceedings of the 2022 ACM conference on fairness, accountability, and transparency*, pp. 850–863, 2022.

Black, E., Koepke, L., Kim, P., Barocas, S., and Hsu, M. The legal duty to search for less discriminatory algorithms. *arXiv preprint arXiv:2406.06817*, 2024.

Breiman, L. Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical science*, 16(3):199–231, 2001.

Collina, N., Goel, S., Gupta, V., and Roth, A. Tractable agreement protocols. In *Proceedings of the 57th Annual ACM Symposium on Theory of Computing*, pp. 1532–1543, 2025.

Collina, N., Globus-Harris, I., Goel, S., Gupta, V., Roth, A., and Shi, M. Collaborative prediction: Tractable information aggregation via agreement. In *Proceedings of the 2026 Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pp. 4712–4798. SIAM, 2026.

Dai, G., Ravishankar, P., Yuan, R., Black, E., and Neill, D. B. Be intentional about fairness!: Fairness, size, and multiplicity in the rashomon set. In *Proceedings of the 5th ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, pp. 42–73, 2025.

Globus-Harris, I., Kearns, M., and Roth, A. An algorithmic framework for bias bounties. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pp. 1106–1124, 2022.

Globus-Harris, I., Gupta, V., Jung, C., Kearns, M., Morgenstern, J., and Roth, A. Multicalibrated regression for downstream fairness. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 259–286, 2023a.

Globus-Harris, I., Harrison, D., Kearns, M., Roth, A., and Sorrell, J. Multicalibration as boosting for regression. *arXiv preprint arXiv:2301.13767*, 2023b.

Gopalan, P., Kalai, A. T., Reingold, O., Sharan, V., and Wieder, U. Omnipredictors. In *13th Innovations in Theoretical Computer Science Conference (ITCS 2022)*, pp. 79–1. Schloss Dagstuhl–Leibniz-Zentrum für Informatik, 2022.

Hays, C., Laufer, B., Barocas, S., and Raghavan, M. Statistical guarantees in the search for less discriminatory algorithms. *arXiv preprint arXiv:2512.23943*, 2025.

Hébert-Johnson, U., Kim, M., Reingold, O., and Rothblum, G. Multicalibration: Calibration for the (computationally-identifiable) masses. In *International Conference on Machine Learning*, pp. 1939–1948. PMLR, 2018.

Kearns, M., Roth, A., and Ryu, E. Networked information aggregation via machine learning. In *Proceedings of the 2026 Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pp. 4799–4845. SIAM, 2026.

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

Laufer, B., Raghavan, M., and Barocas, S. What constitutes a less discriminatory algorithm? In *Proceedings of the 2025 Symposium on Computer Science and Law*, pp. 136–151, 2025.

Roth, A., Tolbert, A., and Weinstein, S. Reconciling individual probability forecasts. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, pp. 101–110, 2023.

Rudin, C., Zhong, C., Semenova, L., Seltzer, M., Parr, R., Liu, J., Katta, S., Donnelly, J., Chen, H., and Boner, Z. Amazing things come from having many good models. *arXiv preprint arXiv:2407.04846*, 2024.

Xin, R., Zhong, C., Chen, Z., Takagi, T., Seltzer, M., and Rudin, C. Exploring the whole rashomon set of sparse decision trees. *Advances in neural information processing systems*, 35:14071–14084, 2022.

A. Formal Preliminaries and Notation

Definition A.1 (Model Loss)**.** *Given a model $f : \mathcal{X} \to \mathcal{Y}$, we will write $L(\mathcal{D}, f)$ to be the loss of a model in expectation over the distribution:*

$$L(\mathcal{D}, f) = \mathbb{E}_{(x,y)\sim\mathcal{D}}[\ell(f(x), y)].$$

We will write $L(D, f)$ to denote the corresponding empirical measure over sample $D \sim \mathcal{D}^n$, viewing D as a uniform distribution over its elements.

In particular we will often be interested in the squared error of a predictor f:

Definition A.2 (Squared Error)**.** *Given a model $f : \mathcal{X} \to [0, 1]$, we will write $\ell_2 : \mathcal{X} \to [0, 1]$ to be the squared error of the model f:*

$$\ell_2(f(x), y) = (f(x) - y)^2$$
$$L_2(\mathcal{D}, f) = \mathbb{E}_{(x,y)\sim\mathcal{D}}[(f(x) - y)^2]$$

We will use the Bayes Optimal model, defined with respect to an arbitrary loss function, as a gold standard comparison to compare the optimality of any given model to:

Definition A.3 (Bayes Optimal Model)**.** *A model f^* is the* Bayes Optimal Model *with respect to pointwise loss function $\ell : \mathcal{X} \to \mathcal{Y}$ and distribution $\mathcal{D}$ if it satisfies the following*

$$f(x) \in \arg\min_{f:\mathcal{X}\to\mathcal{Y}} \mathbb{E}_{(x,y)\sim\mathcal{D}}[\ell(f(x), y)].$$

We will also consider the loss minimizing model when models are constrained to some collection of hypotheses $\mathcal{H}$, which we will usually write as h^*.

Definition A.4 (Risk minimizer with respect to $\mathcal{H}$)**.** *Let $\mathcal{H}$ be a class of models $h : \mathcal{X} \to \mathcal{Y}$. Then h^* is a risk minimizer with respect to $\mathcal{H}$ and loss function ℓ if*

$$h^* \in \arg\min_{h\in\mathcal{H}} \mathbb{E}_{(x,y)\sim\mathcal{D}}[\ell(h(x), y)].$$

Definition A.5 (Group Function)**.** *Let $g : \mathcal{X} \to [0, 1]$ denote a (demographic, e.g.) group that is identifiable by the features. We will say that point x "belongs to " group g if $g(x) = 1$ and will denote the weight of g as $\mu_g = \mathbb{P}[g(x) = 1]$.*

Definition A.6 (Group error)**.** *Given a distribution $\mathcal{D}$ over $\mathcal{X}\times\mathcal{Y}$, a model $f : \mathcal{X} \to \mathcal{Y}$, and a group indicator $g : \mathcal{X} \to \{0, 1\}$, the* group error *of f on g is*

$$\mathbb{E}_g\big[\ell(f(x), y)\big] := \mathbb{E}_{(x,y)\sim\mathcal{D}}\big[\ell(f(x), y) \mid g(x) = 1\big].$$

Definition A.7 (Rashomon set)**.** *Given a hypothesis class $\mathcal{H}$, distribution $\mathcal{D}$, and $\epsilon \in (0, 1/2]$, the ϵ-Rashomon set is the collection of models in $\mathcal{H}$ with error ϵ-close to the risk minimizer h^* : $\mathcal{R}(\mathcal{H}) = \{h \in \mathcal{H} : \mathbb{E}_{(x,y)\sim\mathcal{D}}[\ell(h(x), y)] \leq \mathbb{E}_{(x,y)\sim\mathcal{D}}[\ell(h^*(x), y)] + \epsilon$.*

B. Deferred Proofs from Section 3

Lemma B.1 (Meaningful Multiplicity implies Disagreement)**.** *Let $h^* = \arg\min_{h\in\mathcal{H}} \mathbb{E}\big[(h(x) - y)^2\big]$ be the squared-error minimizing model in $\mathcal{H}$. Suppose there exists a group g with density μ_g and a model $h_g \in \mathcal{H}$ such that*

$$\mathbb{E}_g\big[(h_g(x) - y)^2\big] \leq \mathbb{E}_g\big[(h^*(x) - y)^2\big] - \varepsilon.$$

Then the expected squared disagreement between h_g and h^* on g is at least $\mathbb{E}_g\big[(h_g(x) - h^*(x))^2\big] \geq \frac{\varepsilon^2}{4}$, and overall $\mathbb{E}\big[(h_g(x) - h^*(x))^2\big] \geq \mu_g \cdot \frac{\varepsilon^2}{4}$.

Proof. Note that for any $a, b, y \in [0, 1]$ we have

$$\begin{aligned}
|(a - y)^2 - (b - y)^2| &= |(a^2 - 2ay + y^2) - (b^2 - 2by + y^2)| \\
&= |(a^2 - b^2) + (2by - 2ay)| \\
&= |(a - b)(a + b) + 2y(b - a)| \\
&= |b - a||a + b - 2y| \\
&\leq 2|a - b|,
\end{aligned}$$

so

$$|a-b| \geq \frac{1}{2}|(a-y)^2 - (b-y)^2|.$$

Squaring this yields the inequality

$$(a-b)^2 \geq \frac{1}{4}((a-y)^2 - (b-y)^2)^2. \tag{2}$$

Let $h^* = \arg\min_{h \in \mathcal{H}} \mathbb{E}\big[(h(x)-y)^2\big]$ be the squared-error minimizing model in $\mathcal{H}$. Let g be a group with density μ_g, and let $h_g \in \mathcal{H}$ be a model such that its error on g is better than that of h^* by at least ε. I.e.,

$$\mathbb{E}_g\big[(h_g(x)-y)^2\big] \leq \mathbb{E}_g\big[(h^*(x)-y)^2\big] - \varepsilon.$$

Note that since $h^*(x), h_g(x)$, and y are all in the range $[0,1]$, we can apply Equation 2 and say that for any value of x,

$$(h_g(x) - h^*(x))^2 \geq \frac{1}{4}((h_g - y)^2 - (h^* - y)^2)^2.$$

Since this applies pointwise, it also holds in expectation over group g. Then, applying Jensen's inequality, we find that

$$\begin{aligned}
\mathbb{E}_g\big[(h_g - h^*)^2\big] &\geq \frac{1}{4}\mathbb{E}_g\Big[\big((h_g-y)^2 - (h^*-y)^2\big)^2\Big] \\
&\geq \frac{1}{4}\Big(\mathbb{E}_g\big[(h^*-y)^2 - (h_g-y)^2\big]\Big)^2 \\
&= \frac{1}{4}\Big(\mathbb{E}_g\big[(h^*-y)^2\big] - \mathbb{E}_g\big[(h_g-y)^2\big]\Big)^2 \\
&= \frac{1}{4}\varepsilon^2
\end{aligned}$$

which gives the group bound. Multiplying by μ_g yields the overall bound:

$$\begin{aligned}
\mathbb{E}\big[(h_g(x) - h^*(x))^2\big] &= \mathbb{P}(g(x)=1) \cdot \mathbb{E}_g\big[(h_g(x)-h^*(x))^2\big] + \mathbb{P}(g(x)=0) \cdot \mathbb{E}_{\neg g}\big[(h_g(x)-h^*(x))^2\big] \\
&\geq \mu_g \cdot \frac{\varepsilon^2}{4}.
\end{aligned}$$

□

Lemma B.2. *A model h which witnesses multiplicity with respect to h^* on group g with density $\mu_g < 1/2$ may not belong to the ϵ-Rashomon set of h^* for any $\varepsilon \in (0,1)$ such that $\varepsilon < 1 - 2\mu_g$.*

Proof. Consider a binary prediction problem for labels $\mathcal{Y} \in \{0,1\}$. Let $\mathcal{H}$ be a hypothesis class with only two models in it: h_1, which predicts perfectly for points not in g and predicts maximally incorrectly for points in g, and a model h_2 which predicts perfectly for points in g and incorrectly for points not in g. I.e.,

$$h_1(x) = \begin{cases} y & g(x) = 0, \\ 1-y & g(x) = 1. \end{cases}$$

$$h_2(x) = \begin{cases} y & g(x) = 1, \\ 1-y & g(x) = 0. \end{cases}$$

Then, if g has density μ_g, since $y \in \{0,1\}$, each mistake incurs a cost in squared error of 1 and so the overall expected squared error of h_1 will be μ_g and the overall expected squared error of h_2 will be $1 - \mu_g$. Hence, the gap in their squared error will be

$$\begin{aligned}\mathbb{E}\big[(h_2(x)-y)^2\big]-\mathbb{E}\big[(h_1(x)-y)^2\big] &= (1-\mu_g)-\mu_g\\ &= 1-2\mu_g\\ &> \epsilon\end{aligned}$$

Hence, h_1 is the error-minimizing model h^* for the hypothesis class, and h_2 will not be in its ε-Rashomon set. However, h_2 does witness meaningful model multiplicity on group g, since its error on g is 0 while the error of h_1 on g is 1. □

C. Deferred Proofs from Section 4

Lemma C.1 (Squared Error Decomposition (e.g. (Kearns et al., 2026))). *For any p and q,*

$$\begin{aligned}&\mathbb{E}[(p-y)^2]-\mathbb{E}[(q-y)^2]\\ &\quad= 2\mathbb{E}[(p-q)(q-y)]+\mathbb{E}[(p-q)^2]\\ &\quad= \mathbb{E}[(p-q)^2]+2\mathbb{E}[p(q-y)]-2\mathbb{E}[q(q-y)].\end{aligned}$$

Definition C.2 (Convex hypothesis class). *We will say that $\mathcal{H}$ is a convex hypothesis class if for any $h_1, h_2 \in \mathcal{H}$, the class $\mathcal{H}$ is closed under their linear combination: $h' \in \mathcal{H}$, where $h'(x) = \lambda h_1(x) + (1-\lambda)h_2(x)$ for any $\lambda \in [0,1]$.*

Lemma C.3 (Squared Error Decomposition (e.g. (Kearns et al., 2026))). *For any p and q,*

$$\begin{aligned}&\mathbb{E}[(p-y)^2]-\mathbb{E}[(q-y)^2]\\ &\quad= 2\mathbb{E}[(p-q)(q-y)]+\mathbb{E}[(p-q)^2]\\ &\quad= \mathbb{E}[(p-q)^2]+2\mathbb{E}[p(q-y)]-2\mathbb{E}[q(q-y)].\end{aligned}$$

Proof. The proof proceeds through straightforward algebraic manipulation:

$$\begin{aligned}\mathbb{E}[(p-y)^2]-\mathbb{E}[(q-y)^2] &= \mathbb{E}[p^2-2py+y^2-(q^2-2qy+y^2)]\\ &= \mathbb{E}[(p^2-2pq+q^2)+2pq-2q^2-2py+2qy]\\ &= \mathbb{E}[(p-q)^2]+2\mathbb{E}[q(p-q)-y(p-q)]\\ &= \mathbb{E}[(p-q)^2]+2\mathbb{E}[(p-q)(q-y)]\\ &= \mathbb{E}[(p-q)^2]+2\mathbb{E}[p(q-y)]-2\mathbb{E}[q(q-y)].\end{aligned}$$

□

Lemma C.4 ((Case of disjoint groups). *Let h^* be the error-minimizing model in convex hypothesis class $\mathcal{H}$ on distribution $\mathcal{D}$ and let f^* be the Bayes optimal model. Let $g_1, \ldots, g_k$ be disjoint groups, where g_i has probability mass μ_i. Let $h_{g_i} \in \arg\min_{h\in\mathcal{H}} \mathbb{E}_{g_i}[\ell(h(x),y)]$ be the error-minimizing model for group g_i, and let $\delta_i = \mathbb{E}_{g_i}[(h^*(x)-h_{g_i}(x))^2]$ be the expected squared disagreement of h^* and h_{g_i} on group g_i. Then*

$$\mathbb{E}[(h^*(x)-y)^2]-\mathbb{E}[(f^*(x)-y)^2] \geq \sum_{i\in[k]} \mu_i\delta_i.$$

Proof. Consider the ensemble model f which uses h_{g_i} for group g_i and otherwise uses the general minimizer h^*:

$$f(x) = \begin{cases} h_{g_1}(x) & g_1(x)=1,\\ \vdots & \\ h_{g_k}(x) & g_k(x)=1,\\ h^*(x) & \text{else.}\end{cases}$$

By construction,

$$\begin{aligned}\mathbb{E}_{g_i}[(h^*(x)-y)^2]-\mathbb{E}_{g_i}[(f(x)-y)^2] &= \mathbb{E}_{g_i}[(h^*(x)-y)^2]-\mathbb{E}_{g_i}[(h_{g_i}(x)-y)^2] \\ &= \mathbb{E}_{g_i}[(h^*(x)-h_{g_i}(x))^2]+2\mathbb{E}[(h^*(x)-h_{g_i}(x))(h_{g_i}(x)-y)] \qquad \text{(Lemma C.1)} \\ &\geq \delta_i + 2\mathbb{E}[(h^*(x)-h_{g_i}(x))(h_{g_i}(x)-y)].\end{aligned}$$

Let $h'(x) = (1-\lambda)h_{g_i}(x) + \lambda h^*(x)$ for $\lambda \in [0,1]$. Since $\mathcal{H}$ is convex, $h' \in \mathcal{H}$. Note that since h_{g_i} is by construction the error minimizer of $\mathcal{H}$ constrained to group g, its squared error on this subgroup must be lower than that of h'. I.e.,

$$\begin{aligned}\mathbb{E}_{g_i}[(h_{g_i}(x)-y)^2] &\leq \mathbb{E}_{g_i}[(h'(x)-y)^2] \\ &= \mathbb{E}_{g_i}[((1-\lambda)h_{g_i}(x)+\lambda h^*(x)-y)^2].\end{aligned}$$

Taking the derivative of both sides with respect to λ and then setting λ equal to 0 gives us a lower bound on the expected correlation between h^* and h_{g_i} on the residuals of h_{g_i} on the subgroup:

$$0 \leq 2\mathbb{E}_{g_i}[(h^*-h_{g_i})(h_{g_i}-y)].$$

Hence,

$$\begin{aligned}\mathbb{E}_{g_i}[(h^*(x)-y)^2]-\mathbb{E}_{g_i}[(h_{g_i}(x)-y)^2] &= \mathbb{E}_{g_i}[(h^*(x)-h_{g_i})^2]+2\mathbb{E}_{g_i}[(h^*(x)-h_{g_i}(x))(h_{g_i}-y)] \\ &\geq \mathbb{E}_{g_i}[(h^*(x)-h_{g_i})^2] \\ &= \delta_i\end{aligned}$$

Thus,

$$\mathbb{E}_{g_i}[(h^*(x)-y)^2]-\mathbb{E}_{g_i}[(f(x)-y)^2] \geq \delta_i.$$

Summing over the groups, this lets us bound the distance between h^* and the Bayes optimal predictor f^*, since f cannot have lower squared error than the Bayes optimal predictor:

$$\begin{aligned}\mathbb{E}_{\mathcal{D}}[(h^*(x)-y)^2]-\mathbb{E}_{\mathcal{D}}[(f^*(x)-y)^2] &\geq \mathbb{E}_{\mathcal{D}}[(h^*(x)-y)^2]-\mathbb{E}_{\mathcal{D}}[(f(x)-y)^2] \\ &= \sum_{i\in[k]} \mathbb{P}(x \in g_i)\delta_i + \mathbb{P}(x \neq \cup_{i\in[k]} g_i)\cdot 0 \\ &= \sum_{i\in[k]} \mu_i \delta_i\end{aligned}$$

□

Lemma C.5 ((Overlapping groups)). *Let h^* be the error-minimizing model in convex hypothesis class $\mathcal{H}$ on distribution $\mathcal{D}$ and let f^* be the Bayes optimal model. Let $g_1, \ldots, g_k$ be a collection of groups such that $m = \max_{x\in\mathcal{X}} |\mathcal{S}(x)|$ is the largest number of groups any point in $\mathcal{X}$ belongs to and where g_i has probability mass μ_i. Let $h_{g_i} \in \arg\min_{h\in\mathcal{H}} \mathbb{E}_{g_i}[\ell(h(x),y)]$ be the error-minimizing model for group g_i, and let $\delta_i = \mathbb{E}_{g_i}[(h^*(x)-h_{g_i}(x))^2]$ be the expected squared disagreement of h^* and h_{g_i} on group g_i. Then*

$$\mathbb{E}[(h^*(x)-y)^2]-\mathbb{E}[(f^*(x)-y)^2] \geq \frac{1}{m}\sum_{i\in[k]} \mu_i\delta_i.$$

Proof. As in the proof of Lemma 4.1, we will construct a function which "routes" each point in $\mathcal{X}$ to the error minimizer of each group in order to bound the error of h^* away from optimal. However, here we will have to be more careful, due to the fact that the groups overlap. First, we will develop some notation to identify all of the overlapping regions of the space.

For any $x \in \mathcal{X}$, let $\mathcal{S} : \mathcal{X} \to 2^{[k]}$ be the function mapping points x to the set of groups in $g_1, \ldots, g_k$ which contain point x: $\mathcal{S}(x) = \{i \in [k] : g_i(x) = 1\}$. Let $A_S = \{x \in \mathcal{X} : \mathcal{S}(x) = S\}$ be the set of all points in $\mathcal{X}$ which belong exactly to the collection of groups S. Then, enumerating over all $S \in 2^{[k]}$ induces a disjoint partitioning of $\mathcal{X}$: $\mathcal{X} = \cup_{S\in 2^{[k]}} A_S$.

We will now learn which candidate model for each non-empty subgroup A_S is best, picking between h^* and the models h_{g_i} for $i \in S$,

$$j^*(S) \in \arg\min_{j \in S \cup \{*\}} \mathbb{E}_{A_S}[(h_{g_j}(x) - y)^2],$$

where to avoid too much more notational gymnastics, $h_{g_*} := h^*$. We then define f to be the model which routes to the best model in A_S for each disjoint A_S:

$$f(x) = \begin{cases} h_{j^*(\mathcal{S}(x))}(x) & \mathcal{S}(x) \neq \emptyset, \\ h^*(x) & \text{else.} \end{cases}$$

For each group g_i, let $\Delta_{i,S}$ be the expected improvement in squared error of h_{g_i} on the subgroup A_S and let μ_S be the density of S in distribution $\mathcal{D}$:

$$\begin{aligned} \Delta_{i,S} &= \mathbb{E}_{A_S}[(h^*(x) - y)^2 - (h_{g_i} - y)^2)] \\ &= \mathbb{E}_{\mathcal{D}}[(h^*(x) - y)^2 - (h_{g_i} - y)^2) | \mathcal{S}(x) = S], \end{aligned}$$

$$\mu_S = \mathbb{P}_{(x,y)\sim\mathcal{D}}[\mathcal{S}(x) = S].$$

Since f can also choose to route to h^* on S in addition to $h_{g_1}, \ldots, h_{g_k}$,the improvement of f on S will be

$$\Delta_S = \max\left(0, \max_{i \in S} \Delta_{i,S}\right).$$

Since the maximum over a set is always at least its average, we can then bound Δ_S by the average over $\Delta_{i,S}$:

$$\Delta_S \geq \frac{1}{|S|} \sum_{i \in S} \Delta_{i,S}.$$

Then,

$$\begin{aligned} \mathbb{E}[(h^*(x) - y)^2 - (f(x) - y)^2] &= \sum_{S \in 2^{[k]} : A_S \neq \emptyset} \mu_S \Delta_S \\ &\geq \sum_{S \in 2^{[k]} : A_S \neq \emptyset} \frac{\mu_S}{|S|} \sum_{i \in S} \Delta_{i,S}. \end{aligned}$$

Recall $m = \max_{x \in \mathcal{X}} |\mathcal{S}(x)|$ is the largest number of groups any point in $\mathcal{X}$ belongs to. Then, for all S, $|S| \leq r$, and

$$\mathbb{E}[(h^*(x) - y)^2 - (f(x) - y)^2] \geq \frac{1}{r} \sum_{S \in 2^{[k]} : A_S \neq \emptyset} \mu_S \sum_{i \in S} \Delta_{i,S}. \tag{3}$$

Note that each group g_i can be expressed as a disjoint union of some collection of subgroups A_S,

$$\{x : g_i(x) = 1\} = \{x \in A_S : i \in S\}_{S \in 2^{[k]}},$$

and that since $\mathcal{H}$ is convex and Lemma C.1 still applies, it is the case that for each $i \in [k]$,

$$\mathbb{E}_{g_i}[(h^*(x) - y)^2 - (h_{g_i}(x) - y)^2] \geq \delta_i.$$

Since f can only improve on g_i compared to h_{g_i}, it follows that

$$\sum_{i \in S} \mu_S \Delta_{i,S} \geq \mu_i \delta_i.$$

Combining this with Equation 3 and noting that the Bayes optimal predictor can only outperform f gives

$$\begin{aligned}\mathbb{E}[(h^*(x)-y)^2-(f^*(x)-y)^2] &\geq \mathbb{E}[(h^*(x)-y)^2-(f(x)-y)^2] \\ &\geq \frac{1}{r}\sum_{S\in 2^{[k]}:A_S\neq\emptyset}\mu_S\sum_{i\in S}\Delta_{i,S} \\ &\geq \frac{1}{r}\sum_{i=1}^{k}\mu_i\delta_i.\end{aligned}$$

□

D. Algorithms and Deferred Proofs from Section 5.1 for Decision List Ensembles

For ease of reference, this section describes the algorithms and relevant proofs as stated in (Globus-Harris et al., 2022) for constructing an ensemble of models which guarantees that groupwise errors decrease as models are appended. The two algorithms, Algorithms 1 and 2, are subroutine to the second Algorithm 3. Tuples of groups g_i and models h_i are fed in iteratively, and the overall models' error on g_i is compared to that of h_i using the 2 subroutine. If h_i improves, it is appended to the model using 1, and then the algorithm recursively checks if any other group errors have increased, in which case the model repairs these by appending an additional tuple to the front of the decision list. Since overall error drops any time the submissions are accepted, the process terminates, allowing standard generalization bounds.

The following statements are directly from (Globus-Harris et al., 2022), which is an algorithmic framework which is agnostic to the loss function used, with small changes for expositional clarity.

Definition D.1 (Model Loss on Subgroups)**.** *We write $L(\mathcal{D}, f, g)$ to denote the loss on f conditional on membership in g:*

$$L(\mathcal{D}, f, g) = \mathbb{E}_{(x,y)\sim\mathcal{D}}[\ell(f(x), y)|g(x) = 1].$$

Given a dataset D, we write $L(D, f, g)$ to denote the corresponding empirical loss on D.

Definition D.2. *A group indicator function $g : \mathcal{X} \to \{0, 1\}$ together with a model $h : \mathcal{X} \to \mathcal{Y}$ form a (μ, Δ)-certificate of sub-optimality for a model f under distribution $\mathcal{D}$ if:*

1. *Group g has probability mass at least μ under $\mathcal{D}$: $\mu_{\mathcal{D}}(g) \geq \mu$, and*
2. *h improves on the performance of f on group g by at least Δ: $L(\mathcal{D}, f, g) \geq L(\mathcal{D}, h, g) + \Delta$*

We say that (g, h) form a certificate of sub-optimality for f if they form a (μ, Δ)-certificate of optimality for f for any constants $\mu, \Delta > 0$.

Theorem D.3 (Theorems 12 and 14 in (Globus-Harris et al., 2022))**.** *Fix any $\gamma, \delta > 0$. Let $\mathcal{D} \in \Delta(\mathcal{X} \times \mathcal{Y})$ be any distribution over labelled examples, and let $D \sim \mathcal{D}^n$ be a holdout dataset consisting of n i.i.d. samples from $\mathcal{D}$. Suppose:*

$$n \geq \frac{64\ln\left(\frac{2(U+\frac{8}{\gamma^3})}{\delta'}\right)}{\gamma^3}.$$

Then for any (possibly adaptive) process generating a sequence of at most U submissions $\{(g_i, h_i)\}_{i=1}^{U}$, with probability at least $1 - \delta$, we have that MonotoneFalsifyAndUpdate$(\gamma, D, \cdots)$ satisfies:

1. *If (g_i, h_i) is rejected, then (g_i, h_i) is not a (μ, Δ)-certificate of sub-optimality for f_t, where f_t is the current model at the time of submission i, for any μ, Δ such that $\mu \cdot \Delta \geq \gamma$.*
2. *If (g_i, h_i) is accepted, then (g_i, h_i) is a (μ, Δ)-certificate of sub-optimality for f_t, where f_t is the current model at the time of submission i, for some μ, Δ such that $\mu \cdot \Delta \geq \frac{\gamma}{2}$. Moreover, the new model f_{t+1} output satisfies $L(\mathcal{D}, f_{t+1}, g_i) = L(\mathcal{D}, h_i, g_i)$ and $L(\mathcal{D}, f_{t+1}) \leq L(\mathcal{D}, f_t) - \frac{\gamma}{2}$.*
3. *MonotoneFalsifyAndUpdate does not halt before receiving all U submissions.*

4. *MonotoneFalsifyAndUpdate satisfies group error monotonicity: Consider any model f_t that is output, and any group $g_j \in \mathcal{G}(f_t)$. Then:*

$$L(\mathcal{D}, f_t, g_j) \leq \min_{\ell<t} L(\mathcal{D}, f_\ell, g_j) + \frac{\gamma}{\mu_{g_j}}$$

Theorem 5.1 (Mitigating Multiplicity with Bias Bounties)**.** *Let $(g_1, h_{g_1}), \ldots, (g_k, h_{g_k})$ be a collection of group indicators g_i and models h_{g_i}. Let f be the model output by Algorithm 3 with $(g_1, h_1), \ldots, (g_k, h_k)$ as input tuples, initial model f_0, and error tolerance γ. Then $h_{g_1}, \ldots, h_{g_k}$ will not witness meaningful model multiplicity on $g_1, \ldots, g_k$ on the updated model f.*

Proof. Since f is output by Algorithm 3, by statement 4 of Theorem D.3 it follows that for any submitted group g_i with accompanying model h_{g_i},

$$\begin{aligned}
\mathbb{E}_{g_i}[(f_t(x) - y)^2] &\leq \min_{\ell<t} \mathbb{E}_{g_i}[(f_\ell - y)^2] + \frac{\gamma}{\mu_{g_i}} \\
&= \mathbb{E}_{g_i}[(h_\ell(x) - y)^2] + \frac{\gamma}{\mu_{g_i}} \quad \text{(Statement 2 of Theorem D.3)} \\
&\leq \mathbb{E}_{g_i}[(h_{g_i}(x) - y)^2] + \frac{\gamma}{\mu_{g_i}}.
\end{aligned}$$

Therefor, no h_{g_i} submitted to Algorithm 3 witnesses multiplicity at level $\frac{\gamma}{\mu_{g_i}}$ with respect to f on g_i. □

Algorithm 1 ListUpdate $(f_t, (g_{t+1}, h_{t+1}))$

Input: Model f_t and certificate of suboptimality (g_{t+1}, h_{t+1}).

Define f_{t+1} as follows:

$$f_{t+1}(x) = \begin{cases} f_t(x) & \text{if } g_{t+1}(x) = 0 \\ h_{t+1}(x) & \text{if } g_{t+1}(x) = 1 \end{cases}$$

Output: f_{t+1}

Algorithm 2 CertificateChecker$(\gamma, D, U, (f_1, g_1, h_1), \ldots)$

Input: Holdout dataset D, Accuracy target γ, and a stream of submissions $(f_1, g_1, h_1), (f_2, g_2, h_2), \ldots$ of length at most U

NumberAccepted $= 0$

while NumberAccepted $\leq 2/\gamma$ and $i \leq U$ **do**

  Consider the next submission (f_i, g_i, h_i):

  Compute $\mu_i = \mu_{g_i}$ and $\Delta_i = L(D, f_i, g_i) - L(D, h_i, g_i)$

  if $\mu_i \cdot \Delta_i < \frac{3\gamma}{4}$ **then**

    Output: $\pi_i = \perp$ (Submission Rejected)

  else

    Output: $\pi_i = \top$ (Submission Accepted)

    NumberAccepted $=$ NumberAccepted + 1.

  end if

end while

Description of Algorithm 2: CertificateChecker takes as input a stream of submissions (f_i, g_i, h_i) and checks if (g_i, h_i) sufficiently improves the overall error of the model (i.e. if it is a sufficient certificate of suboptimality for f_i). If it does, it will accept the submission, outputting $\top$. Otherwise, it will reject it. It will output a message *every round* until either U submissions are reached or the number of accepted submissions exceeds $2/\gamma$.

Algorithm 3 MonotoneFalsifyAndUpdate($\gamma, D, (f_1, g_1, h_1), \ldots$)

Input: A model f_0, a holdout dataset D, Target γ, and a stream of submissions $(g_1, h_1), (g_2, h_2), \ldots$ of length at most U
Let $t = 0$.
Initialize an instance of CertificateChecker($\gamma, D, U + \frac{8}{\gamma^3}, \ldots$)
while CertificateChecker has not halted **do**
  Consider the next submission (g_i, h_i)
  Feed the triple (f_t, g_i, h_i) to CertificateChecker and receive $\pi_i \in \{\perp, \top\}$
  if $\pi_i = \perp$ **then**
    Output: Submission (g_i, h_i) is rejected.
  else
    MonotoneProgress=FALSE
    $(g_u, h_u) = (g_i, h_i)$
    $t' = t$ and $f'_{t'} = f_t$
    while MonotoneProgress = FALSE and CertificateChecker has not halted **do**
      Let $t' = t' + 1$ and let $f'_{t'} = \text{ListUpdate}(f'_{t'-1}, (g_u, h_u))$
      MonotoneProgress = TRUE
      for each pair $\ell < t$, $g_j \in \mathcal{G}(f_t)$ **do**
        Feed the triple (f_t, g_j, f_ℓ) to CertificateChecker and receive $\pi_u \in \{\perp, \top\}$
        if $\pi_u = \top$ (Submission accepted) **then**
          MonotoneProgress = FALSE
          $(g_u, h_u) = (g_j, f_\ell)$
          Break the for loop.
        end if
      end for
    end while
    Let $t = t + 1$ and let $f_t = f'_{t'}$
    Output: Submission (g_i, h_i) is accepted. The new model is f_t
  end if
end while

Description of Algorithm 3: Decision list construction which guarantees approximate group-wise error monotonicity. Here, i indexes the stream of submissions of length at most U, while t indexes the *accepted* submissions, which there are at most $2/\gamma$ of. After any submission (g_i, h_i) is accepted, all previously accepted groups $g_j \in \mathcal{G}(f_t)$ are sent again to CertificateChecker and are reincorporated into the model if there is a violation of monotonicity.

E. Deferred Proofs from Section 5.2

First, we note that the best linear combination with respect to squared error will always be orthogonal to its features.

Lemma E.1. *Let $f(x) = \beta \cdot (\phi_1(x), \ldots, \phi_m(x))$ be the best linear combination of m features $\phi_1, \ldots, \phi_m$ according to squared error on distribution $\mathcal{D}$. I.e.,*

$$\beta = \arg\min_{\beta' \in \mathbb{R}^m} \mathbb{E}_{\mathcal{D}}[(y - \beta' \cdot (\phi_1(x), \ldots, \phi_m(x))^2].$$

Then, for each feature ϕ_i,

$$\mathbb{E}_{\mathcal{D}}[\phi_i(x)(y - f(x))] = 0.$$

Proof. Since f minimizes squared error, its partial derivative with respect to any coefficient β_i for $i \in [m]$ will be 0. Hence,

$$0 = \frac{\partial}{\partial \beta_i} \mathbb{E}_{\mathcal{D}}[(y - f(x))^2] = -2\mathbb{E}_{\mathcal{D}}[\phi_i(x)(y - f(x))].$$

□

We can then leverage this to build a linear combination of models which witness multiplicity and group indicator functions as follows.

Theorem 5.2. *Let $(h_{g_1}, \ldots, h_{g_k})$ be a collection of models h_{g_i} which witness meaningful multiplicity with respect to h^* on groups $g_1, \ldots, g_k$ at level ϵ. Let f be the best linear combination of models weighted by group membership on distribution $\mathcal{D}$:*

$$f(x) = \beta_0 h^*(x) + \sum_{i=1}^{k} \beta_{i_1} h_{g_i}(x) g_i(x) + \beta_{i_2} h^*(x) g_i(x), \tag{1}$$

where

$$\beta = \arg \min_{\beta' \in \mathbb{R}^{2k+1}} \mathbb{E}_{\mathcal{D}} \left[\left(y - \beta'_0 h^*(x) - \sum_{i=1}^{k} \beta'_{i_1} h_{g_i}(x) g_i(x) - \beta'_{i_2} h^*(x) g_i(x) \right)^2 \right].$$

Then, for all $i \in [k]$, h_{g_i} does not witness multiplicity with respect to f on g_i. Moreover, if $\mathbb{E}_{g_i}[(h_{g_i}(x) - f(x))^2] > 0$, then f strictly outperforms h_{g_i} on g_i.

Proof. Consider model h_{g_i}'s performance on group g_i as opposed to f's. From Lemma C.1,

$$\mathbb{E}_{g_i}[(h_{g_i}(x) - y)^2] = \mathbb{E}_{g_i}[(h_{g_i}(x) - f(x))^2] + 2\mathbb{E}_{g_i}[h_{g_i}(x)(f(x) - y)] - 2\mathbb{E}_{g_i}[(f(x)(f(x) - y)] + \mathbb{E}_{g_i}[(f(x) - y)^2].$$

Here, we can directly apply Lemma E.1 to set the second term to zero: since $h_{g_i}(x)g_i(x)$ is a feature of f, $\mathbb{E}[h_{g_i}(x)g_i(x)(f(x) - y)] = 0$ and hence $\mathbb{E}_{g_i}[h_{g_i}(x)(f(x) - y)] = 0$ as well. For the third term, note that $f(x)g_i(x)$ is itself a linear combination of f's features and hence, applying linearity of expectations, this term also evaluates to zero.

Thus, we find that

$$\mathbb{E}_{g_i}[(h_{g_i}(x) - y)^2] - \mathbb{E}_{g_i}[(f(x) - y)^2] = \mathbb{E}_{g_i}[(h_{g_i}(x) - f(x))^2] > 0, \tag{4}$$

and h_{g_i} does not witness multiplicity with respect to f on g_i. □

Remark E.2. *Applying the same argument to h^* in place of h_{g_i} yields*

$$\mathbb{E}_{g_i}[(h^*(x) - y)^2] - \mathbb{E}_{g_i}[(f(x) - y)^2] = \mathbb{E}_{g_i}[(h^*(x) - f(x))^2]. \tag{5}$$

Subtracting (4) from (5) gives

$$\mathbb{E}_{g_i}[(h^*(x) - y)^2] - \mathbb{E}_{g_i}[(h_{g_i}(x) - y)^2] = \mathbb{E}_{g_i}[(h^*(x) - f(x))^2] - \mathbb{E}_{g_i}[(h_{g_i}(x) - f(x))^2],$$

expressing how much worse h^* is than h_{g_i} on g_i in terms of their respective distances to f.

F. Algorithms and Deferred Proofs from Section 6

Theorem 6.3. *Let $\mathcal{G}$ be any (potentially uncountable) collection of group indicator functions and let $\mathcal{H}$ be a hypothesis class mapping to real-valued predictions $\mathcal{Y}$. Let $f : \mathcal{X} \to \mathcal{Y}$ be a η-multiaccurate with respect to the product class $\{g \cdot h : g \in \mathcal{G} \text{ and } h \in \mathcal{H}\}$ and η-self-orthogonal with respect to $\mathcal{G}$. Then there will be no model $h \in \mathcal{H}$ which witnesses model multiplicity with respect to $g \in \mathcal{G}$ with density μ_g at a level $\varepsilon > 4\eta/\mu_g$.*

Proof. Since f is multiaccurate with respect to $\{g \cdot h : g \in \mathcal{G} \text{ and } h \in \mathcal{H}\}$, $\forall h \in \mathcal{H}$ and $g \in \mathcal{G}$,

$$\eta \geq \mathbb{E}[h(x)g(x)(f(x) - y) = \mu_g \mathbb{E}_g[h(x)(f(x) - y)]. \tag{6}$$

Similarly, from the self-orthogonality of f, it follows that $\forall g \in \mathcal{G}$,

$$\eta \geq |\mathbb{E}[f(x)g(x)(f(x) - y)]| = |\mu_g \mathbb{E}_g[f(x)(f(x) - y)]|\,. \tag{7}$$

Decomposing the squared error of h in terms of f following Lemma C.1,

$$\begin{aligned}
\mathbb{E}_g[(h(x) - y)^2] &= \mathbb{E}_g[(f(x) - y)^2] + 2\mathbb{E}_g[h(x)(f(x) - y)] - 2\mathbb{E}_g[f(x)(f(x) - y)] + \mathbb{E}_g[(f(x) - h(x))^2] \\
&\geq \mathbb{E}_g[(f(x) - y)^2] + 4\eta/\mu_g + \mathbb{E}_g[(f(x) - h(x))^2] \qquad \text{(Equations 6 and 7)} \\
&\geq \mathbb{E}_g[(f(x) - y)^2] + 4\eta/\mu_g \\
&\geq \mathbb{E}_g[(f(x) - y)^2] + \varepsilon.
\end{aligned}$$

□

G. Definitions and Lemmas for all Constraint-Based Fairness Notions in Section 7

Definitions from (Globus-Harris et al., 2023a)

G.1. Equalizing False Positive Rates

Definition G.1. *The false positive rate of a classifier $c : \mathcal{X} \to \mathcal{Y}$ on a group g is:*

$$\rho(c, g, \mathcal{D}) = \mathbb{P}_{(x,y)\sim\mathcal{D}}[c(x) \neq y | y = 0, g(x) = 1]$$

When c is a randomized classifier, the probabilities are computed over the randomness of c as well.

Definition G.2. *We say that classifier $c : \mathcal{X} \to \mathcal{Y}$ satisfies γ-False Positive (FP) Fairness with respect to $\mathcal{D}$ and $\mathcal{G}$ if for all $g \in \mathcal{G}$,*

$$w_g^{FP} \left|\rho_g(c) - \rho(c)\right| \leq \gamma.$$

where $w_g^{FP} = \mathbb{P}_{(x,y)\sim\mathcal{D}}[g(x) = 1, y = 0]$.

Lemma G.3. *An equivalent formulation of false positive fairness is that for all $g \in \mathcal{G}$,*

$$|\mathbb{E}[(1 - f^*(x)\ell(c(x), 0)g(x)] - \beta_g^{FP}\mathbb{E}[(1 - f^*(x))\ell(c(x), 0)]| \leq \gamma,$$

where $\beta_g^{FP} = \mathbb{P}[g(x) = 1 | y = 0]$.

G.2. Equalizing False Negative Rates

Definition G.4. *The false negative rate of a classifier $c : \mathcal{X} \to \mathcal{Y}$ on a group g is:*

$$\rho_{FN}(c, g, \mathcal{D}) = \mathbb{P}_{(x,y)\sim\mathcal{D}}[c(x) \neq y | y = 1, g(x) = 1]$$

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

Definition G.5. *We say that classifier $c : \mathcal{X} \to \mathcal{Y}$ satisfies γ-False Negative (FN) Fairness with respect to $\mathcal{D}$ and $\mathcal{G}$ if for all $g \in \mathcal{G}$,*

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

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

Lemma G.6. *An equivalent formulation of false negative fairness is that for all $g \in \mathcal{G}$,*

$$\left|\mathbb{E}[\ell(c(x), 1)g(x)f(x)] - \beta_g^{FN}\mathbb{E}\left[\ell(c(x), 1)f(x)\right]\right| \leq \gamma,$$

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

G.3. Statistical Parity

Definition G.7. *We say that classifier $c : \mathcal{X} \to \mathcal{Y}$ satisfies γ-Statistical Parity (SP) Fairness with respect to $\mathcal{D}$ and $\mathcal{G}$ if for all $g \in \mathcal{G}$,*

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

Lemma G.8. *An equivalent formulation of statistical parity fairness is that for all $g \in \mathcal{G}$,*

$$|\mathbb{E}[c(x)g(x)] - \beta_g^{SP}\mathbb{E}[c(x)]| \leq \gamma,$$

where $\beta_g^{SP} = \mathbb{P}[g(x) = 1]$.

G.4. Equalizing Group Error Rates

Definition G.9. *We say that classifier $c : \mathcal{X} \to \mathcal{Y}$ satisfies γ-Error (E) Fairness with respect to $\mathcal{D}$ and $\mathcal{G}$ if for all $g \in \mathcal{G}$,*

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

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

Lemma G.10. *An equivalent formulation of error fairness is that for all $g \in \mathcal{G}$,*

$$|\mathbb{E}[\ell(c(x), 1)g(x)f^*(x) + \ell(c(x), 0)g(x)(1 - f^*(x)) - \beta_g^E\ell(h(x), 0)(1 - f^*(x))]| \leq \gamma,$$

where $\beta_g^E = \mathbb{P}[g(x) = 1]$.

G.5. Multicalibration Definitions

G.5.1. Approximate Multicalibration

Definition G.11 (Multicalibration in Expectation (Hébert-Johnson et al., 2018; Gopalan et al., 2022))**.** *Fix a distribution $\mathcal{D}$ and let $\mathcal{C}$ be a collection of functions $c : \mathcal{X} \to \{0, 1\}$. We say that a predictor $f : \mathcal{X} \to R$ which maps to some discrete domain $R \subseteq [0, 1]$ is α-approximately multicalibrated with respect to $\mathcal{C}$ if for every $c \in \mathcal{C}$,*

$$\sum_{v\in R} \mathbb{P}[f(x) = v] \left|\mathbb{E}[(c(x)(v - y)|f(x) = v)]\right| \leq \alpha.$$

G.5.2. Joint Multicalibration on Threshold Function Classes for Downstream Fairness

Definition G.12 (Set of thresholding functions $\mathcal{B}(C)$)**.** *Let $x_{\mathcal{G}} \in \{0, 1\}^{|\mathcal{G}|}$ denote the group membership indicator vector of a point x. Then, for any λ, x, β, define s_λ to be the following thresholding function for each group fairness notion.*

$$s_\lambda^{FP}(x, v) = \mathbb{1}\left[\langle\lambda, x_{\mathcal{G}} - \beta\rangle \geq \frac{2v - 1}{1 - v}\right],$$

$$s_\lambda^{FN}(x, v) = \mathbb{1}\left[\langle\lambda, x_{\mathcal{G}} - \beta\rangle \geq \frac{1 - 2v}{v}\right],$$

$$s_\lambda^{SP}(x, v) = \mathbb{1}\left[\langle\lambda, x_{\mathcal{G}} - \beta\rangle \geq 2v - 1\right],$$

$$s_\lambda^{E}(x, v) = \mathbb{1}\left[\langle\lambda, x_{\mathcal{G}} - \beta\rangle \geq \frac{2v - 1}{1 - 2v}\right].$$

Define $\mathcal{B}(C)$ to be the collection of such thresholding functions with norm-bounded $\lambda \in \Lambda(C) = \{\lambda \in \mathbb{R}^{2\mathcal{G}} | \|\lambda\|_1 \leq C\}$ and $\beta \in \boldsymbol{\beta} = \{\beta_{g_1}, \ldots, \beta_{g_{|\mathcal{G}|}}\}$, where β_g is defined for each fairness notion in Lemmas G.3, G.6, G.8, and G.10 respectively. Then, $\mathcal{B}(C)$ may be defined as follows.

$$\mathcal{B}(C) = \{s_\lambda | \lambda \in \Lambda(C), \beta \in \boldsymbol{\beta}\}.$$

Definition G.13 (Joint Multicalibration in Expectation ((Globus-Harris et al., 2023a))). *Fix a distribution $\mathcal{D} \in \Delta\mathcal{Z}$ and a model $f : \mathcal{X} \to [0, 1]$ that maps onto a countable subset of its range $R(f)$. We say f is α-approximately jointly multicalibrated with respect to a class of functions $b : \mathcal{X} \times \mathcal{Y} \to \{0, 1\}$ if for every $b \in \mathcal{B}$ and $v \in R(f)$:*

$$\sum_{v \in R(f)} \mathbb{P}[f(x) \in v, b(x, v) = 1] \cdot \left|\mathbb{E}_{(x,y)\sim\mathcal{D}}\left[f(x) - y \middle| f(x) = v, b(x, v) = 1\right]\right| \leq \alpha.$$