LEARNING WITH MULTI-GROUP GUARANTEES FOR CLUSTERABLE SUBPOPULATIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

A canonical desideratum for prediction problems is that performance guarantees should hold not just on average over the population, but also for meaningful subpopulations within the overall population. But what constitutes a meaningful subpopulation? In this work, we take the perspective that relevant subpopulations should be defined with respect to the clusters that naturally emerge from the distribution of individuals for which predictions are being made. In this perspective, a population refers to a mixture model whose components constitute the relevant subpopulations. We suggest two formalisms for capturing per-subgroup guarantees: first, by attributing each individual to the component from which they were most likely drawn, given their features; and second, by attributing each individual to all components in proportion to their relative likelihood of having been drawn from each component. Using online calibration for Gaussian mixture models as a case study, we study a multi-objective algorithm that provides guarantees for each of these formalisms by handling all plausible underlying subpopulation structures simultaneously, and achieve an $O(T^{1/2})$ rate even when the subpopulations are not well-separated. In comparison, the more natural *cluster-then-predict* approach that first recovers the structure of the subpopulations and then makes predictions suffers from a $O(T^{2/3})$ rate and requires the subpopulations to be separable. Along the way, we prove that providing per-subgroup calibration guarantees for underlying clusters can be easier than learning the clusters: separation between median subgroup features is required for the latter but not the former.

1 INTRODUCTION

For systems that make predictions about individuals, it is well-understood that good performance on average across the population may not imply good performance at an individual level. On the other hand, while the ideal system might be one that can provide per-individual performance guarantees, such a system may be intractable to learn from data, if it exists at all. To address these challenges, *per-subpopulation* guarantees have emerged as a widely-accepted approach that balances tractability with ensuring good performance across subpopulations (e.g., Blum et al. (2017); Hébert-Johnson et al. (2018); Hashimoto et al. (2018); Lahoti et al. (2020); Wang et al. (2020); Haghtalab et al. (2022)). Such guarantees may also be desirable for normative or regulatory reasons to capture notions of fairness, or because domain shift often involves changes in the proportions of subgroups. Therefore, the subpopulations for which guarantees are provided should be those that are deemed especially significant, salient, or relevant.

What, then, defines a relevant subpopulation? One influential perspective considers a subpopulation as a predefined combination of feature values, where individuals are represented as feature vectors (e.g., Hébert-Johnson et al. (2018)). In our work, we take an alternative view on what constitutes a subpopulation of interest. We propose that the relevant subgroups for a particular prediction task should be exactly those subgroups that emerge endogenously within the distribution of the individuals being considered for that task. This means that the group membership(s) of any individual cannot be determined through their features alone; instead, their group identity can be understood only by placing their individual features in the context of the rest of the population. In effect, rather than being defined *a priori*, these subgroups must be *learned* about from data in an unsupervised sense. We discuss further motivations for such an approach in Appendix A.

To operationalize our approach, our measures of per-group performance must handle the fact that any individual's group membership can be at best approximately inferred, e.g. as probabilities representing the likelihood that that individual belonged to each group. Accordingly, we study two natural approaches to measuring per-group error. The first, which we refer to as *discriminant error*, attributes the error an individual experiences only to the group that the individual *most likely* belongs to. This corresponds to typical notions of clustering error and is often used in existing approaches to handling uncertainty in group membership, which is effectively to ignore it (see, for example, discussion in Dong et al. (2024)). We also study a probabilistic alternative, which we term *likelihood error*, where we attribute the error an individual experiences to every group, but weighted according to the likelihoods of membership in each group. This likelihood-based notion of per-group error explicitly acknowledges the existence of meaningful uncertainty in group membership. As a consequence, likelihood error also provides some reasonable robustness properties (e.g., to changes in the relative proportions of subgroups), and, relative to discriminant error, improves guarantees for subgroups that comprise a smaller proportion of the total population.

Both of these measures, however, require knowledge of the subgroup distributions—that is, the likelihood that any particular individual (feature vector) belongs to (was drawn from) any particular group—which depends on the population distribution for the prediction task, and are therefore initially unknown. The natural strategy for addressing the problems of unknown subgroups and unknown labels is to first complete the unsupervised task of learning the subgroups, then for each of the learned distributions complete the supervised task of learning to predict; we call this the *cluster-then-predict* approach. The overall prediction quality of this approach critically depends on how well the "clustering" stage can be performed—a task that often requires a large number of observations and separation between subpopulations. We circumvent this problem through a *multi-objective* approach where, instead of learning the exact underlying clustering, we construct a class of plausible clusterings and provide high-quality per-group predictions for all of them simultaneously. What clusterings are "plausible," and how can we provide a solution that works for all of them? These constitute the central technical thrusts of our work, which we outline below.

1) Understanding the structure of subpopulations. We instantiate our model of subpopulations for Gaussian mixture models and consider the class of all plausible group membership functions corresponding to the two formalisms of discriminant error and likelihood error. We give upper bounds on the covering number of these classes for Gaussian mixture models, which we expect to be useful for the broader community and beyond multi-group learning.

2) Multi-objective algorithms for per-subgroup guarantees. Using calibration as a case study, we derive algorithms providing per-subgroup guarantees in our model. We leverage recent results connecting online multicalibration to online min-max optimization to minimize calibration error simultaneously over all clusters in a cover of the entire function class. For both discriminant and likelihood calibration error, our multi-objective approach achieves $O(T^{1/2})$ online error without requiring separability in the underlying clusters. This is in contrast to the error rates of the cluster-then-predict approach, for which we demonstrate $O(T^{2/3})$ error rates even under separability assumptions.

3) Towards statistically-identifiable subpopulations. Beyond the technical approach, we view our work as an important step towards reasoning about group membership in context of the actual population on which predictions are being made. Our work argues that subpopulations should be defined endogenously, rather than characterized by explicit combinations of feature values. In Appendix A, we further discuss the normative implications of viewing subpopulations as such. Our framework also provides a language for formalizing the relationship between explicitly learning subgroups, as opposed to providing high quality predictions for them: in fact, the former (which often requires separation between subpopulation means) is not necessary for the latter.

1.1 Related work

Fair machine learning. The fair machine literature has developed various approaches to handling uncertainty in group membership. One strategy is to avoid enumerating subgroups entirely, and instead focus on identifying subsets of the domain where prediction error is high (Hashimoto et al., 2018; Lahoti et al., 2020). A separate line of work considers learning when demographic labels are available but noisy (Awasthi et al., 2020; Wang et al., 2020). Yet another approach is to use a separate estimator for group membership (Chen et al., 2019; Awasthi et al., 2021; Kallus

et al., 2022); we note that our notions of per-group performance could be applicable to these methods as well, even if our definitions of "group" are different. Liu et al. (2023) also propose a means of understanding group identity in context with the rest of the population, in this case through social networks.

Multicalibration. Calibration is a well-studied objective in online forecasting (Dawid, 1982; Hart, 2022), with classical literature having studied calibration across multiple sub-populations (Foster & Kakade, 2006) and recent literature having studied calibration across computationally-identifiable feature groups (Hébert-Johnson et al., 2018). The latter thread of work, known as multicalibration, has found a wide range of connections to Bayes optimality, conformal predictions, and computational indistinguishability (Hébert-Johnson et al., 2018; Jung et al., 2021; Gopalan et al., 2022; Gupta et al., 2022; Dwork et al., 2021; Jung et al., 2023). We use online multicalibration algorithms (Gupta et al., 2021; Haghtalab et al., 2023) as a building block for efficiently obtaining per-group guarantees in our model.

2 PRELIMINARIES

Our generative model. Let $\mathcal{X} \in \mathbb{R}^d$ denote a d-dimensional instance space and $\mathcal{Y} = \{0, 1\}$ denote the *label* or *outcome* space. We consider a generative model over $\mathcal{X} \times \mathcal{Y}$, where instances are generated from a mixture of k distributions and the conditional outcome distribution is independent of the component from which the instance is generated. Formally, we define a discrete hidden-state *endogenous subgroups generative model* f, such that

$$f(x, y) \propto f(y \mid x) f(x \mid g) f(g),$$

where $f(g) = w_g$ is the distribution over $[k]$ corresponding to mixing weights $w_g \in [0, 1]$ with $\sum_{g \in [k]} w_g = 1$; $f(x \mid g)$ is the density of component g, and $f(y \mid x)$ is a conditional label distribution that is independent of g. In this work, we focus on the case where $f(x \mid g)$ corresponds to Gaussian distribution $\mathcal{N}(\mu_g, \Sigma_g)$. That is, (x, y) is generated by first sampling integer $g \in [k]$ according to weights $(w_1, \ldots, w_k)$, then sampling $x \sim \mathcal{N}(\mu_g, \Sigma_g)$, and finally sampling y according to $f(y \mid x)$. For clarity, we will often suppress w_g in the following exposition, but our results follow without loss of generality as long as all w_g are bounded below by a constant.

Online prediction. Our high level goal is to take high-quality actions for instances that are generated from an unknown endogenous subgroups model. Let $\mathcal{A}$ denote the action space. Examples of action spaces for prediction tasks include $\mathcal{A} = \{0, 1\}$ where an action refers to a predicted label, or $\mathcal{A} = [0, 1]$ where an action refers to predicting the probability that the label is 1. We consider an online prediction problem where a sequence of instance-outcome pairs $(x_1, y_1), \ldots, (x_T, y_T)$ is generated i.i.d. from an unknown generative model f supported on $\mathcal{X} \times \mathcal{Y}$. At time t, the learner must take an irrevocable action $a_t \in \mathcal{A}$ having seen only $x_{1:t}$ and $y_{1:t-1}$. Equivalently, a learner can be thought of as choosing a function $p_t : \mathcal{X} \to \mathcal{A}$ that maps any feature x to an action a. From this perspective, at time t the learner chooses p_t having only observed $x_{1:t-1}$ and $y_{1:t-1}$, after which (x_t, y_t) is observed and the learner takes action $p_t(x_t)$.

Performance on subpopulations. We evaluate the quality of the learner's actions using a vector-valued loss function $\ell : \mathcal{A} \times \mathcal{Y} \to \mathcal{E}$ where $\mathcal{E}$ is some Euclidean space. Examples of loss functions include the scalar binary loss function $\ell(a, y) := \mathbb{1}[a \neq y]$ and the vector-based calibration loss $\ell(a, y)_v = \mathbb{1}[a = v](a - y)$.

In the style of Blackwell approachability (Blackwell, 1956), the learner's overall goal is to produce actions that lead to small cumulative loss on all the relevant subpopulations in the sequence $(x_1, y_1), \ldots, (x_T, y_T)$, as measured by a norm $\|\cdot\|$. We envision relevant subpopulations to be exactly the mixture components of our generative model. However, it is not possible to determine the component that generates an instance (x, y) in a mixture model. Instead, we consider two notions of performing well on subpopulations. In the first, we purely attribute each (x, y) to the component $g = \arg\max_{j \in [k]} f(j \mid x, y)$ that was most likely responsible for producing (x, y); that is, we aim to minimize

$$\max_{g \in [k]} \left\| \sum_{t=1}^{T} \mathbb{1}\left[g = \arg\max_{j \in [k]} f(j \mid x_t, y_t) \right] \cdot \ell(a_t, y_t) \right\|.$$

The attribution of (x, y) to the most likely component corresponds to the usual task of clustering as is done in practice.

In the second, we attribute each (x, y) to a subpopulation g with probability $f(g \mid x, y)$; that is, we aim to minimize

$$\max_{g\in[k]} \Big\| \sum_{t=1}^{T} f(g \mid x_t, y_t) \cdot \ell(a_t, y_t) \Big\|.$$

Note that by definition, $f(g \mid x, y) = \frac{f(x|g)w_g}{\sum_{j\in[k]} f(x|j)w_j}$ is the probability g was indeed responsible for producing (x, y). This objective considers the contribution of an individual to subgroup error in proportion to the uncertainty of that individual's "group membership." Additionally, note that this objective is robust to reweightings of subpopulations, as $\mathbb{E}\left[f(g \mid x_t, y_t)\ell(a_t, y_t)\right] = f(g)\,\mathbb{E}\left[\ell(a_t, y_t)\,|g\right]$.

2.1 Model instantiation for calibration loss

For concreteness, this paper focuses on studying an instantiation of our model for the task of producing predictions that are *calibrated* with respect to clusterable subpopulations. However, our approach extends beyond calibration to a number of other settings that can be studied under online approachability, such as online *conformal prediction* and *calibeating* (Jung et al., 2023; Lee et al., 2022).

For studying calibrated predictions, we work with action space $\mathcal{A} = [0, 1]$ and let $\widehat{y} \in \mathcal{A}$ correspond to the predicted probability that $y = 1$. Calibration is a common requirement on predictors, necessitating their predictions to be unbiased conditioned on the predicted value. For technical reasons such as dealing with the fact that predicted values can take any real values, calibration is more conveniently defined by considering buckets of predicted values. Formally, we define a set of buckets $V_\lambda = \{0, 1/\lambda, 2/\lambda, \ldots, 1\}$ and say that prediction $\widehat{y}$ belongs to bucket v (denoted by $\widehat{y} \in v$) when $|\widehat{y} - v| \le 1/2\lambda$. Then, the calibration error of a sequence of predictors $p_1, \ldots, p_T$ on instance-outcome pairs $(x_1, y_1), \ldots, (x_T, y_T)$ is defined by $\max_{v\in V_\lambda} \left| \sum_{t=1}^{T} \mathbb{1}\left[p_t(x_t) \in v\right] \cdot (p_t(x_t) - y_t) \right|$. In other words, calibration loss is the cumulative ℓ_∞ norm of the objective $\ell(a, y) = [\mathbb{1}[a \in v] \cdot (a - y)]_{v\in V_\lambda}$.

We can accordingly define two variants of the calibration error that account for miscalibration as experienced by each component. In these definitions, we take $\lambda > 0$ to be fixed and clear from the context and suppress it in the notations.

In our first definition, called the *discriminant calibration error*, an instance (x, y) is purely attributed to the component $g = \arg\max_{j\in[k]} f(j \mid x, y)$ that was most likely responsible for producing (x, y).

Definition 1 (Discriminant Calibration Error)**.** *Given a sequence of instance-outcome pairs $(x_1, y_1), \cdots, (x_T, y_T)$, the* discriminant calibration error *of predicted probabilities $\widehat{y}_1, \ldots, \widehat{y}_T$ with respect to the endogenous subgroups model f — as specified by distributions $f(y|x)$, $f(x|g)$, and $f(g)$ — is defined as*

$$\mathbf{DCE}_f(\widehat{y}_{1:T}, x_{1:T}, y_{1:T}) \coloneqq \max_{g\in[k]} \max_{v\in V_\lambda} \left| \sum_{t=1}^{T} \mathbb{1}\Big[g = \arg\max_{j\in[k]} f(j \mid x_t, y_t)\Big] \cdot \mathbb{1}\left[\widehat{y}_t \in v\right] \cdot (\widehat{y}_t - y_t) \right|,$$

When predictors $p_1, \ldots, p_T$ are used for making predictions $\widehat{y}_t = p_t(x_t)$, we denote the corresponding discriminant calibration error by $\mathbf{DCE}_f(p_{1:T}, x_{1:T}, y_{1:T})$.

In our second approach, *likelihood calibration error*, (x, y) is attributed to any component g with likelihood $f(g \mid x, y)$.

Definition 2 (Likelihood Calibration Error)**.** *Given a sequence of instance-outcome pairs $(x_1, y_1), \cdots, (x_T, y_T)$, the* likelihood calibration error *of predicted probabilities $\widehat{y}_1, \ldots, \widehat{y}_T$ with respect to the endogenous subgroups model f — as specified by distributions $f(y|x)$, $f(x|g)$, and $f(g)$) — is defined as*

$$\mathbf{LCE}_f(\widehat{y}_{1:T}, x_{1:T}, y_{1:T}) \coloneqq \max_{g\in[k]} \max_{v\in V_\lambda} \left| \sum_{t=1}^{T} f(g \mid x_t, y_t) \cdot \mathbb{1}\left[\widehat{y}_t \in v\right] \cdot (\widehat{y}_t - y_t) \right|.$$

When predictors $p_1, \ldots, p_T$ are used for making predictions $\widehat{y}_t = p_t(x_t)$, we denote the corresponding likelihood calibration error by $\mathbf{LCE}_f(p_{1:T}, x_{1:T}, y_{1:T})$.

3 A FIRST ATTEMPT: CLUSTER-THEN-PREDICT

As a warmup, we consider the natural algorithmic approach: to first spend some timesteps to estimate the underlying group structure, and then provide guarantees for the estimated groups. We will focus on minimizing discriminant calibration error for a simplified problem setting, then highlight some challenges involved in extending the approach to (a) more general problem settings and (b) to minimizing likelihood calibration error.

Minimizing discriminant calibration error via cluster-then-predict. To instantiate cluster-then-predict for discriminant calibration error, we leverage two common types of algorithms. For the first phase, we use a clustering algorithm that, given a Gaussian mixture model, outputs a mapping $F : \mathcal{X} \to \{1, 2\}$ indicating the group memberships. In particular, we use the algorithm of Azizyan et al. (2013) to obtain F for which $F(x) = \arg\max_{j \in \{1,2\}} f(j \mid x)$ for all but an ε fraction of the underlying distribution, after having made $O(1/\varepsilon^2)$ observations. For the second phase, we can instantiate one (online) calibration algorithm that provides marginal calibration guarantees for its predictions on each of the two clusters. For example, Foster & Vohra (1998) guarantees at most $T^{1/2}$ calibration error.[1]

Cluster-Then-Predict Algorithm for Minimizing DCE

For the first $T' < T$ timesteps, make arbitrary predictions and collect observed features $x_1, \ldots, x_{T'}$. Apply a clustering algorithm, such as the Azizyan et al. (2013) estimator, to the observed features to partition the domain into cluster assignments $F : \mathcal{X} \to \{1, 2\}$.

Then, instantiate two calibrated prediction algorithms (e.g., the Foster & Vohra (1998) algorithm), one for each cluster. For every subsequent timestep $t = T' + 1, \ldots, T$, observe x_t and predict $\widehat{y}_t$ by applying a calibrated online forecasting algorithm to the transcript $\{(x_\tau, y_\tau) \mid T' < \tau < t, F(x_\tau) = F(x_t)\}$ consisting only of datapoints with the same predicted cluster assignment.

We formalize the guarantees of this approach in Proposition 3.1.

Proposition 3.1. *Let f be an unknown endogenous subgroups model whose Gaussian components are isotropic with $\|\mu_1 - \mu_2\| \geq \gamma$. Then, with probability $1 - \delta$, the* Cluster-Then-Predict *algorithm attains discriminant calibration error of $O(d^{1/3}T^{2/3}\gamma^{-4/3} + \sqrt{T \log(1/\delta)})$, when it is run with an appropriate choice of $T' = \Theta(d^{1/3}T^{2/3}\gamma^{-4/3})$.*

Proof Sketch. This $T^{2/3}$ rate is typical for two-stage online algorithms, such as explore-then-commit, and its proof is also similar. By Azizyan et al. (2013) (see Theorem C.1), learning a cluster assignment F that has ε error takes $T' = \Theta(d/\gamma\varepsilon^2)$ samples. Note that, the DCE of this algorithm is at most $T' + \varepsilon T + \sqrt{T \ln(1/\delta)}$, where the second term accounts for the clustering mistakes and the third term accounts for the calibration error of the "predict" stage. Setting $T' = \Theta(d^{1/3}T^{2/3}\gamma^{-4/3})$ gives the desired bound. □

Remark 3.2. *In addition to the upper bound, the result of Azizyan et al. (2013) also implies that, for this class of cluster-then-predict algorithms, the $O(T^{2/3})$ rate is in fact minimax optimal: it is impossible to learn F to any higher accuracy with any fewer samples.*

Remark 3.3. *Another consequence of the cluster-then-predict approach is that any hardness in learning cluster membership is inherited. For example, the γ-separation dependence of Proposition 3.1 is unavoidable in the task of learning cluster assignments, and by extension any instantiation of the cluster-then-predict approach.*

Beyond 2-component isotropic mixtures. For discriminant calibration error, extending these results beyond 2 components to general k-component mixtures complicates the analysis in two ways. First, the ground-truth cluster

[1] The guarantees of Foster & Vohra (1998) hold even when x_t are adversarial; an even simpler algorithm would suffice for our i.i.d case. For example, one could take the naive approach of using some additional timesteps to estimate $\mathbb{E}[y \mid g]$ to sufficient accuracy and use those estimated means for the remainder of time.

assignment functions $x \mapsto \arg\max_{i \in [k]} f(i \mid x)$ are no longer halfspaces but rather Voronoi diagrams for $k > 2$. Second, the minimax optimal accuracy rate for estimating the cluster assignment function of a k-component isotropic Gaussian mixture model is not known exactly, aside from being polynomial (Belkin & Sinha, 2015). Extending these results to mixtures of non-isotropic Gaussians or to non-uniform mixtures is also non-trivial, even for $k = 2$; minimax optimal rates are similarly unknown for these generalizations. One challenge for proving such a result is that the cluster assignment functions are no longer halfspaces, but rather non-linear boundaries, as illustrated in Figure 1.

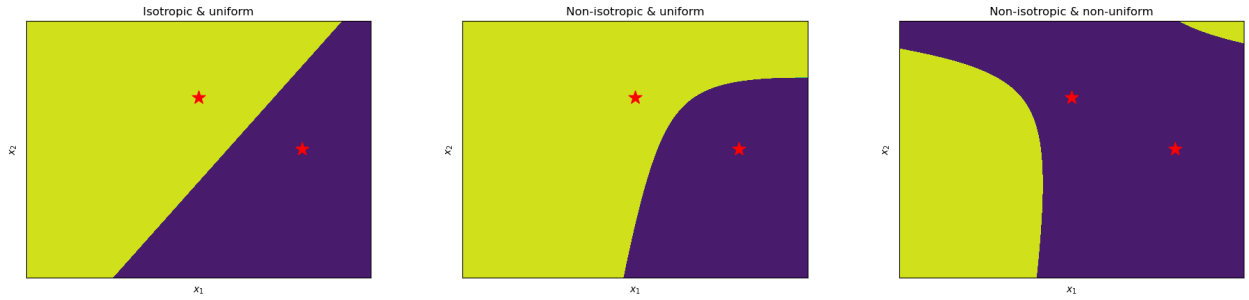

Figure 1: Illustration of the non-linear boundaries that arise in the cluster assignment functions of non-isotropic Gaussian mixtures for $k = 2$. Means of each component are marked with red stars.

Extension to likelihood calibration error. Applying the cluster-then-predict approach to minimize likelihood calibration error again provides a $T^{2/3}$ rate.

Proposition 3.4. *Let f be an unknown endogenous subgroups model whose Gaussian components are isotropic, and where μ_1 and μ_2 are separated by a constant in every dimension. Then, with probability $1 - \delta$, a* Cluster-Then-Predict Algorithm for Minimizing LCE*, setting $T' = O(T^{2/3})$, incurs likelihood calibration error of $\widetilde{O}\left(T^{2/3}\sqrt{d \log(d/\delta)}\right)$.*

We defer the proof of Proposition 3.4, and the statement of the corresponding algorithm, to Appendix C.

Implementing this approach requires some additional care. In the first phase, we must learn a good likelihood function for each cluster (i.e., $f(g \mid x)$), rather than a cluster assignment function. This can be done by estimating the parameters of each component, then using those estimates to construct likelihood functions; to that end, we can apply existing parameter learning algorithms (e.g. Hardt & Price (2015)). A more significant challenge is that even if a good estimator $\widehat{f}(g \mid x)$ is known, the fact that group membership is real-valued means that we can no longer partition the space and independently calibrate predictions in each partition. Instead, our predictions must handle the fact that each x_t belongs to multiple groups; this motivates the use of online calibration algorithms with multi-group guarantees, called *multicalibration*. For similar reasons, we apply multicalibration algorithms in our multi-objective approach, which we discuss in the following section.

4 Improved Bounds: A Multi-Objective Approach

The cluster-then-predict approach studied in Section 3 necessitates learning the exact subpopulation/cluster structure that underlies the data distribution—that is, learning the binary functions $x \mapsto \arg\max_g f(g \mid x)$ or the conditional likelihoods $f(g \mid x)$ to high accuracy. As an alternative to resolving the underlying structure explicitly, we consider a multi-objective approach where we aim to simultaneously provide subgroup guarantees for a representative uncertainty set—specifically a covering—of all possible subpopulation structures.

Building a covering of possible underlying cluster structures is significantly easier in a statistical sense than learning the true structure directly, which offers two benefits. First, rather than paying the $T^{2/3}$ error rate typical of cluster-then-predict methods, the multi-objective approach provides an optimal $T^{1/2}$ error rate. Second, rather than paying the

inevitable mean separation dependence involved in learning discriminant or likelihood functions, the multi-objective approach provides error rates independent of separation.

In Section 4.2, we demonstrate the key technical result enabling the multi-objective approach for likelihood calibration error: an upper bound on the covering number of cluster likelihood functions. In Section 4.3, we extend the result by showing how to convert a covering of the likelihood functions to a covering of the discriminant functions.

4.1 Simultaneous guarantees under multiple clustering schemes

In this section, we use multicalibration algorithms to simultaneously provide per-subgroup guarantees for multiple hypothetical clustering schemes, a multi-objective approach that does not require resolving the underlying clustering scheme. Multicalibration is a refinement of calibration that requires unbiasedness of predictions not only on distinguishers resolving the level sets of one's predictors, i.e. $\{\mathbb{1}[p_t(x_t) \in v]\}_{v \in V_\lambda}$, but also on distinguishers that identify parts of the domain. We use an adaptation of the original multicalibration definition for real-valued distinguishers of the domain.

Definition 3 (Multicalibration Error)**.** *Given a sequence of instance-outcome pairs $(x_1, y_1), \ldots, (x_T, y_T)$ and class of distinguishers $\mathcal{G} \subset [0,1]^{\mathcal{X}}$, the* multicalibration error *of predicted probabilities $\widehat{y}_1, \ldots, \widehat{y}_T$ with respect to $\mathcal{G}$ is*

$$\mathbf{MCE}(\widehat{y}_{1:T}, x_{1:T}, y_{1:T}; \mathcal{G}) \coloneqq \max_{g \in \mathcal{G}} \max_{v \in V_\lambda} \left| \sum_{t \in [T]} g(x_t) \cdot \mathbb{1}[\widehat{y}_t \in v] \cdot (\widehat{y}_t - y_t) \right|.$$

When predictors $p_1, \ldots, p_T$ are used for making predictions $\widehat{y}_t = p_t(x_t)$, we denote the corresponding multicalibration error by $\mathbf{MCE}(p_{1:T}, x_{1:T}, y_{1:T}; \mathcal{G})$.

To show how multicalibration can be useful, we first establish that we can upper bound likelihood calibration error and discriminant calibration error in terms of multicalibration error for carefully-designed classes of distinguishers.

Fact 4.1. *Let $\mathcal{F}$ be the class of all endogenous subgroups models considered in our setting. Take $f \in \mathcal{F}$ to be an arbitrary endogenous subgroups model and $p_1, \ldots, p_T$ to be an arbitrary sequence of predictors. Then, $\mathbf{DCE}_f(p_{1:T}, x_{1:T}, y_{1:T}) \le \mathbf{MCE}(p_{1:T}, x_{1:T}, y_{1:T}; \mathcal{G})$ defined for any set of distinguishers $\mathcal{G}$ where $\mathcal{G} \supseteq \{x \mapsto \mathbb{1}[g = \arg\max_j f(j \mid x)] \mid g \in [k]\}$. One such choice of $\mathcal{G}$ is*

$$\mathcal{G} = \{x \mapsto \mathbb{1}[g = \arg\max_j f'(j \mid x)] \mid g \in [k], f' \in \mathcal{F}\}. \tag{1}$$

Similarly, $\mathbf{LCE}_f(p_{1:T}, x_{1:T}, y_{1:T}) \le \mathbf{MCE}(p_{1:T}, x_{1:T}, y_{1:T}; \mathcal{G})$, for any set of distinguishers $\mathcal{G}$ where $\mathcal{G} \supseteq \{x \mapsto f(g \mid x) \mid g \in [k]\}$. One such choice of $\mathcal{G}$ is

$$\mathcal{G} = \{x \mapsto f'(g \mid x) \mid g \in [k], f' \in \mathcal{F}\}. \tag{2}$$

For any (potentially infinite) set of real-valued distinguishers $\mathcal{G}$ with finite covering number, the following algorithm achieves $O(\sqrt{T})$ multicalibration error (Proposition 4.2): efficiently cover the space of distinguishers and run a standard online multicalibration algorithm on the cover. In Appendix D, we give explicit example algorithms for each stage—the covering stage (Algorithm 1) and the calibration stage (Algorithm 2).

Online Multicalibration Algorithm for Coverable Distinguishers

For the first $T' = \sqrt{T \log(N_1(\varepsilon/8, \mathcal{G}, 2T))}$ timesteps, make arbitrary predictions and collect observed features $x_1, \ldots, x_{T'}$ and compute a small ε-covering $\mathcal{G}'$ of the set $\mathcal{G}$ on $x_1, \ldots, x_{T'}$, e.g. using Alg. 1.

For the remaining timesteps $t = T' + 1, \ldots, T$, observe x_t and predict y_t by applying any online multicalibration algorithm, e.g. Alg. 2, to the transcript of previously seen datapoints $\{(x_\tau, y_\tau) \mid \tau < t\}$ and distinguishers $\mathcal{G}'$.

In this section, we use $N_1(\varepsilon, \mathcal{G}, m)$ to denote the covering number of $\mathcal{G}$ on m datapoints in ℓ_1-norm with distance ε. Proposition 4.2 bounds the error incurred by the Online Multicalibration Algorithm for Coverable Distinguishers.

Proposition 4.2. *For any real-valued function class $\mathcal{G}$ whose covering number grows sub-square root exponentially, i.e. $N_1(1/96T, \mathcal{F}, 2T) \leq \sqrt{\frac{1}{8}\exp(1/32T)}$, with probability $1-\delta$, the predictions $p_{1:T}$ made by the Online Multicalibration Algorithm for Coverable Distinguishers satisfy* $\mathbf{MCE}(p_{1:T}, x_{1:T}, y_{1:T}; \mathcal{G}) \leq \widetilde{O}\left(\sqrt{T\log(N_1(\varepsilon, \mathcal{G}, T)\lambda/\delta)}\right)$.

The key technical ingredient in the proof of Proposition 4.2 is the following lemma, which states that an empirical cover computed on a finite number of samples is a true ε-cover with high probability. We defer its proof to Appendix D.2.

Lemma 4.3. *Given $\varepsilon, \delta, T > 0$ and a real-valued function class $\mathcal{F}$ where $N_1(\varepsilon/96, \mathcal{F}, 2T) \leq \sqrt{\frac{1}{8}\exp(T\varepsilon^2/32)}$, consider any ε-cover $\mathcal{F}'$ of $\mathcal{F}$ computed on T random datapoints. Then $\mathcal{F}'$ is a 4ε-cover of $\mathcal{F}$ with probability at least $1 - O(N_1\left(\frac{\varepsilon}{8}, \mathcal{F}, 2T\right)^2\exp(-T\varepsilon))$.*

Proof of Proposition 4.2. We will begin by analyzing the multicalibration error incurred on timesteps $T'+1 \ldots T$. Abusing notation, for any $g \in \mathcal{G}$, let $g' \coloneqq \arg\inf_{\in\mathcal{G}} \frac{1}{T'}\sum_{t\in[T']} |g'(x_t) - g(x_t)|$. Such a g' always exists, because $\mathcal{G}'$ was computed using $x_{1:T'}$. Then, by the triangle inequality, we have that for transcript $H = (p_{T'+1:T}, x_{T'+1:T}, y_{T'+1:T})$:

$$\mathbf{MCE}(H;\mathcal{G}) \leq \underbrace{\sup_{g\in\mathcal{G}}\max_{v\in V_\lambda}\sum_{t\in[T':T]} |g(x_t) - g'(x_t)| \cdot \mathbb{1}[\widehat{y}_t \in v] \cdot |\widehat{y}_t - y_t|}_{(A)} + \underbrace{\mathbf{MCE}(H;\mathcal{G}')}_{(B)}. \tag{3}$$

For (B), since the distinguisher class covering is small, specifically $|\mathcal{G}'| \leq N_1(\varepsilon, \mathcal{G}, T')$, standard multicalibration algorithm analysis gives that with probability $1-\delta_1$, we have $(B) \leq \sqrt{\log(N_1(\varepsilon, \mathcal{G}, T')\lambda/\delta_1)(T-T')}$ (Haghtalab et al. (2023) Theorem 4.3; see Theorem D.1).

To bound (A), we would like to apply Lemma 4.3; to do so, we will apply Hoeffding's inequality on the random variables $|g(x_t) - g'(x_t)|$. In particular, note that we can trivially bound $\mathbb{1}[\widehat{y}_t \in v] \cdot |\widehat{y}_t - y_t| \leq 1$ for all t; therefore, $(A) \leq \sup_{g\in\mathcal{G}}\sum_{t\in[T':T]} |g(x_t) - g'(x_t)|$, and Hoeffding's inequality gives us that with probability $1-\delta_2$,

$$\sup_{g\in\mathcal{G}}\sum_{t\in[T':T]} |g(x_t) - g'(x_t)| \leq (T-T')\sup_{g\in\mathcal{G}}\mathbb{E}[|g(x) - g'(x)|] + O(\sqrt{(T-T')\log(1/\delta_2)}).$$

Now, by Lemma 4.3, we have that with probability $1-\delta_2$,

$$\sup_{g\in\mathcal{G}}\mathbb{E}[|g(x) - g'(x)|] \leq \tfrac{8}{T'}\log\left(N_1(\tfrac{\varepsilon}{8}, \mathcal{G}, 2T')/\delta_3\right).$$

Therefore, with probability $1-\delta_2-\delta_3$,

$$\sup_{g\in\mathcal{G}}\sum_{t\in[T':T]} |g(x_t) - g'(x_t)| \leq \tfrac{4(T-T')}{T'}\log\left(N_1(\tfrac{\varepsilon}{8}, \mathcal{G}, 2T')\right) + O(\sqrt{(T-T')\log(1/\delta_2)}).$$

Combining this with (3) and choosing $\delta_1, \delta_2, \delta_3 = \Theta(\delta)$, we have that with sufficiently large T', with probability $1-\delta$,

$$\mathbf{MCE}(p_{T':T}, x_{T':T}, y_{T':T}; \mathcal{G}) \leq O\left(\tfrac{T}{T'}\log\left(N_1(\tfrac{\varepsilon}{8}, \mathcal{G}, 2T')/\delta\right) + \sqrt{T\log(N_1(\varepsilon, \mathcal{G}, T')\lambda/\delta)}\right).$$

The statement follows from noting that $\mathbf{MCE}(p_{1:T'}, x_{1:T'}, y_{1:T'}; \mathcal{G}) \leq T'$ and choosing $T' = \sqrt{T\log\big(N_1(\tfrac{\varepsilon}{8}, \mathcal{G}, 2T)\big)}$. □

4.2 Minimizing likelihood calibration error

We now turn to instantiating our multi-objective approach for minimizing likelihood calibration error: run the online multicalibration algorithm on the distinguisher class $\mathcal{G}$ as defined in Eq. 2, which consists of the group likelihood functions for all possible underlying cluster structures. The online multicalibration algorithm obtains a covering of the distinguisher class—selecting a small uncertainty set of cluster structures to simultaneously provide guarantees for—then runs a standard multicalibration algorithm.

Theorem 4.4. *With probability $1-\delta$, the Online Multicalibration Algorithm for Coverable Distinguishers, run with $\mathcal{G}$ as defined in Eq. 2, attains likelihood calibration error of $O\big(\sqrt{T(d^2k\log(T)\log^3(d^2k)+\log(\lambda/\delta))}\big)$.*

The multi-objective algorithm thus evades the separation dependence of cluster-then-predict (Proposition 3.4) while also obtaining the optimal $\widetilde{O}(\sqrt{T})$ rate. Notably, this error rate is significantly better than the best known rates for learning multivariate Gaussian likelihood functions: Hardt & Price (2015) prove an ε^{-12} lower bound for learning a mixture of two Gaussians, which would suggest a corresponding $T^{11/12}$ rate; a $T^{1/2}$ rate would have only been achievable under the separation assumption of Proposition 3.4. The key technical fact that our analysis relies on is that the combinatorial complexity of Gaussian clustering schemes is bounded and small. This is summarized by the following lemma, which bounds the pseudodimension of $\mathcal{G}$.

Lemma 4.5. *Let $\mathcal{F}$ be the set of possible generative models in our class. The real-valued hypothesis class $\mathcal{G}=\{x\mapsto f'(g\mid x)\mid g\in[k], f'\in\mathcal{F}\}$ has a pseudodimension of at most $(kd(d+1)+k)\log^2(kd(d+1)+k)$.*

Proof. Without loss of generality, fix $g=1$. For any $f\in\mathcal{F}$, we can write $f(1\mid x)$ as:

$$f(1\mid x)=\frac{f(x\mid 1)}{\sum_{j\in[k]}f(x\mid j)}=\frac{1}{\sum_{j\in[k]}\exp(-(x-\mu_j)^\top\Sigma_j(x-\mu_j)+(x-\mu_1)^\top\Sigma_1(x-\mu_1))}.$$

To bound the cover size of $\mathcal{G}$, we will analyze the building blocks of this function one at a time, and utilize convenient properties of composition for pseudodimension and covering numbers.

Fix any $x\in\mathbb{R}^d$. Consider the set of $d(d+1)$ elements $S:=\{x_ix_j\}_{i,j\in[d]}\bigcup\{x_i\}_{i\in[d]}$ and corresponding function class $\mathcal{H}:=\{x^\top(\Sigma-\Sigma')x-x^\top\Sigma\mu+x^\top\Sigma'\mu'+\mu^\top\Sigma\mu-\mu'^\top\Sigma'\mu'\mid\Sigma,\Sigma'\in\mathbb{R}^{d\times d},\mu,\mu'\in\mathbb{R}^d\}$. Then, $\text{Pdim}(\mathcal{H})\le d(d+1)+1$ (Fact B.2). Furthermore, monotonic functions preserve pseudodimension (Fact B.3); since the exp function is monotonic, $\text{Pdim}(\{\exp(h)\mid h\in\mathcal{H}\})\le d(d+1)+1$. Similarly, since $x\mapsto 1/x$ is also monotonic, $\text{Pdim}(\{1/\exp(h)\mid h\in\mathcal{H}\})\le d(d+1)+1$. Finally, we also have that the sum of bounded pseudodimension classes enjoy an almost linear bound in pseudodimension (Attias & Kontorovich (2024) Thm. 1; Fact B.4); thus, $\text{Pdim}(\{\sum_{i\in[k]}1/\exp(h_i)\mid h_1,\dots,h_k\in\mathcal{H}\})\le(kd(d+1)+k)\log^2(kd(d+1)+k)$. □

Proof of Theorem 4.4. By Lemma 4.5, $\text{Pdim}(\mathcal{G})\le(kd(d+1)+k)\log^2(kd(d+1)+k)$. Therefore, $N_1(\varepsilon,\mathcal{G},T)\le O\left(\text{Pdim}(\mathcal{G})(1/\varepsilon)^{\text{Pdim}(\mathcal{G})}\right)\le O(d^2k\log^2(d^2k)(1/\varepsilon)^{d^2k\log^2(d^2k)})$ (Anthony & Bartlett (1999) Thm. 18.4; see Fact B.1). The statement of the theorem follows from Proposition 4.2 combined with Fact 4.1, and choosing $\varepsilon=1/T$. □

4.3 Minimizing discriminant calibration error

We can also instantiate the multi-objective approach for minimizing discriminant calibration error to realize a similar $O(\sqrt{T})$ rate. In this case, we apply online multicalibration to the class of distinguishers $\mathcal{G}$ as defined in Eq. 1 to consist of the group discriminant functions for all possible underlying cluster structures.

Theorem 4.6 formalizes the guarantee of this approach when cluster centers are well-separated. In contrast to the Cluster-then-Predict guarantee of Proposition 3.1, Theorem 4.6 obtains an improved error rate of $\widetilde{O}(\sqrt{T})$ versus $\widetilde{O}(T^{2/3})$ and does not require isotropic assumptions.

Within this section, we will call a set of likelihood functions $\{f'(g\mid x)\mid g\in[k]\}$ *legal* if there exists a Gaussian mixture model giving rise to the likelihood functions, i.e., for all $x\in\mathcal{X}$, we have $\sum_{g\in[k]}f'(g\mid x)=1$ and $f'(g\mid x)\in[0,1]$. Then, given a legal combination of functions of form $\{f'(g\mid x)\mid g\in[k]\}$, we write $F_{f'}(x)=\mathbb{1}[1=\arg\max_{i\in[k]}f'(i\mid x)]$ to denote the corresponding arg-max function for $g=1$.

Theorem 4.6. *Assume covariance matrices $\Sigma_1,\dots,\Sigma_k$ have constant eigenvalues and that, for all $i\neq j$, $\|\mu_i-\mu_j\|\ge\gamma=\Omega(\sqrt{d+\log(k+T)})$. Then, with probability $1-\delta$, the Online Multicalibra-*

tion Algorithm for Coverable Distinguishers, with $\mathcal{G}$ as in Eq. 1, attains discriminant calibration error of $O(\sqrt{T(\exp(-\gamma)d^2k^2\log^3(d^2k)\log(T)+\log(\lambda/\delta))})$.

The statement of Theorem 4.6 follows from Proposition 4.2, choosing $\varepsilon = 1/T$, and the following bound on the covering number of discriminant functions.

Lemma 4.7. *Let $\mathcal{F}$ be the set of possible generative models in our class. The binary-valued hypothesis class $\mathcal{G} = \{x \mapsto \mathbb{1}[g = \arg\max_{g' \in [k]} f'(g' \mid x)] \mid g \in [k], f' \in \mathcal{F}\}$ has a covering number of $N_1(\varepsilon, \mathcal{G}, m) \in O((d^2k\log^2(d^2k)(k/\gamma\varepsilon)^{d^2k\log^2(d^2k)})^k)$ under the assumptions of Theorem 4.6.*

When cluster centers are not well-separated, we can still obtain a similar rate under isotropic assumptions.

Theorem 4.8. *Let f be an unknown discrete hidden-state generative model whose Gaussian components are isotropic and equally weighted. With probability $1-\delta$, the Online Multicalibration Algorithm for Coverable Distinguishers, run with $\mathcal{G}$ as defined in Eq. 1, attains discriminant calibration error of $O\left(\sqrt{T\left(dk^2\log(T)\log(k)+\log(\lambda/\delta)\right)}\right)$.*

Gap between learning sub-group guarantees and learning subgroups. Theorem 4.8 provides an $O(\sqrt{T})$ bound independent of component mean separation, even though learning the subgroup discriminator function requires sample complexity that scales with component mean separation.

Theorem 4.9 (Azizyan et al. (2013), Theorem 2)**.** *Let f be an unknown endogenous subgroups model with two isotropic Gaussian components with $d \geq 9$. The sample complexity of learning the cluster assignment function is $\Omega(\frac{d}{\varepsilon^2\gamma^6})$.*

Notably, this lower bound holds even for the case of having two equal weight isotropic Gaussians. This sample complexity is paid for, for example, in the first stage of the Cluster-then-Predict approach. That Theorem 4.8 obtains a better bound highlights the surprising fact that *providing clusterable group guarantees can be easier than clustering*, lending further motivation to multi-objective approaches over cluster-then-predict.

5 Discussion

This work has focused on a particular instantiation of our model—for (online) calibration as the objective, and Gaussian mixtures as the subgroup structure underlying our endogenous subgroups model. However, as discussed in Section 2, the results presented in this paper for calibration extend to other problems that can be formalized in the language of Blackwell approachability, such as online conformal prediction. Another extension is to handle other families of subgroup structures beyond Gaussian mixture models. Many of our results rely on known analyses of Gaussian densities, but in principle, similar technical analysis can be performed for other unsupervised learning models of bounded combinatorial complexity. In fact, our approaches—and notions of discriminant calibration error and likelihood calibration error—can apply to any setting where group membership can at best be estimated.

More generally, our formalization of an unsupervised notion of multi-group guarantees provides a language for understanding an important downstream application of clustering. Our results demonstrate that being intentional about how learned clusters will be used, rather than treating clustering and learning as distinct stages, is significant both conceptually and for attaining optimal theoretical rates. First, resolving the exact clustering structure of one's data is inefficient, and results in the same theoretical sub-optimality as explore-then-commit algorithms in bandit/reinforcement learning literature—namely, $O(T^{2/3})$ rather than $O(T^{1/2})$ rates. Second, the task of learning with guarantees for subgroups can be surprisingly easier than learning the subgroups themselves. The most striking example of this appears in our results for discriminant calibration error, for which we show that learning cluster assignment functions has an inevitable dependence on cluster separation, whereas separation can be ignored when pursuing per-cluster guarantees. Moreover, as our improved rates for the multi-objective approach suggest, it is not just that learning subgroups may be harder: it is also not necessary to learn the subgroups exactly if the ultimate goal is to provide guarantees across them.

REFERENCES

Martin Anthony and Peter L. Bartlett. *Neural Network Learning: Theoretical Foundations*. Cambridge University Press, 1999.

Jamil Arbas, Hassan Ashtiani, and Christopher Liaw. Polynomial time and private learning of unbounded gaussian mixture models. In *International Conference on Machine Learning*, pp. 1018–1040. PMLR, 2023.

Idan Attias and Aryeh Kontorovich. Fat-shattering dimension of k-fold aggregations. *Journal of Machine Learning Research*, 25(144):1–29, 2024.

Pranjal Awasthi, Matthäus Kleindessner, and Jamie Morgenstern. Equalized odds postprocessing under imperfect group information. In *International conference on artificial intelligence and statistics*, pp. 1770–1780. PMLR, 2020.

Pranjal Awasthi, Alex Beutel, Matthäus Kleindessner, Jamie Morgenstern, and Xuezhi Wang. Evaluating fairness of machine learning models under uncertain and incomplete information. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pp. 206–214, 2021.

Martin Azizyan, Aarti Singh, and Larry Wasserman. Minimax Theory for High-dimensional Gaussian Mixtures with Sparse Mean Separation, June 2013.

Mikhail Belkin and Kaushik Sinha. Polynomial learning of distribution families. *SIAM J. Comput.*, 44(4):889–911, 2015. doi: 10.1137/13090818X. URL `https://doi.org/10.1137/13090818X`.

Sebastian Benthall and Bruce D Haynes. Racial categories in machine learning. In *Proceedings of the conference on fairness, accountability, and transparency*, pp. 289–298, 2019.

David Blackwell. An analog of the minimax theorem for vector payoffs. *Pacific Journal of Mathematics*, 6(1):1 – 8, 1956.

Avrim Blum, Nika Haghtalab, Ariel D Procaccia, and Mingda Qiao. Collaborative pac learning. *Advances in Neural Information Processing Systems*, 30, 2017.

Jiahao Chen, Nathan Kallus, Xiaojie Mao, Geoffry Svacha, and Madeleine Udell. Fairness under unawareness: Assessing disparity when protected class is unobserved. In *Proceedings of the conference on fairness, accountability, and transparency*, pp. 339–348, 2019.

A Philip Dawid. The well-calibrated Bayesian. *Journal of the American Statistical Association*, 77(379):605–610, 1982.

Luc Devroye, László Györfi, and Gábor Lugosi. *A probabilistic theory of pattern recognition*, volume 31. Springer Science & Business Media, 2013.

Evan Dong, Aaron Schein, Yixin Wang, and Nikhil Garg. Addressing Discretization-Induced Bias in Demographic Prediction, May 2024.

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

Elizabeth Mitchell Elder and Matthew Hayes. Signaling race, ethnicity, and gender with names: Challenges and recommendations. *The Journal of Politics*, 85(2):764–770, 2023.

Marc N Elliott, Peter A Morrison, Allen Fremont, Daniel F McCaffrey, Philip Pantoja, and Nicole Lurie. Using the census bureau's surname list to improve estimates of race/ethnicity and associated disparities. *Health Services and Outcomes Research Methodology*, 9:69–83, 2009.

Dean P. Foster and Sham M. Kakade. Calibration via regression. In Gadiel Seroussi and Alfredo Viola (eds.), *Proceedings of 2006 IEEE Information Theory Workshop*, pp. 82–86. IEEE, 2006.

Dean P. Foster and Rakesh V. Vohra. Asymptotic calibration. *Biometrika*, 85(2):379–390, 1998. ISSN 00063444, 14643510. URL http://www.jstor.org/stable/2337364.

Parikshit Gopalan, Adam Tauman Kalai, Omer Reingold, Vatsal Sharan, and Udi Wieder. Omnipredictors. In Mark Braverman (ed.), *Proceedings of the ACM Conference on Innovations in Theoretical Computer Science Conference (ITCS)*, pp. 79:1–79:21, 2022.

Varun Gupta, Christopher Jung, Georgy Noarov, Mallesh M Pai, and Aaron Roth. Online multivalid learning: Means, moments, and prediction intervals. *arXiv preprint arXiv:2101.01739*, 2021.

Varun Gupta, Christopher Jung, Georgy Noarov, Mallesh M. Pai, and Aaron Roth. Online multivalid learning: Means, moments, and prediction intervals. In Mark Braverman (ed.), *Proceedings of the ACM Conference on Innovations in Theoretical Computer Science Conference (ITCS)*, pp. 82:1–82:24. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2022.

Nika Haghtalab, Michael Jordan, and Eric Zhao. On-demand sampling: Learning optimally from multiple distributions. *Advances in Neural Information Processing Systems*, 35:406–419, 2022.

Nika Haghtalab, Michael Jordan, and Eric Zhao. A unifying perspective on multi-calibration: Game dynamics for multi-objective learning. *Advances in Neural Information Processing Systems*, 36, 2023.

Moritz Hardt and Eric Price. Tight bounds for learning a mixture of two gaussians. In *Proceedings of the forty-seventh annual ACM symposium on Theory of computing*, pp. 753–760, 2015.

Sergiu Hart. Calibrated forecasts: The minimax proof, 2022.

Tatsunori Hashimoto, Megha Srivastava, Hongseok Namkoong, and Percy Liang. Fairness without demographics in repeated loss minimization. In *International Conference on Machine Learning*, pp. 1929–1938. PMLR, 2018.

David Haussler and Emo Welzl. Epsilon-nets and simplex range queries. In *Proceedings of the second annual symposium on Computational geometry*, pp. 61–71, 1986.

Úrsula Hébert-Johnson, Michael P. Kim, Omer Reingold, and Guy N. Rothblum. Calibration for the (Computationally-Identifiable) Masses, March 2018.

Lily Hu. What is "race" in algorithmic discrimination on the basis of race? *Journal of Moral Philosophy*, 1(aop):1–26, 2023.

Lily Hu and Issa Kohler-Hausmann. What's Sex Got To Do With Fair Machine Learning? In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pp. 513–513, January 2020. doi: 10.1145/3351095.3375674.

Christopher Jung, Changhwa Lee, Mallesh M. Pai, Aaron Roth, and Rakesh Vohra. Moment multicalibration for uncertainty estimation. In Mikhail Belkin and Samory Kpotufe (eds.), *Proceedings of the Conference on Learning Theory (COLT)*, Proceedings of Machine Learning Research, pp. 2634–2678. PMLR, 2021.

Christopher Jung, Georgy Noarov, Ramya Ramalingam, and Aaron Roth. Batch multivalid conformal prediction. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. URL https://openreview.net/forum?id=Dk7QQp8jHEo.

Nathan Kallus, Xiaojie Mao, and Angela Zhou. Assessing algorithmic fairness with unobserved protected class using data combination. *Management Science*, 68(3):1959–1981, 2022.

Sonia K Kang, Katherine A DeCelles, András Tilcsik, and Sora Jun. Whitened résumés: Race and self-presentation in the labor market. *Administrative science quarterly*, 61(3):469–502, 2016.

Preethi Lahoti, Alex Beutel, Jilin Chen, Kang Lee, Flavien Prost, Nithum Thain, Xuezhi Wang, and Ed Chi. Fairness without demographics through adversarially reweighted learning. *Advances in neural information processing systems*, 33:728–740, 2020.

Daniel Lee, Georgy Noarov, Mallesh M. Pai, and Aaron Roth. Online minimax multiobjective optimization: Multicalibeating and other applications. In Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022. URL http://papers.nips.cc/paper_files/paper/2022/hash/ba942323c447c9bbb9d4b638eadefab9-Abstract-Conference.html.

David Liu, Virginie Do, Nicolas Usunier, and Maximilian Nickel. Group fairness without demographics using social networks. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, pp. 1432–1449, 2023.

Anaelia Ovalle, Arjun Subramonian, Vagrant Gautam, Gilbert Gee, and Kai-Wei Chang. Factoring the matrix of domination: A critical review and reimagination of intersectionality in ai fairness. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 496–511, 2023.

Angelina Wang, Vikram V Ramaswamy, and Olga Russakovsky. Towards intersectionality in machine learning: Including more identities, handling underrepresentation, and performing evaluation. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pp. 336–349, 2022.

Serena Wang, Wenshuo Guo, Harikrishna Narasimhan, Andrew Cotter, Maya Gupta, and Michael Jordan. Robust optimization for fairness with noisy protected groups. *Advances in neural information processing systems*, 33: 5190–5203, 2020.

Kyra Wilson and Aylin Caliskan. Gender, race, and intersectional bias in resume screening via language model retrieval. *arXiv preprint arXiv:2407.20371*, 2024.

Yan Zhang. Assessing fair lending risks using race/ethnicity proxies. *Management Science*, 64(1):178–197, 2018.

A Additional motivation and related work

One reason to think of group membership in context of the population rather than as deterministic functions of individual features is a normative provocation. From a modeling perspective, one common critique of standard practice is that observable (demographic) features are only approximations of more complex phenomena that are related to—but not directly causal of—shared life experience. Therefore, demanding "equal performance" across rigid (demographic) categories does not necessarily imply "fairness" in a normative sense (see e.g. Benthall & Haynes (2019); Hu & Kohler-Hausmann (2020); Hu (2023) for more extended discussion). In some sense, our approach can be seen as an attempt to develop a more constructivist perspective on defining subpopulations—placing individuals in context with others for whom those predictions are made, and allowing group definitions to vary based on the particular prediction task—as opposed to an essentialist one. Therefore, while the (U.S.) legal system relies on such categories to instantiate concrete discussions of discrimination, those categories are not necessarily the only or most salient ways to understand a population. Of course, we cannot claim to fully resolve these normative challenges or realize these goals; at the very least, however, we think of them as a reason to explore different ways of understanding the relationship between groups and individuals.

On the other hand, we argue that the subgroups we consider have a natural correspondence to those subgroups which have practical significance. Because we cannot determine group membership solely based on an individual's feature vector, our problem setting requires some structure on the domain; in particular, features must be clusterable. Then, if all one initially knows about the population is that it is comprised of multiple subpopulations where group membership affects feature realizations, determining subgroup membership based only on those realizations is the best one can expect to do. Our focus on statistically identifiable groups is in contrast with the computationally-identifiable groups studied when subpopulations are defined as combinations of feature values. In those settings, it is necessary to ensure that membership can be distinguished as *efficiently* as possible (e.g., through low circuit complexity, as multicalibration was initially described in Hébert-Johnson et al. (2018)); in our setting, the key challenge is to instead identify membership as *accurately* as possible, because group membership itself is uncertain.

Of course, for our setting, accuracy in our model can be formalized either in the "discriminant" sense or the "likelihood" sense. When might one prefer one over another? It is clear that, when all groups are well-separated, discriminant calibration error and likelihood calibration error are approximately equivalent; furthermore, in these cases, there is no meaningful uncertainty in group membership. Our model is therefore most salient exactly when the Gaussian components of the endogenous subgroups model are *not* always well-separated. In this case, likelihood calibration error is advantageous over discriminant calibration error for several reasons. For one, likelihood calibration error handles underrepresented groups more gracefully than discriminant calibration error; more generally, likelihood calibration error captures uncertainty in group membership at a more granular level than discriminant calibration error which is necessarily coarser. Our technical results also point to preferring likelihood calibration error—Theorem 4.4 does not require the same assumptions as 4.6 (separation) and 4.8 (isotropic).

We also note that finding statistically identifiable subpopulations from data (in the sense of learning membership likelihoods), and using those subgroups downstream, is not a new idea. When true subgroup labels are unknown, inferring or estimating group membership is a natural (and sometimes even necessary) approach. For example, it is well-known that names are often associated with demographic identity (Elder & Hayes, 2023), and audits of resume screening systems in practice often use those assumed associations rather than explicitly-stated demographic identity (e.g. Kang et al. (2016); Wilson & Caliskan (2024)). More generally, an extensive literature discusses how demographic labels might be imputed from data—e.g., name and census tract, in the well-known "BISG" (Bayesian Improved Surname Geocoding) approach (Elliott et al., 2009) and its variants; how those labels might be used for downstream purposes (e.g. auditing lending decisions (Zhang, 2018)); and how those estimates ought to be incorporated in a mathematical sense to those downstream applications (e.g., (Dong et al., 2024) and others).

Finally, though intersectionality *per se* is not the focus of our work, our model of subpopulations raises interesting questions in this direction. For instance, one suggestion in Wang et al. (2022) is to explicitly consider intrinsic structure in covariates across groups. More generally, to the extent that the notion of power is central to what "defines" a group (Ovalle et al., 2023), this appears more clearly in our model of subgroups.

B FACTS, REFERENCES, AND RESTATEMENTS

Fact B.1 (Anthony & Bartlett (1999), Theorem 18.4)**.** *Let $\mathcal{G}$ be a real-valued function class. Then $N_1(\varepsilon, \mathcal{G}, m) \leq O\left(Pdim(\mathcal{G})(1/\varepsilon)^{Pdim(\mathcal{G})}\right)$. Note that, if $\mathcal{G}$ is a binary function, the bound holds for VC($\mathcal{G}$).*

Fact B.2 (Anthony & Bartlett (1999), Theorem 11.4)**.** *The pseudodimension of a k-dimensional vector space of real valued functions is $k+1$.*

Fact B.3 (Anthony & Bartlett (1999), Theorem 11.3)**.** *Let f be a monotonic function and $\mathcal{G}$ be a real-valued function class with pseudodimension d. Then the pseudodimension of $\{f \circ g \mid g \in \mathcal{G}\}$ is d.*

Fact B.4 (Attias & Kontorovich (2024), Theorem 1)**.** *Let $\mathcal{F}_1, \ldots, \mathcal{F}_k$ be real-valued function classes each with a pseudodimension of d. Then the pseudodimension of $\{\sum_i f_i \mid f_1 \in \mathcal{F}_1, \ldots, f_k \in \mathcal{F}_k\}$ is $kd\log^2(kd)$.*

Fact B.5 (Devroye et al. (2013), Theorem 21.5)**.** *The VC dimension of the class of k-cell Voronoi diagrams in $\mathbb{R}^d$ is upper bounded by $k + (d+1)k^2 \log k$.*

C PROOFS FOR SECTION 3

C.1 PROPOSITION 3.1

Let $\Psi_\gamma = \{\mu_1, \mu_2 \in \mathcal{X} \mid \|\mu_1 - \mu_2\| \geq \gamma\}$ denote the space of possible component means with at least γ separation. Let $\mathcal{F}_n$ be the class of all mixture model estimators; formally, we define $\mathcal{F}_n$ as the set of all functions mapping from n-length datasets $(\mathcal{X})^n$ to functions $\{1,2\}^{\mathcal{X}}$. Azizyan et al. (2013) provides an estimator for the Gaussian mixture model that achieves the minimax optimal error rate, with a guarantee as follows.

Theorem C.1 (Minimax Gaussian clustering rates (Azizyan et al., 2013))**.** *For $n \geq \max\{68, 4d\}$, the minimax optimal accuracy for the estimator of a two-component isotropic Gaussian mixture model with separation γ is*

$$\inf_{F \in \mathcal{F}_n} \sup_{\theta \in \Psi_\gamma} \mathop{\mathbb{E}}_{\{x_1,\ldots,x_n\} \sim \mathcal{D}_\theta^n} \left[\Pr_{x \sim \mathcal{D}_\theta} \left[F(x_1, \ldots, x_n)(x) \neq \arg\max_{i \in \{1,2\}} f(g = i \mid x) \right] \right] \in \widetilde{\Theta}\left(\frac{1}{\gamma^2} \sqrt{\frac{d}{n}} \right)$$

where $\widetilde{\Theta}$ suppresses logarithmic factors, $\mathcal{D}_{\mu_1,\mu_2}$ denotes the uniform mixture of $\mathcal{N}(\mu_1, I_d)$ and $\mathcal{N}(\mu_2, I_d)$, and $\mathcal{D}^n_{\mu_1,\mu_2}$ denotes n i.i.d. samples from $\mathcal{D}_{\mu_1,\mu_2}$.

Proof of Proposition 3.1. We instantiate the cluster-then-predict algorithm with the Gaussian mixture model estimator of (Azizyan et al., 2013) for the first phase. For the second phase, we use the multicalibration algorithm of Algorithm 2; we will instantiate Algorithm 2 with the trivial distinguisher set $\mathcal{G}_1 = \{x \to 1\}$ because we only need calibration with marginal guarantees within each bucket. By Theorem C.1, the expected clustering error attained by the (Azizyan et al., 2013) estimator learned with T' samples is $O\left(\frac{1}{\gamma^2}\sqrt{\frac{d}{T'}}\right)$. Let the resulting cluster assignment function be denoted F. Using Hoeffding's inequality, this implies that with probability at least $1-\delta$,

$$\begin{aligned}\sum_{t=T'}^{T} \mathbb{1}\left[F(x_t) \neq \arg\max_{j \in [k]} f(j \mid x_t, y_t)\right] &\leq \mathbb{E}\left[\sum_{t=T'}^{T} \mathbb{1}\left[F(x_t) \neq \arg\max_{j \in [k]} f(j \mid x_t, y_t)\right]\right] + \sqrt{(T-T')\log(1/\delta)} \\ &\leq O\left(\tfrac{1}{\gamma^2}\sqrt{\tfrac{d}{T'}}(T-T') + \sqrt{(T-T')\log(1/\delta)}\right). \end{aligned} \tag{4}$$

With a slight abuse of notation, let us define the quantity $\mathbf{DCE}_F(\widehat{y}_{T':T}, x_{T':T}, y_{T':T})$ as the discriminant calibration error that *would have* been incurred had $F : \mathcal{X} \to \{1,2\}$ been the true discriminant with respect to f, that is,[2]

$$\mathbf{DCE}_F(\widehat{y}_{T':T}, x_{T':T}, y_{T':T}) \coloneqq \max_{g \in [k]} \max_{v \in V_\lambda} \left| \sum_{t=1}^{T} \mathbb{1}\left[g = F(x_t)\right] \cdot \mathbb{1}\left[\widehat{y}_t \in v\right] \cdot (\widehat{y}_t - y_t) \right|.$$

[2]Note that the in the usual definition of $\mathbf{DCE}$, the only information needed about the endogenous subgroups model f is the corresponding discriminant function $\arg\max_j f(j|x,y)$, and $f(j|x,y)$ is independent of y given x.

By Theorem D.1, the calibration error on timesteps $t > T'$ where cluster 1 is predicted, i.e. $F(x_t) = 1$, is bounded with probability at least $1 - \delta$ by $O(\sqrt{T_1 \log(1/\delta)})$ where $T_1 = \sum_{\tau=T'+1}^{T} \mathbb{1}[F(x_t) = 1]$ is the number of timesteps where cluster 1 is predicted. With a similar bound holding for cluster 2, by union bound, we have that

$$\mathbf{DCE}_F(\widehat{y}_{T':T}, x_{T':T}, y_{T':T}) \in O(\sqrt{(T - T')\log(1/\delta)}). \tag{5}$$

By triangle inequality and an additional union bound, combining (4) and (5) gives

$$\mathbf{DCE}_f(\widehat{y}_{1:T}, x_{1:T}, y_{1:T}) \leq O\left(T' + \tfrac{1}{\gamma^2}\sqrt{\tfrac{d}{T'}}(T - T') + \sqrt{(T - T')\log(1/\delta)}\right).$$

Choosing $T' = \Theta(d^{1/3}T^{2/3}\gamma^{-4/3})$ gives the desired upper bound of

$$\mathbf{DCE}_f(\widehat{y}_{1:T}, x_{1:T}, y_{1:T}) \leq O\left(d^{1/3}T^{2/3}\gamma^{-4/3} + \sqrt{T\log(1/\delta)}\right).$$

□

C.2 Proposition 3.4

In this section, we prove a generalization of Proposition 3.4: Theorem C.2.

Cluster-Then-Predict Algorithm for Minimizing $\mathbf{LCE}$

For the first $T' < T$ timesteps, make arbitrary predictions and collect observed features $x_1, \ldots, x_{T'}$. Apply a parameter-learning algorithm, such as the (Hardt & Price, 2015) method, to the observed features to obtain estimates of the per-component likelihoods $\widehat{f}(x \mid g)$ for each $g \in [k]$.

Then, instantiate Algorithm 2 with distinguishers $\mathcal{G} = \left\{x \mapsto \widehat{f}(g \mid x) \mid g \in [k]\right\}$. For each timestep $t = T' + 1, \ldots, T$, observe x_t and predict y_t by applying Algorithm 2 to the transcript of previously seen datapoints $\{(x_\tau, y_\tau) \mid T' < \tau < t\}$.

While the algorithm is written for general k and f, we focus on the case where $k = 2$ and $w_1 = w_2 = 1/2$.

Theorem C.2. *Let $k = 2$ and $w_1 = w_2 = 1/2$. Define $\sigma = \|\mu_1 - \mu_2\|_\infty^2 + \|\Sigma_1\|_\infty + \|\Sigma_2\|_\infty$. If we have that $\min_{j\in[d]} |\mu_{1,j} - \mu_{2,j}| \geq \Omega(\sigma)$, then, with probability $1 - \delta$, the* Cluster-Then-Predict Algorithm for Minimizing $\mathbf{LCE}$*, setting $T' = O(T^{2/3})$, incurs*

$$\mathbf{LCE}_f(p_{1:T}, x_{1:T}, y_{1:T}) \leq \widetilde{O}\left(T^{2/3}\sqrt{d}\log^{1/2}(d/\delta)\right).$$

On the other hand, without separation, we must set $T' = O(T^{12/13})$ and incur

$$\mathbf{LCE}_f(p_{1:T}, x_{1:T}, y_{1:T}) \leq \widetilde{O}(T^{12/13}\sqrt{d}\log^{1/2}(d/\delta)).$$

To prove Theorem C.2, we will need some facts relating parameter estimation to errors in estimating group membership are bounded by $O(\varepsilon)$.

Fact C.3. *When parameters (μ_i, Σ_i) of mixture component i are ε, δ-learned in the sense of (Hardt & Price, 2015), we have that with probability $1 - \delta$, $\mathrm{TV}\left(f(x \mid g_i), \widehat{f}(x \mid g_i)\right) \leq O(\varepsilon\sqrt{d})$.*

Proof of Fact C.3. To see this, note that ε, δ-learning in (Hardt & Price, 2015) is defined in terms of ℓ_∞-norm on the estimated parameters relative to the variance of the mixture, i.e.

$$\max_{i=\{1,2\}} \max\left(\|\mu_i - \widehat{\mu}_i\|_\infty^2, \|\Sigma_i - \widehat{\Sigma}_i\|_\infty\right) \leq \varepsilon^2\left(\|\mu_1 - \mu_2\|_\infty^2 + \|\Sigma_1\|_\infty + \|\Sigma_2\|_\infty\right).$$

Theorem 1.8 of (Arbas et al., 2023) shows that $\mathrm{TV}\left(\mathcal{N}(\mu,\Sigma),\mathcal{N}(\widehat{\mu},\widehat{\Sigma})\right) = \Theta(\Delta)$ where Δ is parameter distance in ℓ_2; specifically,

$$\Delta = \max\left(\|\Sigma^{-1/2}\widehat{\Sigma}\Sigma^{-1/2} - I_d\|_F, \|\Sigma^{-1/2}(\mu-\widehat{\mu})\|_2\right).$$

Translating between the ℓ_2 and ℓ_∞ norms incurs a $O(\sqrt{d})$ penalty. □

Fact C.4. *Suppose that for each component g_i, we have estimates of the density of x conditioned on membership in g_i, i.e. $\widehat{f}(x \mid g_i)$, such that $\mathrm{TV}\left(f(x \mid g_i), \widehat{f}(x \mid g_i)\right) \le \varepsilon$. Then, mistakes in estimating group membership can be bounded as $\mathbb{E}_{x\sim f}\left[\left|f(g_i \mid x) - \widehat{f}(g_i \mid x)\right|\right] \le 3\varepsilon$, and the overall TV distance between the true mixture and the estimated mixture can be bounded as $\mathrm{TV}(f(x), \widehat{f}(x)) \le \varepsilon$.*

Proof of Fact C.4. By the definition of conditional probability, we have $f(g_i \mid x) = \frac{f(x,g_i)}{f(x)}$ for every g_i. Then, we can write

$$\begin{aligned}
\mathbb{E}\left[\left|f(g_i \mid x) - \widehat{f}(g_i \mid x)\right|\right] &= \int \left|f(g_i \mid x) - \widehat{f}(g_i \mid x)\right| f(x)dx \\
&= \int \left|f(x,g_i) - \widehat{f}(x,g_i) - \left(\frac{f(x)}{\widehat{f}(x)} - 1\right)\widehat{f}(x,g_i)\right| dx \\
&\le \underbrace{\int \left|f(x,g_i) - \widehat{f}(x,g_i)\right| dx}_{(A)} + \underbrace{\int \left|\left(\frac{f(x)}{\widehat{f}(x)} - 1\right)\widehat{f}(x,g_i)\right| dx}_{(B)}.
\end{aligned}$$

Again by the definition of conditional probability, $f(x,g_i) = f(x \mid g_i)\cdot f(g_i)$, and likewise for the estimated quantity. Recall that the marginal likelihood of g_i (i.e. its mixing weight) is $f(g_i) = \frac{1}{2}$. Then, (A) reduces to $2\cdot\frac{1}{2}\cdot\mathrm{TV}\left(f(x \mid g_i), \widehat{f}(x \mid g_i)\right) \le \varepsilon$. For (B), noting that $\widehat{f}(g_i \mid x) \le 1$ for all x, we have

$$\begin{aligned}
\int \left|\left(\frac{f(x)}{\widehat{f}(x)} - 1\right)\widehat{f}(x,g_i)\right| dx &= \int \left|f(x) - \widehat{f}(x)\right| \frac{\widehat{f}(x,g_i)}{\widehat{f}(x)} dx \\
&= \int \left|f(x) - \widehat{f}(x)\right| \widehat{f}(g_i \mid x) dx \\
&\le \int \left|f(x) - \widehat{f}(x)\right| dx.
\end{aligned}$$

Recalling that $f(x) = \sum_{i\in[k]} f(x \mid g_i)\cdot f(g_i)$ and $k = 2$, we can bound the final quantity in the above display by 2ε. □

A direct consequence of Fact C.4 is that the $\mathbf{LCE}$ incurred by a sequence of predictors $p_1,\dots,p_T$ on samples (x,y) from the true distribution is close to the $\mathbf{LCE}$ that would have been incurred had each (x,y) been sampled from the estimated distribution.

Lemma C.5. *Suppose that for each component g_i, we have estimates of the density of x conditioned on membership in g_i, i.e. $\widehat{f}(x \mid g_i)$, such that $\mathrm{TV}\left(f(x \mid g_i), \widehat{f}(x \mid g_i)\right) \le \varepsilon$. Then, with probability $1-\delta$, the $\mathbf{LCE}$ incurred by a fixed sequence of predictors $p_1,\dots,p_T$ on datapoints $(x,y) \sim f$ can be bounded as*

$$\mathbf{LCE}_f(p_{1:T}, x_{1:T}, y_{1:T}) \le \mathbb{E}_{\widehat{f}}[\mathbf{LCE}_{\widehat{f}}(p_{1:T}, x_{1:T}, y_{1:T})] + O(\sqrt{T})\ln(1/\delta) + 5T\varepsilon.$$

Proof. First, by definition, we have

$$\mathbf{LCE}(p_1, \dots, p_T) = \max_{g \in [k]} \max_{v \in V_\lambda} \left| \sum_{t \in [T]} f(g \mid x_t) \cdot \mathbb{1}[p_t(x_t) \in v] \cdot (p_t(x_t) - y_t) \right|.$$

Note that at each timestep t, and for every g and v, the quantity $f(g \mid x_t) \cdot \mathbb{1}[p_t(x_t) \in v] \cdot (p_t(x_t) - y_t)$ is a random variable bounded in $[0, 1]$. Therefore, for any g and v, we can relate the realized sum to its expected sum; in particular, with probability $1 - \delta$,

$$\begin{aligned}
\left| \sum_{t \in [T]} f(g \mid x_t) \cdot \mathbb{1}[\widehat{y}_t \in v] \cdot (\widehat{y}_t - y_t) \right| &\leq O(\sqrt{T}) \ln(1/\delta) + \left| \mathbb{E} \left[\sum_{t \in [T]} f(g \mid x_t) \cdot \mathbb{1}[p_t(x_t) \in v] \cdot (p_t(x_t) - y_t) \right] \right| \\
&= O(\sqrt{T}) \ln(1/\delta) + \left| \sum_{t \in [T]} \int f(g \mid x) \cdot \mathbb{1}[p_t(x) \in v] \cdot (p_t(x) - y) f(x) dx \right| \\
&\leq O(\sqrt{T}) \ln(1/\delta) \\
&\quad + T \int \left| f(g_i \mid x) - \widehat{f}(g_i \mid x) \right| \cdot \mathbb{1}[p_t(x) \in v] \cdot |p_t(x) - y| f(x) dx \\
&\quad + \left| \sum_{t \in [T]} \int \widehat{f}(g_i \mid x) \cdot \mathbb{1}[p_t(x) \in v] \cdot (p_t(x) - y) \cdot f(x) dx \right|, \qquad (6)
\end{aligned}$$

where Eq. 6 is due to the triangle inequality. Now, we can express the first term in terms of the TV distance between the true and the estimated group-conditional densities. Combining Fact C.4 (for the first term of (6)) and the triangle inequality (for the second term of (6)), we have

$$\begin{aligned}
(6) &\leq O(\sqrt{T}) \ln(1/\delta) + 3T\varepsilon + \left| \sum_{t \in [T]} \int \widehat{f}(g \mid x) \cdot \mathbb{1}[p_t(x) \in v] \cdot (p_t(x) - y) \cdot \widehat{f}(x) dx \right| \\
&\quad + \sum_{t \in [T]} \int \widehat{f}(g \mid x) \cdot \mathbb{1}[p_t(x) \in v] \cdot |p_t(x) - y| \cdot \left| f(x) - \widehat{f}(x) \right| dx \\
&\leq O(\sqrt{T}) \ln(1/\delta) + 5T\varepsilon + \left| \sum_{t \in [T]} \int \widehat{f}(g \mid x) \cdot \mathbb{1}[p_t(x) \in v] \cdot (p_t(x) - y) \cdot \widehat{f}(x) dx \right| \qquad (7) \\
&= O(\sqrt{T}) \ln(1/\delta) + 5T\varepsilon + \left| \mathbb{E}_{\widehat{f}} \left[\sum_{t \in [T]} \widehat{f}(g \mid x) \cdot \mathbb{1}[p_t(x) \in v] \cdot (p_t(x) - y) \right] \right| \\
&\leq O(\sqrt{T}) \ln(1/\delta) + 5T\varepsilon + \mathbb{E}_{\widehat{f}} \left[\left| \sum_{t \in [T]} \widehat{f}(g \mid x) \cdot \mathbb{1}[p_t(x) \in v] \cdot (p_t(x) - y) \right| \right], \qquad (8)
\end{aligned}$$

where Eq. (7) again comes from combining Fact C.4 with the triangle inequality and Eq. (8) is due to Jensen's inequality. The statement of the lemma follows by noting that, because this held for any g and v, it also holds for the max g and v;

furthermore, by another application of Jensen's inequality,

$$\begin{aligned}
&\max_{g\in[k]}\max_{v\in V_\lambda}\mathbb{E}_{\widehat{f}}\left[\left|\sum_{t\in[T]}\int \widehat{f}(g\mid x_t,y_t)\cdot\mathbb{1}[\widehat{y}_t\in v]\cdot(\widehat{y}_t-y_t)\right|\right]\\
&\quad\le \mathbb{E}_{\widehat{f}}\left[\max_{g\in[k]}\max_{v\in V_\lambda}\left|\sum_{t\in[T]}\int \widehat{f}(g\mid x_t,y_t)\cdot\mathbb{1}[\widehat{y}_t\in v]\cdot(\widehat{y}_t-y_t)\right|\right]\\
&\quad= \mathbb{E}_{\widehat{f}}[\mathbf{LCE}_{\widehat{f}}(p_{1:T},x_{1:T},y_{1:T})].
\end{aligned}$$

□

We conclude this section with the proof of Theorem C.2.

Proof of Theorem C.2. The proof of Theorem C.2 proceeds in three steps. In Step 1, we analyze the estimation phase and show that spending T' samples allows us to estimate $\widehat{f}(\cdot)$ with sufficiently small error. In Step 2, we analyze the calibration phase and relate the error incurred when using $\widehat{f}(\cdot)$ to the true error. Step 3 completes the argument.

Step 1: Analyzing the estimation phase. In the clustering phase of *Cluster-Then-Predict Algorithm for Minimizing* $\mathbf{LCE}$, T' samples are used to learn $\widehat{f}(x\mid g_i)$ for $i\in\{1,2\}$. Let δ_1 be the likelihood that parameters are successfully learned in this step.

Let $\sigma=\|\mu_1-\mu_2\|_\infty^2+\|\Sigma_1\|_\infty+\|\Sigma_2\|_\infty$. If we have that $\min_{j\in[d]}|\mu_{1,j}-\mu_{2,j}|\ge\Omega(\sigma)$, i.e., that μ_1 and μ_2 are sufficiently separated in all dimensions, then set $T'=\lceil T^{2/3}\rceil$ and the algorithm of (Hardt & Price, 2015) will ε,δ_1-learn the parameters (μ_1,Σ_1) and (μ_2,Σ_2) with $\widetilde{O}(\varepsilon^{-2}\log(d/\delta))$ samples (where the $\widetilde{O}$ suppresses a $\log\log(1/\varepsilon)$ term). Setting $T^{2/3}=\widetilde{O}\left(\varepsilon^{-2}\log(d/\delta_1)\right)$, we have $\varepsilon=\widetilde{O}\left(T^{-1/3}\log^{1/2}(d/\delta_1)\right)$.

On the other hand, when μ_1 and μ_2 are not separated, we need $\widetilde{O}(\varepsilon^{-12}\log(d/\delta_1))$ samples to learn parameters. We set $T'=\lceil T^{12/13}\rceil$ and instead have $\varepsilon=\widetilde{O}(T^{-1/13}\log^{1/12}(d/\delta_1))$.

Finally, to analyze the error incurred in this phase, note that since the predictor $p_t(x_t)=\frac{1}{2}$ for each $t\le T_1$,

$$\left|\sum_{t\in[T']} f(g_i\mid x_t)\cdot\mathbb{1}[p_t(x_t)\in v]\cdot(p_t(x_t)-y_t)\right|\le\tfrac{1}{2}T^{2/3}$$

for any i and v.

Step 2: Analyzing error incurred in the calibration phase when using $\widehat{f}(\cdot)$. In the calibration phase of *Cluster-Then-Predict Algorithm for Minimizing* $\mathbf{LCE}$, the predictor p_t is updated using the estimated densities $\widehat{f}(\cdot)$. Note that ε error in parameter learning translates to $\varepsilon\sqrt{d}$ error in TV distance between the estimated $\widehat{f}(x\mid g_i)$ and the true $f(x\mid g_i)$, by Fact C.3. Therefore, in the event that all parameters are learned within additive error of ε (which occurs with probability $1-\delta_1$), Lemma C.5 combined with Fact C.3 gives us that, with probability $1-\delta_2$, the error incurred from $t=T'+1\dots T$ is at most

$$\left|\sum_{t=T',\dots,T} f(g\mid x_t)\cdot\mathbb{1}[\widehat{y}_t\in v]\cdot(\widehat{y}_t-y_t)\right|\le\mathbb{E}[\mathbf{LCE}_{\widehat{f}}(p_{T':T},x_{T':T},y_{T':T})]+5T\varepsilon\sqrt{d}+O(\log(1/\delta_2)\sqrt{T-T'}).$$

Recall that we have defined $\mathcal{G}=\left\{x\mapsto\widehat{f}(g\mid x)\mid g\in[k]\right\}$; note that $|\mathcal{G}|=k$. Then, by definition, $\mathbb{E}[\mathbf{LCE}_{\widehat{f}}(p_{T':T},x_{T':T},y_{T':T})]=\mathbb{E}[\mathbf{MCE}(p_{T':T},x_{T':T},y_{T':T};\mathcal{G})]$.

We will now extend Theorem D.1 to hold for expected multicalibration error, i.e. $\mathbb{E}[\mathbf{MCE}(p_{T':T}, x_{T':T}, y_{T':T}; \mathcal{G})]$, rather than for specific realizations of x_t, y_t. In particular, integrating over $\delta \in (0,1)$, we have that $\mathbb{E}[\mathbf{MCE}(p_{T':T}, x_{T':T}, y_{T':T}; \mathcal{G})] \leq O(\eta\sqrt{T - T'\log(k)})$ for $T - T' \geq C\eta^{-2}\log(k\lambda)$. Solving for η gives $\eta \leq C\log^{1/2}(k\lambda\delta_3)T^{-1/2}$.

Therefore, $\mathbb{E}[\mathbf{LCE}_{\widehat{f}}(p_{T':T}, x_{T':T}, y_{T':T})] \leq O(T^{1/2}\log^{1/2}(k\lambda/\delta))$, and when μ_1 and μ_2 are sufficiently separated,

$$\left|\sum_{t=T',\ldots,T} f(g_i \mid x_t)\cdot \mathbb{1}[p_t(x_t)\in v]\cdot(p_t(x_t)-y_t)\right| \leq \widetilde{O}\left(T^{1/2}(\log(1/\delta_2)+\log^{1/2}(k\lambda/\delta))\right) + \widetilde{O}\left(T^{2/3}\sqrt{d}\log^{1/2}(d/\delta_1)\right).$$

Otherwise, without separation, we have

$$\left|\sum_{t=T',\ldots,T} f(g_i \mid x_t)\cdot \mathbb{1}[p_t(x_t)\in v]\cdot(p_t(x_t)-y_t)\right| \leq \widetilde{O}\left(T^{1/2}(\log(1/\delta_2)+\log^{1/2}(k\lambda/\delta))\right) + \widetilde{O}\left(T^{12/13}\sqrt{d}\log^{1/2}(d/\delta_1)\right).$$

Step 3: Combining Steps 1 and 2. We now have that with probability $1-\delta_1$, all parameters are learned within additive error (from step 1); and conditioning on that event, with probability $1-\delta_2$ that the error incurred in the prediction phase can be bounded as argued in Step 2. We can therefore write the total error incurred over all T samples for any g_i and v when means are separated as, with probability $(1-\delta_1)(1-\delta_2) \geq 1-\delta_1-\delta_2$,

$$\begin{aligned}\left|\sum_{t\in[T]} f(g_i \mid x_t)\cdot \mathbb{1}[p_t(x_t)\in v]\cdot(p_t(x_t)-y_t)\right| &\leq \left|\sum_{t\in[T']} f(g_i \mid x_t)\cdot \mathbb{1}[p_t(x_t)\in v]\cdot(p_t(x_t)-y_t)\right| \\ &\quad + \left|\sum_{t=T',\ldots,T} f(g_i \mid x_t)\cdot \mathbb{1}[p_t(x_t)\in v]\cdot(p_t(x_t)-y_t)\right| \\ &\leq O\left(T^{2/3}\right) + \widetilde{O}\left(T^{2/3}\sqrt{d}\log^{1/2}(d/\delta_1)\right),\end{aligned}$$

and without separation as

$$\left|\sum_{t\in[T]} f(g_i \mid x_t)\cdot \mathbb{1}[p_t(x_t)\in v]\cdot(p_t(x_t)-y_t)\right| \leq O\left(T^{2/3}\right) + \widetilde{O}\left(T^{12/13}\sqrt{d}\log^{1/2}(d/\delta_1)\right).$$

The statement of the theorem follows from noting that this holds for any g_i and v, and therefore also holds for the maximum g_i and v. □

D Supplemental Material for Section 4

D.1 Algorithms 1 and 2

Here, we give example algorithms that can be used to instantiate a version of the Online Multicalibration Algorithm for Coverable Distinguishers.

Algorithm 1 computes an empirical cover on samples $x_1, \ldots, x_T$, inspired by the classical approach (for binary-valued functions) of Haussler & Welzl (1986).

Algorithm 1: Algorithm for computing a cover

1: Input: function class $\mathcal{G} \subset [0,1]^{\mathcal{X}}$, $\varepsilon \in (0,1)$, samples $x_{1:T}$;
2: Initialize empty class $\mathcal{G}'$;
3: For every possible labeling $y_{1:T} \in V_\varepsilon^T$, add to $\mathcal{G}'$ any $g \in \mathcal{G}$ where $g(x_t) \in y_t$ for all $t \in [T]$;
4: **return** $\mathcal{G}'$

We also give Algorithm 2, the online multicalibration algorithm of Haghtalab et al. (2023).

Algorithm 2: Online multicalibration algorithm

1: Input: $\mathcal{G} \subset [0,1]^{\mathcal{X}}$, $\varepsilon \in (0,1)$, $\lambda, T \in \mathbb{Z}_+$;
2: Initialize Hedge iterate $\ell^{(1)} = \text{Uniform}(\{\pm 1\} \times \mathcal{G} \times V_\lambda)$;
3: **for** $t = 1$ to T **do**
4: Compute $p^{(t)}(x) \coloneqq \min_{p^*(x) \in \Delta([0,\varepsilon/4\lambda,\ldots,1])} \max_{y \in [0,1]} \mathbb{E}_{\widehat{y} \sim p^*(x)} \left[\mathbb{E}_{\ell_{i,g,v} \sim \ell^{(t)}} [i \cdot g(x) \cdot \mathbb{1}[y \in v] \cdot (\widehat{y} - y)] \right]$;
5: Announce predictor $p^{(t)}$ to Nature and observe Nature's choice $(x^{(t)}, y^{(t)})$;
6: Update $\ell^{(t+1)} \coloneqq \text{Hedge}(\widetilde{\ell}^{(1:t)})$ where $\widetilde{\ell}^{(t)}(i,g,v) \coloneqq 1 - \frac{1}{2} \mathbb{E}_{\widehat{y} \sim p_t(x_t)} [1 + i \cdot g(x_t) \cdot \mathbb{1}[y_t \in v] \cdot (\widehat{y} - y_t)]$;
7: **end for**

Algorithm 2 enjoys the following guarantee on multicalibration error.

Theorem D.1 (Haghtalab et al. (2023) Theorem 4.3)**.** *Fix $\varepsilon > 0$, $\lambda \in \mathbb{Z}_+$, and distinguishers $\mathcal{G} \subset [0,1]^{\mathcal{X}}$. With probability $1 - \delta$, Algorithm 2 guarantees $\mathbf{MCE}(p_{1:T}, x_{1:T}, y_{1:T}; \mathcal{G}) \le \varepsilon T$ for $T \ge O(\varepsilon^{-2} \ln(|\mathcal{G}|\, \lambda/\delta))$.*

D.2 Proof of Lemma 4.3

We first recall the following one-sided testing form of Bernstein's inequality.

Fact D.2. *Let $X_1, \ldots, X_T$ be i.i.d. random variables supported on $[0,1]$ with mean $\mu = \mathbb{E}[X_i]$, and let $\widehat{\mu} = \frac{1}{T}\sum_{i=1}^T X_i$ be the sample mean. Then, for any $\varepsilon > 0$, we have:*

$$\text{If } \mu > 3\varepsilon, \text{ then } \Pr(\widehat{\mu} \le 2\varepsilon) \le \exp\left(-\frac{T\varepsilon}{8}\right).$$

Proof of Fact D.2. Applying Bernstein's inequality with deviation $t = \mu - 2\varepsilon \ge \mu/3 \ge \varepsilon$ gives:

$$\Pr(\widehat{\mu} \le 2\varepsilon) = \Pr(\mu - \widehat{\mu} \ge t) \le \exp\left(\frac{-Tt^2}{2(\mu + \frac{t}{3})}\right) \le \exp\left(\frac{-\frac{1}{3}T\mu\varepsilon}{2(\frac{4}{3}\mu - \frac{2}{3}\varepsilon)}\right) = \exp\left(\frac{-T\mu\varepsilon}{4(2\mu - \varepsilon)}\right) \le \exp\left(-\frac{T\varepsilon}{8}\right).$$

□

We now turn to proving Lemma 4.3, which states that a small covering can be easily obtained via random sampling for any real-valued function class with small covering number; this follows a similar approach to Haussler & Welzl (1986).

Lemma 4.3. *Given $\varepsilon, \delta, T > 0$ and a real-valued function class $\mathcal{F}$ where $N_1(\varepsilon/96, \mathcal{F}, 2T) \le \sqrt{\frac{1}{8}\exp(T\varepsilon^2/32)}$, consider any ε-cover $\mathcal{F}'$ of $\mathcal{F}$ computed on T random datapoints. Then $\mathcal{F}'$ is a 4ε-cover of $\mathcal{F}$ with probability at least $1 - O(N_1\left(\frac{\varepsilon}{8}, \mathcal{F}, 2T\right)^2 \exp(-T\varepsilon))$.*

Proof of Lemma 4.3. **Existence of covering points.**

First, note that for any ε and m, $N_1(\varepsilon, \{f(x) - g(x) \mid f, g \in \mathcal{F}\}, m) \le N_1(\varepsilon, \mathcal{F}, m)^2$. Then, recalling that bounded covering number implies uniform convergence,

$$\Pr\left(\sup_{f,g \in \mathcal{F}} \mathbb{E}\left[|f(x) - g(x)|\right] - \frac{1}{T}\sum_{t=1}^T |f(x'_t) - g(x'_t)| \ge \varepsilon/8\right) \le 4N_1(\varepsilon/(16 \cdot 8), \mathcal{F}, 2T)^2 \exp\left(-T\varepsilon^2/32\right).$$

Since $8N_1(\varepsilon/(16\cdot 8), \mathcal{F}, 2T)^2 \le \exp\big(T\varepsilon^2/32\big)$:

$$\Pr\left(\sup_{f,g\in\mathcal{F}} \mathbb{E}\left[|f(x)-g(x)|\right] - \tfrac{1}{T'}\sum_{t=1}^{T'} |f(x'_t)-g(x'_t)| \ge \varepsilon/8\right) \le \tfrac{1}{2}.$$

Thus, by the probabilistic method, there exists some sequence of datapoints $x_1, \ldots, x'_T$ such that

$$\sup_{f,g\in\mathcal{F}} \mathbb{E}\left[|f(x)-g(x)|\right] - \tfrac{1}{T}\sum_{t=1}^{T} |f(x'_t)-g(x'_t)| \le \varepsilon/8. \tag{9}$$

Covering failure events. Let $x_1, \ldots, x_T$ denote i.i.d. samples from distribution $\mathcal{D}$. By definition, there is a subset $\mathcal{F}' \subset \mathcal{F}$ of size $N_1(\varepsilon, \mathcal{F}, T)$ such that for all $f \in \mathcal{F}$, there is a $f' \in \mathcal{F}'$ such that $\frac{1}{T}\sum_{t=1}^{T} |f(x_t) - f'(x_t)| \le \varepsilon$. Note that this subset $\mathcal{F}'$ is not dependent on $\mathcal{D}$ and can be computed from $x_1, \ldots, x_T$.

We next define, for any $f, g \in \mathcal{F}$, the event $E_{f,g}$ to be the event that both $\mathbb{E}\left[|f(x)-g(x)|\right] \ge 4\varepsilon$ and $\frac{1}{T}\sum_{t=1}^{T} |f(x_t)-g(x_t)| \le \varepsilon$.

Condition on none of the events in $\{E_{f,g} \mid f, g \in \mathcal{F}\}$ occurring. Fix any $f \in \mathcal{F}$. Since we covered $\mathcal{F}$ on $x_1, \ldots, x_T$ with a tolerance of ε, there is a $g \in \mathcal{F}'$ such that $\frac{1}{T}\sum_{t=1}^{T} |f(x_t)-g(x_t)| \le \varepsilon$. Since $E_{f,g}$ did not occur, we know that $\mathbb{E}\left[|f(x)-g(x)|\right] \le 4\varepsilon$. This implies that $\mathcal{F}'$ is a 4ε-net as desired.

It thus suffices to upper bound $\Pr\Big[\bigcup_{f,g\in\mathcal{F}} E_{f,g}\Big]$.

Bounding $\bigcup \Pr[E_{f,g}]$ by covering $\mathcal{F}$. We now define, for any $f, g \in \mathcal{F}$, the event $\widetilde{E}_{f,g}$ to be the event that both $\mathbb{E}\left[|f(x)-g(x)|\right] \ge 3\varepsilon$ and $\frac{1}{T}\sum_{t=1}^{T} |f(x_t)-g(x_t)| \le 2\varepsilon$.

Let $\widehat{\mathcal{F}}$ be a $\big(\frac{\varepsilon}{4}\big)$-covering of $\mathcal{F}$ on the datapoints $x'_1, \ldots, x'_T, [x_1, \ldots, x_T]_M$. Therefore, $\widehat{\mathcal{F}}$ is a cover of size $N_1(\frac{\varepsilon}{8}, \mathcal{F}, 2T)$. This means for every $f \in \mathcal{F}$ there is a $\widehat{f} \in \widehat{\mathcal{F}}$ such that

$$\tfrac{1}{2T}\left(\sum_{t=1}^{T}\left|f(x'_t) - \widehat{f}(x'_t)\right| + \sum_{t=1}^{T}\left|f(x_t) - \widehat{f}(x_t)\right|\right) \le \tfrac{\varepsilon}{8}.$$

We now proceed to show that $\bigcup_{f,g\in\mathcal{F}} E_{f,g} \subseteq \bigcup_{\widehat{f},\widehat{g}\in\widehat{\mathcal{F}}} E_{\widehat{f},\widehat{g}}$.

First, by the triangle inequality,

$$\begin{aligned}\tfrac{1}{T}\sum_{t=1}^{T}\left|\widehat{f}(x_t) - \widehat{g}(x_t)\right| &\le \tfrac{1}{T}\sum_{t=1}^{T}|f(x_t)-g(x_t)| + \left|\tfrac{1}{T}\sum_{t=1}^{T}|f(x_t)-g(x_t)| - \tfrac{1}{T}\sum_{t=1}^{T}\left|\widehat{f}(x_t) - \widehat{g}(x_t)\right|\right| \\ &\le \tfrac{1}{T}\sum_{t=1}^{T}|f(x_t)-g(x_t)| + \varepsilon.\end{aligned}$$

Therefore, $\{\frac{1}{T}\sum_{t=1}^{T} |f(x_t)-g(x_t)| \le \varepsilon\}$ only if $\{\frac{1}{T}\sum_{t=1}^{T}\left|\widehat{f}(x_t) - \widehat{g}(x_t)\right| \le 2\varepsilon.\}$

Next, by the definition of $\widehat{\mathcal{F}}$, we have

$$\tfrac{1}{T}\sum_{t=1}^{T}\left|f(x'_t) - \widehat{f}(x'_t)\right| \le \varepsilon/4 \quad \text{and} \quad \tfrac{1}{T}\sum_{t=1}^{T}\left|f(x_t) - \widehat{f}(x_t)\right| \le \varepsilon/4.$$

By the triangle inequality, for every $f, g \in \mathcal{F}$ and their matching $\widehat{f}, \widehat{g} \in \widehat{\mathcal{F}}$, we have

$$\left|\tfrac{1}{T}\sum_{t=1}^{T}|f(x'_t)-g(x'_t)| - \tfrac{1}{T}\sum_{t=1}^{T}\left|\widehat{f}(x'_t)-\widehat{g}(x'_t)\right|\right| \le \tfrac{\varepsilon}{2} \quad \text{and} \quad \left|\tfrac{1}{T}\sum_{t=1}^{T}|f(x_t)-g(x_t)| - \tfrac{1}{T}\sum_{t=1}^{T}\left|\widehat{f}(x_t)-\widehat{g}(x_t)\right|\right| \le \tfrac{\varepsilon}{2}. \tag{10}$$

Thus, repeatedly applying the triangle inequality (first, third, and fifth transitions below),

$$\begin{aligned}
\mathbb{E}\left[|f(x)-g(x)|\right] &\le \tfrac{1}{T}\sum_{t=1}^{T}|f(x'_t)-g(x'_t)| + \left|\mathbb{E}\left[|f(x)-g(x)|\right] - \tfrac{1}{T}\sum_{t=1}^{T}|f(x'_t)-g(x'_t)|\right| \\
&\le \tfrac{1}{T}\sum_{t=1}^{T}|f(x'_t)-g(x'_t)| + \tfrac{\varepsilon}{4} \\
&\le \tfrac{1}{T}\sum_{t=1}^{T}\left|\widehat{f}(x'_t)-\widehat{g}(x'_t)\right| + \left|\tfrac{1}{T}\sum_{t=1}^{T}|f(x'_t)-g(x'_t)| - \tfrac{1}{T}\sum_{t=1}^{T}\left|\widehat{f}(x'_t)-\widehat{g}(x'_t)\right|\right| + \tfrac{\varepsilon}{4} \\
&\le \tfrac{1}{T}\sum_{t=1}^{T}\left|\widehat{f}(x'_t)-\widehat{g}(x'_t)\right| + \tfrac{3\varepsilon}{4} \\
&\le \mathbb{E}\left[\left|\widehat{f}(x)-\widehat{g}(x)\right|\right] + \left|\mathbb{E}\left[\left|\widehat{f}(x)-\widehat{g}(x)\right|\right] - \tfrac{1}{T}\sum_{t=1}^{T}\left|\widehat{f}(x'_t)-\widehat{g}(x'_t)\right|\right| + \tfrac{3\varepsilon}{4} \\
&\le \varepsilon + \mathbb{E}\left[\left|\widehat{f}(x)-\widehat{g}(x)\right|\right],
\end{aligned}$$

where the second and final transitions are due to (9) and the fourth is due to (10).

Thus, for any $f, g \in \mathcal{F}$, $\mathbb{E}\left[|f(x)-g(x)|\right] \ge 4\varepsilon$ only if the corresponding $\widehat{f}, \widehat{g} \in \mathcal{F}$ satisfies $\mathbb{E}\left[\left|\widehat{f}(x)-\widehat{g}(x)\right|\right] \ge 3\varepsilon$. It follows that $\Pr\left(\bigcup_{f,g\in\mathcal{F}} E_{f,g}\right) \le \Pr\left(\bigcup_{\widehat{f},\widehat{g}\in\widehat{\mathcal{F}}} \widetilde{E}_{\widehat{f},\widehat{g}}\right)$.

Finally, for any fixed choice of f and g, we have by Fact D.2 that $\Pr\left[\widetilde{E}_{f,g}\right] \le \exp\left(-\frac{T\varepsilon}{8}\right)$. Thus, union bounding over all pairs $\widehat{f}, \widehat{g} \in \widehat{\mathcal{F}}$, we have

$$\Pr\left(\bigcup_{f,g\in\mathcal{F}} \widetilde{E}_{f,g}\right) \le \left|\widehat{\mathcal{F}}\right|^2 \exp\left(-\tfrac{T\varepsilon}{8}\right) \le O(N_1(\tfrac{\varepsilon}{8}, \mathcal{F}, 2T)^2 \exp(-T\varepsilon)).$$

□

D.3 Proof of Lemma 4.7

Proof of Lemma 4.7. Let $\mathcal{G}_{\text{LCE}}$ be defined as in (2), and let $\widehat{\mathcal{G}}$ denote an ε'-cover of $\mathcal{G}_{\text{LCE}}$ for some value of ε'. Now, construct the set $\mathcal{G}' = \{F_{f'} \mid f'_1, \ldots, f'_k \in \widehat{\mathcal{G}} \text{ and } \{f'_i\}_{i\in[k]} \text{ is legal}\}$. Because $\mathcal{G}'$ is constructed from $\widehat{\mathcal{G}}$, we have by Lemma 4.5 and Fact B.1 that $|\mathcal{G}'| \le |\widehat{\mathcal{G}}| \le O(d^2k\log^2(d^2k)(1/\varepsilon')^{d^2k\log^2(d^2k)})$.

We now show that $\mathcal{G}'$ is indeed an ε-net for the distinguishers in (1). Recall that f denotes the true likelihoods $f(g \mid x)$ in the underlying endogenous subgroups model. Consider a set of legal likelihoods $\{f'(g \mid x)\}_{g\in[k]}$ and an x where $i = \arg\max_{i\in[k]} f(i \mid x)$. If $f(i \mid x) > f(j \mid x) + 2\varepsilon'$ for all $j \ne i$ and $|f(j;x) - f'(j;x)| \le \varepsilon'$ for all j, then $F_f(x) = F_{f'}(x)$. Since $\mathcal{G}'$ is an ε'-cover of the legal likelihoods, there must exist a set of legal likelihoods f' such that $|f(j;x) - f'(j;x)| \le \varepsilon'$ for all j and x. We also have by Lemma D.3 that, this property is satisfied with probability

$1-\varepsilon$ if $\gamma \geq \Omega(\sqrt{d+\log((k\varepsilon'+1)/\varepsilon)})$. Thus, for $\varepsilon, \varepsilon' \leq 1/T$, we have $\Pr[F_f(x) \neq F_{f'}(x)] \leq \varepsilon$ and $\mathcal{G}'$ is an ε-net for the distinguishers in (1).

Thus, there is an ε-net of size $O((d^2k\log^2(d^2k)(1/\varepsilon')^{d^2k\log^2(d^2k)})^k)$ for $\varepsilon' < O(\varepsilon(\exp(\gamma^2-d)-1)/k)$. This means that the distinguishers $\mathcal{G}$ have a covering number of $N_1(\varepsilon, \mathcal{G}, m) \in O((d^2k\log^2(d^2k)(k/\gamma\varepsilon)^{d^2k\log^2(d^2k)})^k)$. □

Lemma D.3. *Suppose covariance matrices $\Sigma_1, \ldots, \Sigma_k$ have constant eigenvalues and that, for all $i \neq j$, $\|\mu_i - \mu_j\| \geq \Omega(\sqrt{d+\log((k\varepsilon+1)/\delta)})$. Then given a random sample x from their uniform mixture, with probability at least $1-\delta$ there is some $i \in [k]$ such that for all $j \neq i$, $f(i;x) \geq f(j;x) + \varepsilon$.*

Proof. In this proof, let $\|M\|$ denote the spectral norm of a matrix M. Fix $i \in [k]$. We first will apply a change of basis W so that the ith component of the Gaussian mixture is standardized. Note that spectral norm of a transformed matrix can be bounded as $\|WM\| \leq \sqrt{\|W\|\|W^{-1}\|}\|M\|$, whereas the determinant is wholly invariant as $|WM| = |M|$.

By Fact D.4 and Fact D.5, if

$$\begin{aligned}\|\mu_i - \mu_j\| &\geq \sqrt{\max\left\{0, 2\|W\Sigma_j\|(r^2 + \log(w) - \tfrac{1}{2}\log(|W\Sigma_j|))\right\}} \\ &= \sqrt{\max\left\{0, 2\|\Sigma_j\|\sqrt{\|\Sigma_i\|\|\Sigma_i^{-1}\|}(r^2 + \log(w) - \tfrac{1}{2}\log(|\Sigma_j|))\right\}} \\ &\geq \sqrt{\max\left\{0, 2\|\Sigma_j\|\sqrt{\|\Sigma_i\|\|\Sigma_i^{-1}\|}(2d + 8\log(\tfrac{1}{\delta}) + \log(w) - \tfrac{1}{2}\log(|\Sigma_j|))\right\}}\end{aligned}$$

then with probability at least $1-\delta$ on a sample $x \sim f(x;i)$, for any j: $f(i;x)/f(j;x) \geq w$. Thus, for all $j \neq i$, $f(i;x) \geq f(j;x) + \frac{w-1}{w+k-1}$. Thus, if

$$\begin{aligned}\|\mu_i - \mu_j\| &\geq \sqrt{\max\left\{0, 2\|\Sigma_j\|\sqrt{\|\Sigma_i\|\|\Sigma_i^{-1}\|}(2d + 8\log(\tfrac{1}{\delta}) + \log\left(\tfrac{1+\varepsilon k-\varepsilon}{1-\varepsilon}\right) - \tfrac{1}{2}\log(|\Sigma_j|))\right\}} \\ &\geq \Omega(\sqrt{\max\{0, d + \log((1+\varepsilon k)/\delta) - c\}})\end{aligned}$$

for all $i \neq j$, with probability at least $1-\delta$, there is some $i \in [k]$ such that for all $j \neq i$, $f(i;x) \geq f(j;x) + \varepsilon$. □

Fact D.4. *Let $\mathcal{N}(0, I)$ be Gaussian. Then, for any $\varepsilon > 0$, there is $r = \sqrt{d} + 2\sqrt{\log\frac{1}{\varepsilon}}$ such that $\Pr_{x\sim\mathcal{N}(0,I)}(\|x\| \geq r) < \varepsilon$.*

Fact D.5. *Let $\mathcal{N}(0, I)$ and $\mathcal{N}(\mu, \Sigma)$ be two non-isotropic Gaussian distributions, and let $\|\Sigma\|$ denote the spectral norm of covariance matrix $\Sigma > 0$. Suppose $\|\mu\| \geq \sqrt{\max\left\{0, 2\|\Sigma\|(r^2 + \log(w) - \frac{1}{2}\log(|\Sigma|))\right\}}$. Then, the PDF of $\mathcal{N}(0, I)$ is a factor of $w > 1$ larger than the PDF of $\mathcal{N}(\mu, \Sigma)$ at every point within a radius r ball around the origin.*

Proof of Fact D.5. The PDF ratio of $\mathcal{N}(\mu, \Sigma)$ and $\mathcal{N}(0, I)$ at a point x is:

$$\frac{p_2(x)}{p_1(x)} = \frac{1}{|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}\left[(x-\mu)^\top\Sigma^{-1}(x-\mu) - \|x\|^2\right]\right).$$

We can lower bound

$$(x-\mu)^T\Sigma^{-1}(x-\mu) - \|x\| \geq \|x-\mu\|^2\|\Sigma\|^{-1} - r^2.$$

Thus, if $\|x-\mu\|^2 \geq 2\|\Sigma\|(r^2 + \log(w/|\Sigma|^{1/2}))$, then $\frac{p_2(x)}{p_1(x)} \leq \frac{1}{w}$. □

D.4 Proof of Theorem 4.8

The main technical tool for overcoming separation dependence is that, for the uniform and isotropic setting, the combinatorial complexity of discriminant functions can be directly bounded.

Fact D.6. *For a uniform mixture of k isotropic Gaussians, the optimal discriminant function is always a Voronoi diagram, i.e.* $\arg\max_{j\in\{1,\dots,k\}} f(j \mid x) = \mathbb{1}[\|x - c_j\| \leq \|x - c_i\| \text{ for all } i \neq j]$ *for some centers* $c_1, c_2, \dots, c_k \in \mathbb{R}^d$.

Proof. Since the components are uniformly weighted, to find the component j that maximizes $f(j \mid x)$, we can equivalently maximize $f(x \mid j)$. Since the exponential function is monotonically increasing, maximizing $f(x \mid j)$ is equivalent to minimizing $\|x - \mu_j\|^2$, which is a Voronoi diagram. □

Proof of Theorem 4.8. By Fact D.6, $\mathcal{G}$ consists solely of functions that can be expressed as membership in a cell of a Voronoi diagram. Since Voronoi diagrams are of VC dimension $\leq n \coloneqq (d+1)k^2 \log(k) + k$ (Fact B.5), the covering number of $\mathcal{G}$ is bounded by $N_1(\varepsilon, \mathcal{G}, T) \leq O\left(n(1/\varepsilon)^n\right)$ (Fact B.1). The statement of the theorem follows from Proposition 4.2 combined with Fact 4.1, and choosing $\varepsilon = 1/T$. □