REIMAGINING MEANINGFUL MODEL MULTIPLICITY

Anonymous authors
Paper under double-blind review

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 $h^* \in \mathcal{H}$. 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 duty 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 show that models which are suitably multiaccurate preclude 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 approaches 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 differ 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 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. First, we show that within the standard framework of model multiplicity, where a collection of multiplicitous models is provided, we can boost over this collection in order to build a model which has no meaningful multiplicity, while preserving generalization. We then use these ensembling methods as

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

the intuition behind the fact that models which are suitably multiaccurate—which do not require the explicit enumeration of multiplicitous models—will not demonstrate meaningful multiplicity.

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 final portion 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 $\mathcal{D} \in \Delta(\mathcal{X} \times \mathcal{Y})$, where the true underlying label $y \in \{0, 1\}$ is binary. We consider two different prediction mechanisms to learn this label. In Sections 2–5, we consider learning *real-valued* predictors $f \in \mathcal{H} : \mathcal{X} \to [0, 1]$ to estimate the conditional label probability $\mathbb{E}[y|x]$. In Section 6, we shift to classifiers $c \in \mathcal{C} : \mathcal{X} \to \{0, 1\}$—which can be generated, e.g., by thresholding the real-valued predictions of f. We show that even when the ultimate goal is classification, the regression-first approach developed here is best.

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, given finite data, 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 $\ell(h(x), y) = (h(x) - y)^2$ in expectation, 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 a 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 a 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 fewer 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, do not make sense—we take the narrow view of fairness as minimizing groupwise error rates; 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 6, we

show that a suitably multicalibrated label-probability predictor can also be efficiently postprocessed for fairness metrics that are in tension with accuracy.

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 2 through 6, additionally enjoy generalization guarantees, a property that existing approaches to searching the Rashomon set do not in general provide. As we show in Section 2, 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 4 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. The second method shows that when $|\mathcal{G}|$ is not too large, iteratively fitting linear combinations of group-model products also suffices.

Moving Beyond the Rashomon Set: Ideas from Multiaccuracy. 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.

Searching all of $\mathcal{H}$ might seem harder than searching $\mathcal{R}(\mathcal{H})$, but in Section 5 we show that there is a natural criterion for models that achieve this. The key is *multiaccuracy*: a model f that is multiaccurate with respect to $\mathcal{G} \times \mathcal{H}$—no $h \in \mathcal{H}$ can identify a subgroup in $\mathcal{G}$ on which f has systematic prediction error. This, combined with a criterion we call self-orthogonality, is guaranteed to have no competitive outperforming model in $\mathcal{H}$, and thus no model in $\mathcal{R}(\mathcal{H})$, for any group in $\mathcal{G}$.

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, such as when equalizing error rates across groups is in direct tension with overall accuracy, unlike in our group error minimization formulation. In Section 6, we show 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 $\mathcal{H}$.

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 iterates 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 Meaningful Model Multiplicity

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 $\ell : \mathcal{Y} \times \mathcal{Y} \to [0, 1]$— which we will generally assume to be 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 then 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.

There are several challenges to standard approaches: 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 2.1 (Meaningful Multiplicity)**.** Fix a hypothesis class $\mathcal{H}$ of models mapping from $\mathcal{X}$ to $\mathcal{Y}$, benchmark model $f : \mathcal{X} \to \mathcal{Y}$, and a group indicator $g : \mathcal{X} \to \{0, 1\}$. We say that $\mathcal{H}$ exhibits *meaningful multiplicity* on g at level $\varepsilon > 0$ with respect to f if there exists $h \in \mathcal{H}$ such that

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

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

Often, we will consider this notion of meaningful multiplicity specifically with respect to the general empirical risk minimizer $h^* \in \arg\min_{h \in \mathcal{H}} \mathbb{E}[\ell(h(x), y)]$.

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 2.2 (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 h_g witnesses ε meaningful multiplicity with respect to h^* on g. I.e.,*

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

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 2.3. *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 under comparison would make the problem harder. However, Sections 5 and 6 show that direct enumeration of the expanded space is unnecessary: the former gives a population-level certificate and existence result, while the latter gives an efficient postprocessing guarantee from a suitably multicalibrated predictor.

3 Model Multiplicity Implies Models are Far From Optimal

Before we build constructions for searching over the space of models, it is useful to build intuition for what it means to have meaningful model multiplicity: 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 model leverages.

Lemma 3.1 (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 3.2 (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 Lemmas 3.1 and 3.2, 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^*. In this way, $h_{g_1}, \ldots, h_{g_k}$ are witnesses to the suboptimality of h^*.

4 Resolving Model Multiplicity via Ensembling

In the constructions for the bounds in Section 3, an *ensemble* of the models in the multiplicity set is created, and we show that this ensemble is superior to the risk minimizer of $\mathcal{H}$. This provides intuition for how we might resolve meaningful multiplicity: by boosting or ensembling over the collection of disagreeing models. However, the specific constructions above are not scalable to implement in practice: they require statistical evaluations of loss over each subgroup induced by intersections of the multiplicitous models.

Here, we describe two alternative and efficient approaches to ensembling multiplicitous models. In line with the model multiplicity literature, we will assume that we are handed such a collection of "multiplicitous" models ahead of time, which we wish to choose between or ensemble. In subsequent sections we will remove this assumption and provide methods which, without explicitly constructing the full collection of multiplicitous models, still provide guarantees on them.[1] Our first approach

[1]The methodologies discussed here can be considered as complementary to the reconciliation approach recommended by (Roth et al., 2023) and (Behzad et al., 2025), but with fewer operations required.

consists of a decision list over models, treating each as a black box. The second approach provides the natural intuition for the following Sections 5 and 6. The proofs and algorithmic details are deferred to Appendices D and E.

4.1 Ensembling via Decision Lists

The first approach generates a decision list over a globally optimal h^* and group-optimal models $h_{g_1}, \ldots, h_{g_k}$. A naive implementation of such a decision list ensemble might iterate through the full collection of k groups and route each group g_i to h_{g_i}. However, if the groups overlap, this approach is insufficient. For instance, say that group g_i is contained within group g_j. Then, any decision list that has g_j appear as a predicate before g_i will route points in g_i to h_{g_j} rather than h_{g_i}, and it could be the case that h_{g_j} does poorly on this subgroup. Of course, one could instead learn models for all intersections, but this will lead to a total number of models exponential in the number of groups k, and hence generalize poorly!

Here, we offer an alternative solution, based on subroutines leveraged by Globus-Harris et al. (2022) in the context of "bias bounties". There, the authors construct decision lists of group-model pairs such that the total group error per group "accepted" as an addition to the decision list monotonically decreases at every update. Here, we ask for something slightly different: a guarantee that the final ensembled f will not display meaningful model multiplicity with respect to any of the models $h_{g_1}, \ldots, h_{g_k}$. The algorithm, which is described fully as Algorithm 2 (MultiplicityListEnsemble) in Appendix D, proceeds by iteratively checking through (g, h_g) pairs and appending them to a decision list that routes to h_g if x is a member of g, until no such updates are accepted. Once no such update is accepted, the process halts, and the final decision list provided will have no multiplicity with respect to any of the input model-group pairs.[2]

We now formally state its guarantees: first, that when the Algorithm 2 process is run with initial model f_0, and groups $g_1, \ldots, g_k$, then the final model f output by the algorithm will not witness ε-meaningful model multiplicity with respect to any of the input groups. As an immediate corollary, when the initial model f_0 is the risk minimizer h^* of some class $\mathcal{H}$ and the k group-model pairs $(g_i, h_{g_i})_{i\in[k]}$ are all multiplicity witnesses with respect to h^*, then the same claim holds, and moreover, the group-wise error of the final model f will be strictly better than that of h^* for all groups $\{g_i\}_{i\in[k]}$. Analogous finite-sample theorems with generalization guarantees are included in Appendix D.

Theorem 4.1 (Mitigating Multiplicity via Decision List Ensembles)**.** *Let $\mathcal{M} = (g_1, h_{g_1}), \ldots, (g_k, h_{g_k})$ be a collection of group indicators g_i and models h_{g_i} with nonzero support over $\mathcal{D}$. Let $\varepsilon \in (0, 1)$, and let $\mu^* = \min_{i\in[k]} \mu_{g_i}$. Let $f \leftarrow$ MultiplicityListEnsemble$(\mathcal{D}, \mathcal{M}, f^{init}, \gamma)$ be the model output by Algorithm 2 on distribution $\mathcal{D}$, initial model f^{init}, $\mathcal{M}$, and acceptance threshold $\gamma \leq \mu^*\varepsilon$. Then, no h_{g_i} will witness ε-meaningful model multiplicity with respect to f on g_i for all $(g_i, h_{g_i}) \in \mathcal{M}$.*

Corollary 4.2. *Let $h^* = \arg\min_{h\in\mathcal{H}} L(\mathcal{D}, h)$ be the risk minimizer of $\mathcal{H}$ on distribution $\mathcal{D}$. Let $h_1, \ldots, h_k$ be witnesses of ε-meaningful model multiplicity with respect to h^* on $g_1, \ldots, g_k$ respectively, and let $\mathcal{M} = \{(g_1, h_1), \ldots, (g_k, h_k)\}$, $\mu^* = \min_{i\in[k]} \mu_{g_i}$. Then, no h_i is a witness to ε-meaningful model multiplicity with respect to $f \leftarrow$ MultiplicityListEnsemble$(\mathcal{D}, \mathcal{M}, h^*, \gamma)$ for $\gamma \leq \mu^*\varepsilon$. Furthermore, f strictly improves on h^* for each of the groups: for all $i \in [k]$,*

$$L(\mathcal{D}, f, g_i) < L(\mathcal{D}, h_i, g_i) + \varepsilon \leq L(\mathcal{D}, h^*, g_i).$$

4.2 Ensembling via Linear Combinations

An alternative and complementary approach is to take the best linear combination over products of models and groups. Due to the first-order optimality of ordinary least squares, the final model's features will be orthogonal to the models' residuals in expectation—i.e., the error of the linear ensemble is uncorrelated with any of its features. This in turn means that there will be no model in the set of multiplicitous models which has correlation with the error of the linear combination, and hence any meaningful multiplicity must have been removed.

[2]The algorithm and analysis here differ from that of Globus-Harris et al. (2022), while using the same foundational building blocks. In particular, a naive implementation of their approach with $\{g_i\}_{i\in[k]}$ does not provide our guarantees.

More formally, we begin by defining the span of group-model pairs which we will use as features for our linear predictor, and the risk minimizer f over this span:

$$V = \text{span}\left(\{h^*\} \cup \{g_i h_{g_i} : i \in [k]\} \cup \{g_i h^* : i \in [k]\}\right), \\ f \in \arg\min_{v \in V} \mathbb{E}[(y - v(x))^2]. \tag{1}$$

The first-order conditions of ordinary least-squares then guarantee that for each $i \in [k], \mathbb{E}_{\mathcal{D}}[g_i(x)h_{g_i}(x)(f(x) - y)] = 0$, and, when groups are disjoint, $\mathbb{E}_{\mathcal{D}}[g_i(x)f(x)(f(x) - y)] = 0$; the result follows.

Theorem 4.3. *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. Suppose that the groups are disjoint and have nonzero support in the distribution. Let f be the best linear combination of models weighted by group membership on distribution $\mathcal{D}$ as in Equation 1. 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.*

When the groups intersect, this guarantee no longer holds: the orthogonality conditions of ordinary least-squares no longer imply self-orthogonality of the predictor on each group, because the intersection of groups introduces cross terms upon which orthogonality is not guaranteed. This could be resolved by including all group-intersections as features, but this would create an exponential number of features and hence generalization would take a hit. But, hope is not lost! Instead, we use an iterative process similar to that in Section 4.1 and which chooses intersections with sufficiently large correlations in each patch, iteratively refitting a linear predictor with this added feature.

Theorem 4.4. *Let $\mathcal{M} = (h_{g_1}, \ldots, h_{g_k})$ 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 on $\mathcal{D}$, where all groups have nonzero support. Let f^{init} be the best linear combination of models weighted by group membership on distribution $\mathcal{D}$ as in Equation 1. Fix $\gamma > 0$, and let $f \leftarrow$ SequentialLinearEnsemble(). Then, the algorithm terminates after at most $\lceil 1/\gamma^2 \rceil$ updates, and no h_{g_i} witnesses meaningful multiplicity with respect to f on g_i at any level $\varepsilon > 2\gamma/\mu_{g_i}$.*

5 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. This is a standard assumption within the model multiplicity literature, but pre-generating such a collection may be costly in terms of sample complexity. In particular, this approach requires an enumeration of all group identifiers $g \in \mathcal{G}$, which may be costly.

In practice, one might wish to achieve good performance on many complex and intersecting groups $\mathcal{G}$ when their full enumeration is prohibitive. Here, we observe that two relatively weak conditions suffice for a model f to have no meaningful multiplicity with respect to these groups: multiaccuracy with respect to the product class $\mathcal{G} \cdot \mathcal{H}$ and self-orthogonality on each group in $\mathcal{G}$. Thus, if a model in $\mathcal{H}$ disagrees with such an f, its disagreement is not *useful*: it cannot substantially improve on f's groupwise squared error.

Definition 5.1 (Multiaccuracy (Kim et al. (2019)))**.** Fix a distribution $\mathcal{D} \in \Delta\mathcal{Z}$ and a model $f : \mathcal{X} \to [0, 1]$ with finite range. Let $\mathcal{P} \subseteq [-1, 1]^{\mathcal{X}}$ be a class of bounded real-valued auditors. We say that f is α-approximately multiaccurate with respect to $\mathcal{D}$ and $\mathcal{P}$ if, for every $p \in \mathcal{P}$,

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

The other condition we will require is (groupwise) self-orthogonality, following the terminology of Kearns et al. (2026). Here, we demand that a model must be uncorrelated to its residuals, on group-weighted subsets of the distribution.

Definition 5.2 (Groupwise Self-Orthogonality (Kearns et al. (2026)))**.** Fix a collection of group indicators $\mathcal{G}$. We say that a model $f : \mathcal{X} \to [0, 1]$ is β-self-orthogonal with respect to $\mathcal{G}$ if, for every $g \in \mathcal{G}$, $|\mathbb{E}[g(x)f(x)(y - f(x))]| \leq \beta$.

Theorem 5.3. *Let $\mathcal{G}$ be a collection of group indicators, $\mathcal{H}$ a hypothesis class of real-valued predictors, and $f : \mathcal{X} \to [0,1]$ be a predictor with finite range. Say f is α-multiaccurate with respect to $\mathcal{G} \cdot \mathcal{H} := \{g \cdot h : g \in \mathcal{G},\ h \in \mathcal{H}\}$ and is β-self-orthogonal with respect to $\mathcal{G}$. Then, for every $g \in \mathcal{G}$ with nonzero support, no $h \in \mathcal{H}$ witnesses ε-meaningful multiplicity with respect to f on g for $\varepsilon > 2(\alpha+\beta)/\mu_g$. In particular, when $\alpha = \beta = \eta$, the threshold is $4\eta/\mu_g$.*

6 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 6.1 (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 $\bullet \in \{\text{FN}, \text{FP}, \text{E}, \text{SP}\}$ 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 6.2. 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 6.1 may be in direct tension with minimizing model error, which in the classification setting we will take to be zero-one loss.

Definition 6.3 (Zero-one loss)**.**

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

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 6.4. 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 $\text{err}(c') \leq \text{err}(c)$ and $\phi(c') \leq \phi(c)$ with at least one inequality strict.

How might one actually go about finding such a model? One method would be to solve the following constrained optimization problem over $\Delta\mathcal{C}$ for every value of $\gamma \in [0,1]$:

Definition 6.5 (Optimal Group-Constrained γ-Fair Model)**.**

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

where w_g, $\rho_g(c)$, and $\rho(c)$ are defined as in Definition 6.1.

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,

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

and similarly we can write $\phi(c)$ 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 Appendix G for details.

Remark 6.6. 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 6.5 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 6.7. Given predictor $f : \mathcal{X} \to [0, 1]$, let c_f be the solution to the optimization in Definition 6.5 with f substituted for f^*, i.e., using $f(x)$ as a proxy for $\mathbb{E}[y|x]$ in evaluating $\text{err}(c)$ and $\phi(c)$. The explicit substitutions needed for each fairness notion • 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 suitably *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.

Theorem 6.8 (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 • evaluated according to the projected gradient descent algorithm in Globus-Harris et al. (2022).*[3] *Then,*

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

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. Here, however, we now have a method to efficiently learn 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 6.9. *Let f satisfy the multicalibration conditions of Theorem 6.8, and for any $\gamma \in [0, 1]$, let $\bar{c}_\gamma$ denote the output of the optimal postprocessing procedure in 6.8 applied to f with fairness parameter γ. Then, $\bar{c}_\gamma$ lies on the Pareto front $\mathcal{P}(\Delta\mathcal{C})$, 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 Rashomon set $\mathcal{R}(\mathcal{C})$.*

Proof. This follows directly from Theorem 6.8 and the observation that $\mathcal{C} \subset \Delta\mathcal{C}$. □

REFERENCES

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

[3] 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.

Robert J. Aumann. 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.

Tina Behzad, Sílvia Casacuberta, Emily Ruth Diana, and Alexander Williams Tolbert. Reconciling predictive multiplicity in practice. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*, pp. 3350–3369, 2025.

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

Emily Black, Logan Koepke, Pauline Kim, Solon Barocas, and Mingwei Hsu. The legal duty to search for less discriminatory algorithms. *arXiv preprint arXiv:2406.06817*, 2024.

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

Natalie Collina, Surbhi Goel, Varun Gupta, and Aaron Roth. Tractable agreement protocols. In *Proceedings of the 57th Annual ACM Symposium on Theory of Computing*, pp. 1532–1543, 2025.

Natalie Collina, Ira Globus-Harris, Surbhi Goel, Varun Gupta, Aaron Roth, and Mirah Shi. 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.

Gordon Dai, Pavan Ravishankar, Rachel Yuan, Emily Black, and Daniel B Neill. 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.

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

Ira Globus-Harris, Varun Gupta, Christopher Jung, Michael Kearns, Jamie Morgenstern, and Aaron Roth. Multicalibrated regression for downstream fairness. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 259–286, 2023a.

Ira Globus-Harris, Declan Harrison, Michael Kearns, Aaron Roth, and Jessica Sorrell. Multicalibration as boosting for regression. *arXiv preprint arXiv:2301.13767*, 2023b.

Parikshit Gopalan, Adam Tauman Kalai, Omer Reingold, Vatsal Sharan, and Udi Wieder. Omnipredictors. In *13th Innovations in Theoretical Computer Science Conference (ITCS 2022)*, pp. 79–1. Schloss Dagstuhl–Leibniz-Zentrum für Informatik, 2022.

Chris Hays, Ben Laufer, Solon Barocas, and Manish Raghavan. Statistical guarantees in the search for less discriminatory algorithms. *arXiv preprint arXiv:2512.23943*, 2025.

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

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

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*, pp. 247–254, 2019.

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

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

Cynthia Rudin, Chudi Zhong, Lesia Semenova, Margo Seltzer, Ronald Parr, Jiachang Liu, Srikar Katta, Jon Donnelly, Harry Chen, and Zachery Boner. Amazing things come from having many good models. *arXiv preprint arXiv:2407.04846*, 2024.

Rui Xin, Chudi Zhong, Zhi Chen, Takuya Takagi, Margo Seltzer, and Cynthia Rudin. 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

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

As a shorthand, we will also write $L(\mathcal{D}, f, g) := \mathbb{E}_g\big[\ell(f(x), y)\big]$.

Definition A.7 (Rashomon set)**.** Given a hypothesis class $\mathcal{H}$, distribution $\mathcal{D}$, and $\varepsilon \in (0, 1/2]$, the ε-Rashomon set is the collection of models in $\mathcal{H}$ with error ε-close to the risk minimizer h^*:

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

B Deferred Proofs from Section 2

Lemma 2.2 (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 h_g witnesses ε meaningful multiplicity with respect to h^* on g. I.e.,*

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

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

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 2.3. *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 $\varepsilon < 1 - 2\mu_g$.*

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

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

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 3

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

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

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 3.1 (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*

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

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 3.2 (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*

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

Proof. As in the proof of Lemma 3.1, 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 $\mathcal{D}$:

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

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

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 4.1 for Decision List Ensembles

D.1 Preliminaries

We begin with definitions and a sub-algorithm that we will invoke in our framework from Globus-Harris et al. (2022), 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 $\mu, \Delta > 0$.

In Globus-Harris et al. (2022), a simple patching operation is used to ensemble a certificate of sub-optimality (g, h) to an existing model f, which is reproduced below in Algorithm 1.

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}

We then use this patching operation as a component of our ensembling technique to minimize model multiplicity.

D.2 Resolving Model Multiplicity via Decision List Ensembles

Here, each (g_i, h_i) pair in the collection of proposed multiplicitous models is checked iteratively. If one is a certificate of suboptimality in the sense of Definition D.2, then ListUpdate is used to append it to the model ensemble, and the process terminates when none of the multiplicitous models are certificates of suboptimality. We note that while this approach is similar to those proposed by Globus-Harris et al. (2022), it differs in that it tracks *all* submitted models, rather than only the

accepted ones: a naive implementation of the bias bounties approach in Globus-Harris et al. (2022) will not give the guarantee with respect to model multiplicity that we wish for, because a rejected group-model pair may later induce multiplicity, and must be rechecked.

Algorithm 2 MultiplicityListEnsemble($\mathcal{D}, \mathcal{M}, f^{\texttt{init}}, \gamma$)

Input: Distribution $\mathcal{D}$, finite list $\mathcal{M} = ((g_1, h_{g_1}), \ldots, (g_k, h_{g_k}))$ of group-model pairs, initial model $f^{\texttt{init}}$, and acceptance threshold γ.
Let $f_t = f^{\texttt{init}}$ and $\texttt{halt} = \texttt{FALSE}$.
while $\texttt{halt} = \texttt{FALSE}$ **do**
  Set $\texttt{halt} = \texttt{TRUE}$
  for $i = 1, \ldots, k$ **do**
    Let $\Delta_i = (L(\mathcal{D}, f, g_i) - L(\mathcal{D}, h_{g_i}, g_i))$.
    if $\Delta_i \mu_{g_i} \geq \gamma$ **then**
      Let $(g_t, h_t) = (g_i, h_i)$ be the accepted group-model pair.
      Let $f_{t+1} \leftarrow \text{ListUpdate}(f_t, (g_t, h_t))$.
      $\texttt{halt} = \texttt{FALSE}$
      $t = t + 1$
    end if
  end for
end while
Output: f_t.

Theorem 4.1 (Mitigating Multiplicity via Decision List Ensembles)**.** *Let $\mathcal{M} = (g_1, h_{g_1}), \ldots, (g_k, h_{g_k})$ be a collection of group indicators g_i and models h_{g_i} with nonzero support over $\mathcal{D}$. Let $\varepsilon \in (0, 1)$, and let $\mu^* = \min_{i \in [k]} \mu_{g_i}$. Let $f \leftarrow \text{MultiplicityListEnsemble}(\mathcal{D}, \mathcal{M}, f^{init}, \gamma)$ be the model output by Algorithm 2 on distribution $\mathcal{D}$, initial model f^{init}, $\mathcal{M}$, and acceptance threshold $\gamma \leq \mu^* \varepsilon$. Then, no h_{g_i} will witness ε-meaningful model multiplicity with respect to f on g_i for all $(g_i, h_{g_i}) \in \mathcal{M}$.*

Lemma D.3. *Let $f \leftarrow \text{MultiplicityListEnsemble}(\mathcal{D}, \mathcal{M}, f_0, \gamma)$ be the model output by Algorithm 2 on distribution $\mathcal{D}$, initial model f_0, $\mathcal{M} = \{(g_i, h_i)\}_{[k]}$, and acceptance threshold $\gamma \leq \mu^* \varepsilon$. Say at round t group-model pair $(g_t, h_t$ was accepted as an update, yielding the patched model $f_{t+1} = \text{ListUpdate}(f_t, (g_t, h_t))$. Then,*

- *f_{t+1} has strictly smaller distributional loss than f_t:*
$$L(\mathcal{D}, f_{t+1}) \leq L(\mathcal{D}, f_t) - \gamma,$$
- *the process terminates: the total number of accepted updates is $T \leq \lfloor 1/\gamma \rfloor$, and*
- *there are at most $k(N + 1)$ total comparisons between group-model pairs and the current model.*

Proof. Let (g_t, h_t) be the group-model pair accepted at round t of 2, and let $f_{t+1} \leftarrow \text{ListUpdate}(f_t, (g_t, h_t))$. Then, since the model has only been modified on group g_{t+1},

$$L(\mathcal{D}, f_t) - L(\mathcal{D}, f_{t+1}) = \mathbb{P}(g_t(x) = 1)(L(\mathcal{D}, f_t, g_t) - L(\mathcal{D}, h_t, g_t)) = \mu_t \Delta_t \geq \gamma,$$

where μ_t is the weight of group g_t and Δ_t is the error improvement at round t.

Since the output space of the prediction problem is bounded above by 1 the worst starting f_0 will have average squared error of 1. Thus, after $\lfloor 1/\gamma \rfloor$ accepted updates each with an additive improvement of γ to the squared error, the error will be able to be unable to improve without become negative, which it cannot do.

Moreover, if any outer round of update across the k proposed models in $\mathcal{M}$ leads to no accepted update, the process will terminate. Hence, in the worst case there will only be one group per outer pass through the k models which is accepted, and after the up to $\lfloor 1/\gamma \rfloor$ rounds and the final round where no group-model pair is accepted there are a total of $k(N+1)$ total group-model pair checks. □

Proof. (Proof of Theorem D.2) First, note that by Lemma D.3, Algorithm 2 does halt and converge to a model f.

In the terminal round of Algorithm 2, no (g_i, h_i) pair is accepted. Thus, no updates are incurred, and each of the group-model pairs are compared against the same model f_t. Since there were no accepted updates it must be the case that for each of the k groups, $\gamma > \mu_i \Delta_i = \mu_i(L(\mathcal{D}, f, g_i) - L(\mathcal{D}, h_{g_i}, g_i)$. Hence, $L(\mathcal{D}, f, g_i) < L(\mathcal{D}, h_{g_i}, g_i) + \gamma/\mu_i$.

From the statement of Theorem D.2, $\gamma \leq \mu^* \varepsilon$, where $\mu^* = \min_{i \in [k]} \mu_{g_i}$. Thus,

$$\frac{\gamma}{\mu_i} \leq \frac{\mu^* \varepsilon}{\mu_i} \leq \varepsilon,$$

and h_i is not an ε-meaningful model multiplicity witness for f on g_i. □

Corollary 4.2. *Let $h^* = \arg\min_{h \in \mathcal{H}} L(\mathcal{D}, h)$ be the risk minimizer of $\mathcal{H}$ on distribution $\mathcal{D}$. Let $h_1, \ldots, h_k$ be witnesses of ε-meaningful model multiplicity with respect to h^* on $g_1, \ldots, g_k$ respectively, and let $\mathcal{M} = \{(g_1, h_1), \ldots, (g_k, h_k)\}$, $\mu^* = \min_{i \in [k]} \mu_{g_i}$. Then, no h_i is a witness to ε-meaningful model multiplicity with respect to $f \leftarrow$ MultiplicityListEnsemble$(\mathcal{D}, \mathcal{M}, h^*, \gamma)$ for $\gamma \leq \mu^* \varepsilon$. Furthermore, f strictly improves on h^* for each of the groups: for all $i \in [k]$,*

$$L(\mathcal{D}, f, g_i) < L(\mathcal{D}, h_i, g_i) + \varepsilon \leq L(\mathcal{D}, h^*, g_i).$$

Proof. This follows directly from Lemma D.3 and Theorem D.2 with f_0 substituted for h^*: if $\gamma \leq \mu^* \varepsilon$, then for all $i \in [k]$,

$$\frac{\gamma}{\mu_{g_i}} \leq \frac{\mu^* \varepsilon}{\mu_{g_i}} \leq \varepsilon.$$

Thus, no initial model h_{g_i} can improve on the f output by MultiplicityListEnsemble by ε on its group g_i, and if the initial (g, h) pairs *were* witnesses to multiplicity on h^*, then combining the inequalities gives us that $L(\mathcal{D}, f, g_i) < L(\mathcal{D}, h^*, g_i)$. □

D.3 In-Sample Guarantees and Generalization

The above Algorithm 2 takes a distribution $\mathcal{D}$ as input. This could be instantiated on a finite sample $D \sim \mathcal{D}^n$ of n datapoints drawn iid from a distribution $\mathcal{D}$ by running the algorithm on an input distribution which is uniform over the sample D. However, then the guarantees will hold only in-sample, and not with respect to the original distribution $\mathcal{D}$ which the samples were drawn from.

Here, we modify the algorithm slightly and apply uniform convergence in order to get such an out-of-sample generalization for our multiplicity guarantees

Theorem D.4 (Mitigating Multiplicity via Decision List Ensembles with Generalization)**.** *Let $\mathcal{M} = (g_1, h_{g_1}), \ldots, (g_k, h_{g_k})$ be a collection of group indicators g_i and models h_{g_i} with nonzero support over $\mathcal{D}$. Let $\varepsilon \in (0, 1)$, and let $\mu^* = \min_{i \in [k]} \mu_{g_i} > 0$ where $\hat{\mu}_i$ is the weight of group i within sample D. Set $\gamma = \hat{\mu}\varepsilon/2$ and $B = \lfloor 4/3\gamma \rfloor$, and let $f \leftarrow$ MultiplicityListEnsemble$(D, \mathcal{M}, f^{init}, 3\gamma/4)$ be the model output by Algorithm 2 on sample $D \sim \mathcal{D}^n$, $\mathcal{M}$, and initial model f^{init}. Then, if*

$$n \geq \frac{128}{\varepsilon^2 \hat{\mu}^2} \left((B+1) \log(2(k+1)) + \log \frac{4}{\delta} \right),$$

it follows that with probability at least $1 - \delta$, no h_{g_i} will witness ε-meaningful model multiplicity with respect to f on g_i.

Lemma D.5 (Uniform concentration of empirical group weights)**.** *Let $D = ((x_1, y_1), \ldots, (x_n, y_n)) \sim \mathcal{D}^n$. For each $i \in [k]$, let*

$$\mu_i = \mathop{\mathbb{P}}_{(x,y) \sim \mathcal{D}}[g_i(x) = 1] \qquad \textit{and} \qquad \hat{\mu}_i = \frac{1}{n} \sum_{j=1}^{n} g_i(x_j).$$

Then, for any $\delta \in (0,1)$, with probability at least $1-\delta/2$ over the draw of D,

$$|\hat{\mu}_i - \mu_i| \leq \sqrt{\frac{\log(4k/\delta)}{2n}} \qquad \textit{for every } i \in [k].$$

Proof. For each fixed $i \in [k]$, $\hat{\mu}_i$ is the empirical average of the i.i.d. Bernoulli random variables $g_i(x_1), \ldots, g_i(x_n)$, each with mean μ_i. Hence, by Hoeffding's inequality, for every $t > 0$,

$$\mathop{\mathbb{P}}_{D\sim\mathcal{D}^n}\left[|\hat{\mu}_i - \mu_i| > t\right] \leq 2\exp(-2nt^2).$$

Setting

$$t = \sqrt{\frac{\log(4k/\delta)}{2n}},$$

it follows that for each group g_i,

$$\begin{aligned}\mathop{\mathbb{P}}_{D\sim\mathcal{D}^n}\left[|\hat{\mu}_i - \mu_i| > \tau\right] &\leq 2\exp(-2nt^2)\\ &= 2\exp\left(-\log\frac{4k}{\delta}\right)\\ &= \frac{\delta}{2k}.\end{aligned}$$

Thus, with probability $\delta/2k$, each of the empirical averages are close to their distributional analogue. Union bounding over the k groups then gives

$$\begin{aligned}\mathop{\mathbb{P}}_{D\sim\mathcal{D}^n}\left[\exists i \in [k] : |\hat{\mu}_i - \mu_i| > \tau\right] &\leq \sum_{i=1}^{k} \mathop{\mathbb{P}}_{D\sim\mathcal{D}^n}\left[|\hat{\mu}_i - \mu_i| > \tau\right]\\ &\leq k\left(\frac{\delta}{2k}\right)\\ &= \frac{\delta}{2}.\end{aligned}$$

Thus, with probability at least $1-\delta/2$, $|\hat{\mu}_i - \mu_i| \leq t$, for all groups g_i. □

Proof of Theorem D.4. Our goal will be to bound the difference between the empirical weighted improvement used to accept or reject updates on a model f and its distributional analogue. For any f, let

$$\begin{aligned}A_i(D,f) &= \hat{\mu}_i\left(L(D,f,g_i) - L(D,h_{g_i},g_i)\right),\\ A_i(\mathcal{D},f) &= \mu_i\left(L(\mathcal{D},f,g_i) - L(\mathcal{D},h_{g_i},g_i)\right).\end{aligned}$$

Then, applying Hoeffding to any fixed model f that may be reached after at most t accepted updates gives

$$\mathbb{P}\left[|A_i(D,f) - A_i(\mathcal{D},f)| > \tau_t\right] \leq 2\exp\left(-\frac{n\tau_t^2}{2}\right).$$

Note also that since each accepted update must be one of the k group-model pairs, the total number of decision lists reachable after at most t accepted updates is bounded above by $(k+1)^t$. Thus, taking a union over all such models and all $i \in [k]$, and setting

$$\tau_t = \sqrt{\frac{2}{n}\left((t+1)\log(2(k+1)) + \log\frac{4}{\delta}\right)},$$

we find that the probability of a bad event at round t is bounded above by

$$(k+1)^{t+1}2\exp\left(-\frac{n\tau_t^2}{2}\right) \leq \frac{\delta}{2^{t+2}}.$$

Then, summing over all rounds and applying Lemma D.5, we find that, with probability at least $1-\delta$, for every round t, every model f reachable after at most t accepted updates, and every $i \in [k]$,

$$|A_i(D, f) - A_i(\mathcal{D}, f)| \leq \tau_t,$$

and the group-weight concentration guarantee also holds.

We now show that when the algorithm is run on a sufficiently large sample, the desired guarantee holds. Fix a sample on which both concentration guarantees hold and which satisfies the sample-size condition, and let T be the number of accepted updates. Since empirical loss lies in $[0, 1]$ and each accepted update decreases it by at least $3\gamma/4$, we have $T \leq B$. The sample-size condition therefore gives

$$\tau_T \leq \tau_B \leq \frac{\gamma}{4} \qquad \text{and} \qquad \sqrt{\frac{\log(4k/\delta)}{2n}} \leq \frac{\hat{\mu}}{2}.$$

Thus, Lemma D.5 implies

$$\mu_i \geq \hat{\mu}_i - \frac{\hat{\mu}}{2} \geq \frac{\hat{\mu}}{2} > 0, \qquad \gamma = \frac{\varepsilon\hat{\mu}}{2} \leq \varepsilon\mu_i.$$

During the final pass, every group-model pair is rejected against the same final model f, so $A_i(D, f) < 3\gamma/4$. Hence,

$$A_i(\mathcal{D}, f) \leq A_i(D, f) + \tau_T < \gamma \leq \varepsilon\mu_i.$$

Dividing by $\mu_i > 0$ gives

$$L(\mathcal{D}, f, g_i) < L(\mathcal{D}, h_{g_i}, g_i) + \varepsilon,$$

which proves the claim. □

E Deferred Proofs from Section 4.2

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. For disjoint groups, we need only fit to a collection of features with respect to the groups:

$$\begin{aligned} V &= \operatorname{span}\left(\{h^*\} \cup \{g_i h_{g_i} : i \in [k]\} \cup \{g_i h^* : i \in [k]\}\right), \\ f &\in \arg\min_{v \in V} \mathbb{E}[(y - v(x))^2]. \end{aligned} \tag{4}$$

Theorem 4.3. *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. Suppose that the groups are disjoint and have nonzero support in the distribution. Let f be the best linear combination of models weighted by group membership on distribution $\mathcal{D}$ as in Equation 1. 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. Fix $i \in [k]$, and let V and f be as in Equation 1. Since $g_i h_{g_i} \in V$, Lemma E.1 gives

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

Because the groups are disjoint and $g_i^2 = g_i$, multiplication by g_i maps every generator of V back into V. Hence $g_i f \in V$, and another application of Lemma E.1 gives

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

Since g_i has nonzero support, conditioning both equalities on $g_i(x) = 1$ and applying the squared-loss identity yields

$$\begin{aligned}
&\mathbb{E}_{g_i}[(h_{g_i}(x) - y)^2] - \mathbb{E}_{g_i}[(f(x) - y)^2] \\
&\quad = \mathbb{E}_{g_i}[(h_{g_i}(x) - f(x))^2] + 2\mathbb{E}_{g_i}[h_{g_i}(x)(f(x) - y)] - 2\mathbb{E}_{g_i}[f(x)(f(x) - y)] \\
&\quad = \mathbb{E}_{g_i}[(h_{g_i}(x) - f(x))^2] \geq 0.
\end{aligned}$$

Thus h_{g_i} does not witness multiplicity with respect to f on g_i. If $\mathbb{E}_{g_i}[(h_{g_i}(x) - f(x))^2] > 0$, the inequality is strict, so f strictly outperforms h_{g_i} on g_i. □

However, when groups intersect, we have to do extra work: in rounds, we must iteratively add features: at each round, we add a feature within the collection of intersecting groups which has sufficient correlation.

Algorithm 3 SequentialLinearEnsemble($\mathcal{D}, \mathcal{M}, f^{\texttt{init}}, \gamma$)

Input: Distribution $\mathcal{D}$, $\mathcal{M} = ((g_1, h_{g_1}), \ldots, (g_k, h_{g_k}))$ set of group-model pairs, initial benchmark model $f^{\texttt{init}}$, and tolerance $\gamma \geq 0$.
Set $V_0 = \text{span}\left(\{f^{\texttt{init}}\} \cup \{g_i h_{g_i}, g_i f^{\texttt{init}} : i \in [k]\}\right)$ and $t = 0$.
Fit $f_t \in \arg\min_{v \in V_t} \mathbb{E}_{\mathcal{D}}[(y - v(x))^2]$.
while there is an $i \in [k]$ such that $|\mathbb{E}_{\mathcal{D}}[g_i(x)f_t(x)(f_t(x) - y)]| > \gamma$ **do**
  Choose group $i_t \in [k]$ which has correlation with the residuals of the current model, i.e. such that
  $$|\mathbb{E}_{\mathcal{D}}[g_{i_t}(x)f_t(x)(f_t(x) - y)]| > \gamma.$$
  Set $\phi_t = g_{i_t} f_t$ as the new model feature, and append it to the feature span, setting $V_{t+1} = \text{span}\left(V_t \cup \{\phi_t\}\right)$.
  Fit $f_t \in \arg\min_{v \in V_t} \mathbb{E}_{\mathcal{D}}[(y - v(x))^2]$.
  Set $t \leftarrow t + 1$.
end while
Output: f_t.

Theorem 4.4. *Let $\mathcal{M} = (h_{g_1}, \ldots, h_{g_k})$ 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 on $\mathcal{D}$, where all groups have nonzero support. Let f^{init} be the best linear combination of models weighted by group membership on distribution $\mathcal{D}$ as in Equation 1. Fix $\gamma > 0$, and let $f \leftarrow$ SequentialLinearEnsemble(). Then, the algorithm terminates after at most $\lceil 1/\gamma^2 \rceil$ updates, and no h_{g_i} witnesses meaningful multiplicity with respect to f on g_i at any level $\varepsilon > 2\gamma/\mu_{g_i}$.*

Proof. Run Algorithm 3 with input tolerance γ, and let

$$L_t := \mathbb{E}_{\mathcal{D}}\left[(y - f_t(x))^2\right]$$

be the expected squared loss after the least-squares fit at round t. We first show that each update decreases squared error. Let $\phi_t = g_{i_t} f_t$ be the feature selected by the algorithm—note that by definition, this feature must have at least γ correlation with the current model's residuals:

$$|\mathbb{E}_{\mathcal{D}}\left[\phi_t(x)(f_t(x) - y)\right]| > \gamma.$$

Furthermore, since ϕ_t is added to the span V_{t+1}, this space contains a linear function $\widetilde{f}_{t+1} = f_t - a\phi_t$ for every $a \in \mathbb{R}$, which will have loss

$$\mathbb{E}_{\mathcal{D}}\left[(y - f_t(x) + a\phi_t(x))^2\right] = L_t - 2a\mathbb{E}_{\mathcal{D}}[\phi_t(x)(f_t(x) - y)] + a^2\mathbb{E}_{\mathcal{D}}[\phi_t(x)^2].$$

Minimizing this quadratic gives us the constant

$$a_t := \frac{\mathbb{E}_{\mathcal{D}}[\phi_t(x)(f_t(x) - y)]}{\mathbb{E}_{\mathcal{D}}[\phi_t(x)^2]}.$$

The denominator is positive, and hence a_t is well defined, because

$$0 < \gamma^2 < |\mathbb{E}_{\mathcal{D}}[\phi_t(x)(f_t(x) - y)]|^2 \leq \mathbb{E}_{\mathcal{D}}[\phi_t(x)^2] \leq 1.$$

Therefore, the model $\widetilde{f}_{t+1} := f_t - a_t\phi_t$ satisfies

$$\begin{aligned}\mathbb{E}_{\mathcal{D}}\left[(y - \widetilde{f}_{t+1}(x))^2\right] &= L_t - \frac{\left(\mathbb{E}_{\mathcal{D}}[\phi_t(x)(f_t(x) - y)]\right)^2}{\mathbb{E}_{\mathcal{D}}[\phi_t(x)^2]} \\ &< L_t - \gamma^2,\end{aligned}$$

Since f_{t+1} is the least-squares minimizer over V_{t+1}, its loss may be no larger than that of $\widetilde{f}_{t+1}$. Thus every update decreases expected squared loss by more than γ^2, and can run for only at most $\lceil 1/\gamma^2 \rceil$ rounds.

We now show that the final model output, f, does not display meaningful model multiplicity. Note that since every feature $g_i h_{g_i}$ belongs to V_0, it also belongs to the final feature space V_t. Thus, by Lemma E.1,

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

Furthermore, since the process terminated,

$$|\mathbb{E}_{\mathcal{D}}[g_i(x)f(x)(f(x) - y)]| \leq \gamma.$$

Then,

$$\begin{aligned}&\mathbb{E}_{g_i}[(h_{g_i}(x) - y)^2] - \mathbb{E}_{g_i}[(f(x) - y)^2] \\ &= \mathbb{E}_{g_i}[(h_{g_i}(x) - f(x))^2] + 2\mathbb{E}_{g_i}[h_{g_i}(x)(f(x) - y)] - 2\mathbb{E}_{g_i}[f(x)(f(x) - y)] \\ &\geq \mathbb{E}_{g_i}[(h_{g_i}(x) - f(x))^2] - \frac{2\gamma}{\mu_{g_i}},\end{aligned}$$

because the first residual term is zero and the second has absolute value at most γ/μ_{g_i}. Therefore no h_{g_i} witnesses meaningful multiplicity with respect to f on g_i at any level $\varepsilon > 2\gamma/\mu_{g_i}$. □

F Algorithms and Deferred Proofs from Section 5

Theorem 5.3. *Let $\mathcal{G}$ be a collection of group indicators, $\mathcal{H}$ a hypothesis class of real-valued predictors, and $f : \mathcal{X} \to [0, 1]$ be a predictor with finite range. Say f is α-multiaccurate with respect to $\mathcal{G} \cdot \mathcal{H} := \{g \cdot h : g \in \mathcal{G},\ h \in \mathcal{H}\}$ and is β-self-orthogonal with respect to $\mathcal{G}$. Then, for every $g \in \mathcal{G}$ with nonzero support, no $h \in \mathcal{H}$ witnesses ε-meaningful multiplicity with respect to f on g for $\varepsilon > 2(\alpha + \beta)/\mu_g$. In particular, when $\alpha = \beta = \eta$, the threshold is $4\eta/\mu_g$.*

Proof. Fix $g \in \mathcal{G}$ with $\mu_g > 0$ and $h \in \mathcal{H}$. Decomposing the squared error of h around f following Lemma C.1 gives

$$\begin{aligned}L(\mathcal{D}, h, g) - L(\mathcal{D}, f, g) &= \mathbb{E}_g\left[(h - f)^2\right] + 2\mathbb{E}_g\left[(h - f)(f - y)\right] \\ &= \mathbb{E}_g\left[(h - f)^2\right] - \frac{2}{\mu_g}\mathbb{E}\left[gh(y - f)\right] + \frac{2}{\mu_g}\mathbb{E}\left[gf(y - f)\right] \\ &\geq \mathbb{E}_g\left[(h - f)^2\right] - 2(\alpha + \beta)/\mu_g,\end{aligned}$$

where the final inequality follows from product-class multiaccuracy and groupwise self-orthogonality. Since the squared-disagreement of h and f is nonnegative, this in turn means that if $\varepsilon > 2(\alpha+\beta)/\mu_g$, then

$$L(\mathcal{D}, h, g) > L(\mathcal{D}, f, g) - \varepsilon.$$

Setting $\alpha = \beta = \eta$ gives the threshold $4\eta/\mu_g$. □

G Definitions and Lemmas for all Constraint-Based Fairness Notions in Section 6

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:

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

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 $g \in \mathcal{G}$,

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

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

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:

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

$\rho_g^{\text{FN}}(c) \equiv \rho_{\text{FN}}(c, g, \mathcal{D})$, and $\rho_{\text{FN}}(c) \equiv \rho(c, I, \mathcal{D})$.

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 $g \in \mathcal{G}$,

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

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

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 $g \in \mathcal{G}$,

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

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 $g \in \mathcal{G}$,

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

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

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 $c \in \mathcal{C}$,

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

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.

$$\begin{aligned}
s_\lambda^{\text{FP}}(x, v) &= \mathbb{1}\left[\langle \lambda, x_{\mathcal{G}} - \beta \rangle \geq \frac{2v - 1}{1 - v}\right], \\
s_\lambda^{\text{FN}}(x, v) &= \mathbb{1}\left[\langle \lambda, x_{\mathcal{G}} - \beta \rangle \geq \frac{1 - 2v}{v}\right], \\
s_\lambda^{\text{SP}}(x, v) &= \mathbb{1}\left[\langle \lambda, x_{\mathcal{G}} - \beta \rangle \geq 2v - 1\right], \\
s_\lambda^{\text{E}}(x, v) &= \mathbb{1}\left[\langle \lambda, x_{\mathcal{G}} - \beta \rangle \geq \frac{2v - 1}{1 - 2v}\right].
\end{aligned}$$

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 $v \in R(f)$:

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

H Declaration of LLM Usage

Claude Sonnet 4.6 was used for proof verification and proofreading for an initial version of the text and a restructuring of the abstract. GPT-5.6 Sol and 6 Astra was used for subsequent rounds of correctness auditing of the proofs, and suggested fixes as well as notational streamlining were iteratively integrated.