Be Intentional About Fairness!: Fairness, Size, and Multiplicity in the Rashomon Set

Gordon Dai*
New York University
New York, NY, USA
td2568@nyu.edu

Pavan Ravishankar*
New York University
New York, NY, USA
pr2248@nyu.edu

Rachel Yuan
New York University
New York, NY, USA
ry1023@nyu.edu

Emily Black†
New York University
New York, NY, USA
emilyblack@nyu.edu

Daniel B. Neill†
New York University
New York, NY, USA
daniel.neill@nyu.edu

Abstract

When selecting a model from a set of equally performant models, how much unfairness can you really reduce? Is it important to be intentional about fairness when choosing among this set, or is arbitrarily choosing among the set of "good" models good enough? Recent work has highlighted that the phenomenon of model multiplicity—where multiple models with nearly identical predictive accuracy exist for the same task—has both positive and negative implications for fairness, from strengthening the enforcement of civil rights law in AI systems to showcasing arbitrariness in AI decision-making. Despite the enormous implications of model multiplicity, there is little work that explores the properties of sets of equally accurate models, or Rashomon sets, in general. In this paper, we present theoretical and methodological contributions which help us to understand the relatively unexplored properties of the Rashomon set, in particular with regards to fairness. Our contributions include methods for efficiently sampling models from this set and techniques for identifying the fairest models according to key fairness metrics such as statistical parity. We also derive the probability that an individual's prediction will be flipped within the Rashomon set, as well as expressions for the set's size and the distribution of error tolerance used across models. These results lead to policy-relevant takeaways, such as the importance of intentionally looking for fair models within the Rashomon set, and understanding which individuals or groups may be more susceptible to arbitrary decisions.

*Equal Contribution
†Equal Contribution

EAAMO '25, Pittsburgh, PA, USA

ACM ISBN 979-8-4007-2140-3/25/11
https://doi.org/10.1145/3757887.3763011

CCS Concepts

• **Social and professional topics** → **Socio-technical systems**; • **Mathematics of computing** → Mathematical optimization.

Keywords

Model multiplicity, Rashomon set, Fairness in machine learning

ACM Reference Format:
Gordon Dai, Pavan Ravishankar, Rachel Yuan, Emily Black, and Daniel B. Neill. 2025. Be Intentional About Fairness!: Fairness, Size, and Multiplicity in the Rashomon Set. In *Equity and Access in Algorithms, Mechanisms, and Optimization (EAAMO '25), November 05–07, 2025, Pittsburgh, PA, USA.* ACM, New York, NY, USA, 32 pages. https://doi.org/10.1145/3757887.3763011

1 Introduction

Recent work has drawn renewed attention to the fact that there are often many (approximately) equally accurate models available for the same prediction task [5, 7, 23]. This phenomenon—often called the Rashomon effect [7], predictive multiplicity [23], or model multiplicity [5]—has wide-ranging implications for both understanding and improving fairness, as these equally accurate models often differ substantially in other properties such as fairness [22, 29] or model simplicity [31–33].

As prior work has pointed out, this multiplicity of models can be viewed as both a fairness opportunity and a concern [5, 11]. On the positive side, legal scholarship has pointed to the fact that model multiplicity is relevant to how to interpret and enforce U.S. anti-discrimination law, and specifically, can strengthen the disparate impact doctrine to more effectively combat algorithmic discrimination [3]. In a recent paper, Black et al. [3] suggest that the phenomenon of model multiplicity could support a reading of the disparate impact doctrine that requires companies to proactively search the set of equally accurate models for less discriminatory alternatives that have equivalent accuracy to a base model deemed acceptable for deployment from a model performance perspective.

On the negative side, several scholars have pointed out that facially similar models, with equivalent accuracy but differences in their individual predictions, can suggest that some model decisions are arbitrary since they seem to be made on the basis of model choice that does not impact performance (e.g., a <1% change in a model's training set accuracy) [2, 18, 23]. This arbitrariness can impact model explanations and recourse as well: individuals with decisions that are unstable across small model changes may not receive reliable explanations for their model outcome, or ways to change it [4, 6, 26]. Further, if there is a group-based asymmetry of arbitrariness–e.g., if female loan applicants have more arbitrariness in their decisions than male loan applicants— this could lead to a group-based equity concern in and of itself.

Understanding the extent of the benefits and risks of model multiplicity relies upon an understanding of the properties of the *Rashomon set*, or the set of approximately equally accurate models for a given prediction task, i.e., equally accurate up to some error tolerance ϵ. While models in the Rashomon set are considered equivalent from a performance perspective, they may differ substantially in other properties—for the purposes of our paper, we focus on fairness. In order to understand the utility of searching for fairer models within the Rashomon set as suggested by recent legal literature, or the extent of the dangers of arbitrariness surfaced by the algorithmic fairness community, we need to understand more about Rashomon sets themselves. For example, whether companies should be required to search for less discriminatory models [3] rests on the question of how much of the disparity can be reduced by optimizing over the Rashomon set, as compared to choosing an arbitrary model without regard to fairness. In other words, **how much do we gain by being intentional about fairness** when selecting models within the Rashomon set? Similarly, concerns about arbitrariness relate to rates and distributions of the chance that an individual will have their prediction changed—**is this arbitrariness harmful** if only predictions that are very uncertain get flipped, or if all demographic groups have an equal chance of flipping? We can shed light on these important questions by understanding even basic facts about the Rashomon set, such as: what does a randomly sampled model from the Rashomon set look like? What is the **average fairness** for various metrics on the Rashomon set? How might one search through the Rashomon set? Can we find the **fairest model** within the Rashomon set? What is the chance that any one individual might experience a change in prediction in the Rashomon set? Or even, **how large is the Rashomon set?** Despite the enormous implications of model multiplicity, there is little work that explores the properties of Rashomon sets in general.

In this paper, we present six main theoretical and methodological contributions that answer the above questions and more—furthering our understanding of the relatively unexplored properties of the Rashomon set, in particular with regards to fairness:

- First, we define the *largest possible Rashomon set* $R_N(\epsilon)$, for N records drawn from a given data distribution, assuming an allowable error tolerance of ϵ. This novel conceptualization of the Rashomon set, based on deviations from the Bayes-optimal model, enables us to explore fundamental questions related to fairness, set size, model selection, and individual predictions.
- Second, we develop an efficient method for sampling models uniformly at random from within the Rashomon set.
- Third, we present two computationally efficient methods to find the fairest model within the Rashomon set, for statistical parity and error rate balance respectively.
- Fourth, we derive the asymptotic probability that any individual will have their prediction flipped within the models of the Rashomon set for a given ϵ.
- Fifth, we derive a closed-form expression for the size of the Rashomon set for a given N and ϵ.
- Sixth, we show that for sufficiently large datasets and small enough ϵ, models in the Rashomon set will use the full error tolerance (i.e., the average accuracy of models in $R_N(\epsilon)$ converges to the accuracy of the Bayes-optimal model minus ϵ).

These theoretical results create important newfound understanding of the Rashomon set with a focus on fairness and fairness-relevant properties— to our knowledge, there are no results about how to sample randomly from the Rashomon set, Rashomon set size, individual flip probabilities within the Rashomon set, and the distribution of error used in the Rashomon set for any generalized theoretical setup. While concurrent work has shown that finding the fairest model within the Rashomon set is hard (NP-hard) in general [22], we are able to show that under certain conditions we can find the fairest model very efficiently.

Further, our theoretical results lead us to some interesting, policy-relevant takeaways, which we expand on further in Sections 4-6 and support with experiments on three datasets:

A. We can gain a lot of fairness by intentionally searching for fairer models within the set of equally accurate models. Sampling randomly within the Rashomon set— only optimizing for accuracy when selecting a model and hoping that it is fair— will yield a much less fair model than searching for the fairest possible model *even among those that are approximately equally accurate*, so explicitly optimizing for fairness within the Rashomon set is important.
B. We can calculate the probability that any given individual will experience a flip in prediction among models in the (largest possible) Rashomon set. This allows us to shed light on the fates of individuals in the Rashomon set and potential inequities in flip probabilities when viewing inconsistency in the Rashomon set as a source of arbitrariness. We can see what factors–such as the

distribution of prediction certainty and other dataset-specific factors—influence the individual and overall probability of flipping in a given Rashomon set.

C. Finally, our theoretical results allow us to understand the size of the Rashomon set and the amount of error tolerance used on average within the set. In particular, we derive large-sample convergence results for the size of the Rashomon set over N data records, as a function of the error tolerance ϵ. These results point to two takeaways that may influence how companies approach the search for less discriminatory models. First, the size of the Rashomon set increases very quickly in ϵ. Second, as the dataset increases in size, the average model in the Rashomon set uses all of the error tolerance (i.e., has accuracy ϵ less than the base model). Thus, a company may want to set as high an ϵ value as possible, to get a larger set of models in the Rashomon set and maximize their opportunity to find a fairer model, but they should expect the majority of models to use all of the error tolerance ϵ.

The remainder of the paper will proceed as follows: after discussing related work in Section 2, we will outline our theoretical setup and notation in Section 3. We then turn to presenting our theoretical work and policy takeaways together in the next three sections: in Section 4, we present new, efficient optimization and sampling approaches to find the fairest model and to sample a model uniformly at random from the Rashomon set, respectively, and demonstrate how that leads to our results showing the importance of intentionally searching for fair models. Next, in Section 5, we present our results on individual prediction flip probabilities, and how this sheds light on arbitrariness and other fairness properties within the Rashomon set. Finally, in Section 6, we introduce our results on Rashomon set size and use of the error tolerance ϵ, and discuss how they can inform how one might search within the Rashomon set for fairer models. Following this, in Sections 7 and 8, we discuss how our modeling set-up relates to practical searches for less discriminatory models, and conclude the paper.

2 Related work and Legal Background

Related Work. There has been a growing stream of work exploring the phenomenon of multiple approximately equally accurate models existing for the same prediction task [5, 7, 10, 12, 23, 30, 31, 34]. Outside of fairness concerns, a series of papers have demonstrated how model multiplicity can be harnessed to find simpler models within the Rashomon set [13, 31, 33], how the existence of multiple equally accurate models can disrupt model explainability [6, 26], and how sets of equally accurate models can differ greatly in their adversarial robustness [12]. Most related to this work, a series of papers focusing on interpretability of models within the Rashomon set have demonstrated how to search for more interpretable models in practice for particular model classes, e.g., decision trees [24, 35], and have provided empirical observations of Rashomon set size for given model classes [35].

Within literature related to fairness concerns, two main themes have emerged: the optimistic vision of using the variability within the Rashomon set to achieve fairness goals with little impact on accuracy [3, 17, 29], and works bringing to light concerns about the arbitrariness of individual decisions from models with many nearly equally accurate counterparts that differ in their predictions, explanations, or other properties [2, 9, 18, 23]. On the arbitrariness side, many works show how models with minimal differences between them—e.g, a change in random seed or sampling of training data—can result in models with different predictions for certain individuals [2, 9, 18, 23]. In this line of work, perhaps the most related is [9], who show empirically that different individuals have radically different chances of experiencing a change in prediction among approximately equivalent models. In our work, we derive the exact probability that an individual will experience a change in prediction in the Rashomon set, and show that this probability varies as a result of a person's underlying certainty of prediction as well as dataset-dependent factors.

On the fairness side, some of the most related works touch on the details of searching through the Rashomon set for less discriminatory models, or less discriminatory alternatives (LDAs). For example, Gillis et al. [17] outline what an LDA search may look like in practice, and develop an algorithm for searching through the set of *linear* models for the least discriminatory alternative. Perhaps the most closely related work, by Laufer et al. [22], outlines a series of theoretical results related to the search for less discriminatory models within the Rashomon set, such as the computational hardness of finding fairer models within the Rashomon set in general, the theoretical limits of fairness within the Rashomon set, and problems around generalizability of less discriminatory models discovered through search. The paper largely points to difficulties around finding a fairer model within the Rashomon set. In contrast, on a high level, one of the major points of our work is to showcase the *importance* of intentionally searching for fairer models within the Rashomon set, by showing the immense fairness difference between models randomly chosen from the Rashomon set (i.e., on the basis of accuracy alone) and the fairest models within the Rashomon set. More generally, our work presents, for the first time, general properties about the Rashomon set itself— such as the average fairness of models within the Rashomon set, the probability that any individual within the Rashomon set will experience a change in prediction across the models in the set, Rashomon set size, the distribution of model error within the Rashomon set, and others—and discusses how these results influence our understanding of not only how to search for fairer models within the Rashomon set, but also how we think about the arbitrariness of individual decisions within the Rashomon set.

Legal Background and LDA Search. We now discuss some of the legal background necessary to understanding how model multiplicity can strengthen the enforcement of civil

rights law in AI systems— but also raises important questions about the utility of searching through the Rashomon set for fairer models. Multiplicity relates to the interpretation and enforcement of civil rights law most directly through the *disparate impact doctrine*. The disparate impact doctrine applies in decision making systems determining access to credit, housing, and employment opportunities, stating that it is illegal to have a decision-making system that distributes these opportunities across different protected demographic groups at different rates unless it is a "business necessity". In practice, the disparate impact doctrine is enforced through a three-step process. First, a plaintiff finds evidence of a decision-making system within a company that distributes opportunities at different rates among demographic groups, such as a bank that approves loans to more men than women. Next, the company argues that this disparate impact is a business necessity—while there is no exact description of what a business necessity is, a general understanding is that the disparity would be necessary for the business to function. In the case of AI decision-making systems, this is often argued by stating that the algorithm used has the highest accuracy possible, that this accuracy is necessary for business function, and that the observed disparity is necessary to achieve this accuracy. However, even if this business necessity defense is accepted, if the plaintiff can demonstrate that there is a *less discriminatory alternative* decision-making system that satisfies business necessity but reduces disparate impact, the firm can be legally liable for the discrimination they have caused, and forced to use the less discriminatory alternative. In the case of algorithmic systems, i.e., when the alternative decision-making system is another algorithm, we follow [3] in calling the less discriminatory alternative algorithm an LDA. Thus, companies subject to the disparate impact doctrine are theoretically incentivized to search for less discriminatory yet still effective models, for fear of being held liable should another entity find a less discriminatory alternative.[1] Some businesses, mostly financial institutions, do this in practice, though domain experts note that "there is an uneven landscape with respect to how or whether institutions assess their models for discrimination, and the effectiveness of existing programs" [28].

In a recent paper, Black et al. [3] outline a novel interpretation of the disparate impact doctrine that puts even greater pressure on companies to search for LDAs. They suggest that since multiple equally accurate models exist for the same prediction task–some of which will likely have different fairness properties— the business necessity argument fails to make sense, and instead, a company should do a proactive search through the Rashomon set of equally accurate models in order to ensure there is no less discriminatory model easily available. A critical question that this raises, however, is how much of the disparity can be reduced by optimizing over the Rashomon set, as compared to choosing an arbitrary model within that set without regard to fairness. In other words, how much do we gain by being *intentional* about fairness within the Rashomon set—by looking for fair models among those that are approximately equally accurate? In this paper, we show that it is well worth it to search for fairer models within the Rashomon set, and that being intentional about doing so is important, as well as other critical insights about the Rashomon set.

[1]While we are aware that Executive Order (EO) 14281 takes a stance against disparate impact as a theory of discrimination and directs federal agencies to de-prioritize enforcement of disparate impact liability, it is the authors' view that despite this EO, the written law has not changed, the threat of liability remains, and will continue to be important in the future.

3 Preliminaries and Notation

In this section, we introduce the mathematical setup and assumptions behind our theoretical results and discuss the implications of these decisions. To define the Rashomon set of approximately equally accurate models, we consider four questions: (i) how do we define a model? (ii) when are models considered distinct? (iii) how do we measure the accuracy of a model? and (iv) if the Rashomon set consists of all models with accuracy within ϵ of some "optimal" model, how is that model defined?

Basics and Model Definition. To answer the first and second questions above, we consider Rashomon sets in the finite-sample case, i.e., assuming we have a fixed number of data records N. Later in the paper, we present theoretical results in the large-sample case, as N goes to infinity. Additional preliminaries and assumptions necessary for those results are presented in Section 5.1. Let $D_N = \langle d_1, d_2, \ldots, d_N \rangle$ be a set of N data records drawn i.i.d. from distribution D. We focus on the binary classification setting, where each data record $d_i = (x_i, y_i)$, $x_i = \{x_{ij}\}$ represents a set of input features (including a binary *sensitive attribute* which we denote as A_i), and y_i is a binary outcome variable. Thus our models are binary classification models, which predict an outcome in $\{0, 1\}$. We define a predictive model by its classification $\hat{y}_i = f(x_i)$ for each data record d_i, that is, by its mapping from input features x_i to decisions $\{0, 1\}$ on the data D_N. Thus, there are 2^N distinct models possible for a set of data records of size N—note that this is the exhaustive set of all possible mappings defined over the N data records. **Thus we term the Rashomon set $R_N(\epsilon)$ of approximately equally accurate models within this set of 2^N models as the *largest possible Rashomon set* for error tolerance ϵ, because it places no restrictions on the model class, smoothness or consistency of predictions.**

Model Accuracy and Optimal Model. Our answers to the third and fourth questions above rely on the concept of a *Bayes-optimal classifier* $f_{\text{opt}}(x_i)$. This model is assumed to have access to the true probabilities $p_i = \Pr(y = 1 \mid x = x_i)$ but not the observed labels y_i. In other words, the Bayes-optimal classifier has access to the underlying probability that given the available input information, an individual data record will have true outcome $y = 1$ in the classification problem (e.g., the probability that an individual will repay a loan based on their application), but not the actual

outcome (whether or not that individual defaulted on the loan). The Bayes-optimal classifier predicts $f_{\text{opt}}(x_i) = 1$ if $p_i > 0.5$, and $f_{\text{opt}}(x_i) = 0$ otherwise, and has the highest expected classification accuracy, $\mathbb{E}[\max(p_i, 1-p_i)]$, among all classifiers using the same set of features x. Thus, given the data records $D_N = \langle d_1, d_2, \ldots, d_N \rangle$ and the corresponding true probabilities $P_N = \langle p_1, p_2, \ldots, p_N \rangle$, we define the Rashomon set $R_N(\epsilon)$ for error tolerance ϵ as the set of all models with expected classification accuracy greater than or equal to $\mathbb{E}[\max(p_i, 1-p_i)] - \epsilon$.

This definition has the advantage of not allowing models to *overfit* the observed data, since expected error is calculated as a function of the underlying probability p_i of an input x having an outcome of 1. If we instead used the observed labels y_i and computed the empirical accuracy $\mathbb{E}[\mathbf{1}\{f(x_i) = y_i\}]$, a non-Bayes-optimal model (e.g., a classifier trained on the test data D_N) could obtain higher empirical accuracy than the Bayes-optimal model, e.g., by predicting $f(x_i) = 1$ for a data record that was *a priori* unlikely to have $y_i = 1$ (i.e., $p_i < 0.5$) but just happens to have $y_i = 1$ in this instance. We discuss generalizability further in Section 7 below.

Defining Other Models in The Rashomon Set. To more easily determine which of the 2^N possible models (mappings of each d_i, $i \in \{1, \ldots, N\}$, to {0,1}) belong to the Rashomon set $R_N(\epsilon)$, we represent each possible model by a binary *flip vector* representing its changes in prediction from the Bayes-optimal model. This allows us to easily tell which models are in the Rashomon set, since we can easily calculate a model's error difference from the Bayes-optimal model using its flip vector. In particular, we define a flip vector $\theta \in \{0,1\}^N$, where $\theta_i = 1$ if $f(x_i) \neq f_{\text{opt}}(x_i)$, and $\theta_i = 0$ if $f(x_i) = f_{\text{opt}}(x_i)$. The Bayes-optimal model $f_{\text{opt}}(\cdot)$ has a corresponding flip vector θ_0 consisting of N zeros. We can then compute the accuracy of any model $f(\cdot)$ with corresponding flip vector θ, which we denote as $acc(\theta)$, as $acc(\theta) = acc(\theta_0) - \frac{1}{N}\sum_{i=1\ldots N} \theta_i |2p_i - 1|$. This follows from the fact that the Bayes-optimal classifier's probability of predicting y_i correctly is $\max(p_i, 1-p_i)$, while the flipped prediction ($\theta_i = 1$) would be correct with probability $\min(p_i, 1-p_i)$, leading to a difference of $|2p_i - 1|$. We thus define the weight w_i corresponding to probability p_i as $w_i = |2p_i - 1|$. These weights can be thought of as the Bayes-optimal classifier's confidence in each positive or negative prediction, and range from 0 (for $p_i = 0.5$) to 1 (for $p_i = 0$ or $p_i = 1$). Let $W_N = \langle w_1, w_2, \ldots, w_N \rangle$ be the weight vector for data records $\langle d_1, d_2, \ldots, d_N \rangle$, and then we can write:

$$acc(\theta) = acc(\theta_0) - \frac{\theta \cdot W_N}{N}.$$

Finally, for a given error tolerance ϵ, we define the largest possible Rashomon set $R_N(\epsilon)$ as all flip vectors $\theta \in \{0,1\}^N$ with $acc(\theta) \geq acc(\theta_0) - \epsilon$, and thus:

$$R_N(\epsilon) = \left\{ \theta \in \{0,1\}^N : \frac{\theta \cdot W_N}{N} \leq \epsilon \right\}.$$

The critical takeaway here is that we can enumerate all of the models in the Rashomon set by checking to see which of the 2^N possible flip vectors fall within the accuracy constraint ϵ. But since it would be too costly in practice to do this for all 2^N flip vectors, we show below how to randomly sample (efficiently) from the Rashomon set—and how to find the fairest model.

4 The Importance of Intentional Fairness

A natural question that may come up when considering searching for fairer models within the Rashomon set is—is it worth it? While it is clear that intentionally searching for fair models *without* a strict bound on accuracy leads to large fairness gains, it is not obvious *a priori* that this holds true within sets of models that are approximately equally accurate. What if the fairness of all the models in the Rashomon set is more or less the same, and a randomly sampled model—akin to selecting a model solely on the basis of accuracy and not paying attention to fairness—is just as fair as the fairest ones within the set? We show that this is definitely not the case— the fairness difference between the average, or randomly sampled, model within the Rashomon set and the fairest models can be very large. We also show experimentally that how you look for fairer models can influence your success—while searching directly for the fairest model is always the most effective method, whether or not you can reach significantly fairer models by randomly sampling models in the Rashomon set is dataset-dependent, and also depends on how you search.

To compare what we gain by being intentional or arbitrary about fairness within the Rashomon set, we must show both how to draw randomly from the Rashomon set and how to find the fairest model. We present **novel, computationally efficient approaches** for (i) optimizing different fairness metrics over the Rashomon set, as described in Section 4.1; and (ii) sampling models uniformly at random from the Rashomon set, as described in Section 4.2. We also describe a simple baseline for comparison in Section 4.3.1: restricting the model class (here we assume penalized logistic regression models) and learning models from that class with different sources of random variation. We compare the fairness of the models found by the optimization, uniform sampling, and restricted model class approaches in Section 4.3, and explore policy takeaways in Section 4.4.

4.1 Optimizing fairness over the Rashomon set

Despite computational hardness results for finding the fairest model in the Rashomon set [22], we show that under certain conditions, it is possible to find the fairest model within the Rashomon set $R_N(\epsilon)$ defined on N data records in log-linear time, $O(N \log N)$. In particular, when we are concerned with mitigating demographic disparity (i.e., equalizing the positive prediction rate or PPR) between two groups, we show that we can find the exact fairest model within the Rashomon set. For equalizing false positive rate (FPR) or true positive rate

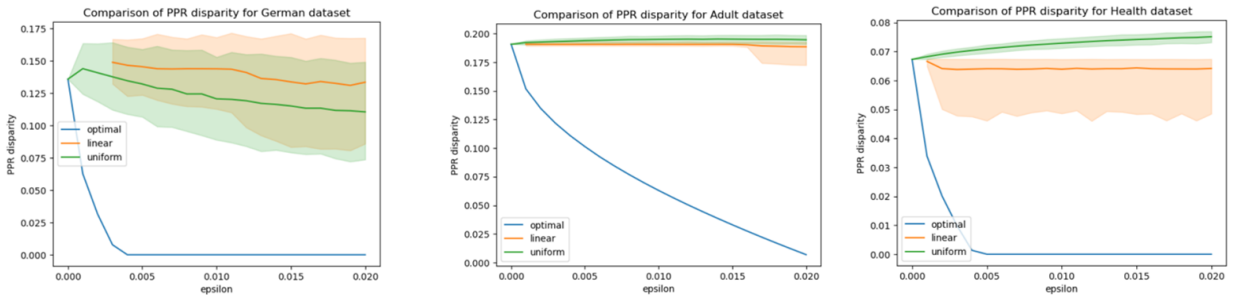

Figure 1: Disparity in positive prediction rate for the German, Adult, and Health datasets, as a function of the error tolerance ϵ. Comparison of methods for optimizing PPR (Section 4.1.1), uniform random sampling (Section 4.2), and sampling linear models (Section 4.3.1) over the Rashomon set $R_N(\epsilon)$.

(TPR) between two groups, we can find a model which is guaranteed to have error rate disparity no more than $O(\frac{1}{N})$ higher than the fairest model. As we show in Section 4.3, using these algorithms on three real-world datasets, we see that in practice, it is often possible to completely eradicate disparities by searching within the Rashomon set for very small ϵ— less than half of a percent in many cases.

For PPR, FPR, and TPR, we can express the optimization of the fairness criterion over flip vectors θ, subject to the constraint that θ is in the Rashomon set $R_N(\epsilon)$, as a *knapsack problem*, where each data record d_i has a weight $w_i = |2p_i - 1|$ corresponding to the error incurred by its flip, and a value v_i corresponding to how much it reduces disparity. A flip occurring, i.e., $\theta_i = 1$, corresponds to the inclusion of element i in the knapsack, adding w_i to the total weight and v_i to the total value. The 0-1 knapsack problem is then the constrained optimization with capacity $N\epsilon$: $\max \sum_i \theta_i v_i$ subject to $\theta_i \in \{0, 1\}$ and $\sum_i \theta_i w_i \leq N\epsilon$.

We note that concurrent work by Laufer et al. [22] formulates the optimization of fairness over the Rashomon set as a subset sum problem (closely related to the knapsack problem) and uses this equivalence to show that their problem (i) is NP-hard to solve in general, and (ii) can be approximated in $O(N^3)$ time. While the knapsack problem is also NP-hard in general, we present efficient $O(N \log N)$ solutions for the special cases below.

4.1.1 Optimizing for statistical parity. We present an efficient, $O(N \log N)$ knapsack approach to find the exact fairest model that minimizes PPR disparity over the Rashomon set $R_N(\epsilon)$, as described in detail in Appendix A.1, Algorithm 1. The goal of this algorithm is to find the individual predictions to flip (setting $\theta_i = 1$) to reduce disparity, until we either use up the entire error tolerance ϵ or completely remove the disparity. Intuitively, we want to flip individuals who will increase the error as little as possible (low weights w_i) and reduce the disparity as much as possible (high values v_i).

The key idea for making this efficient is that there are only two distinct values of v_i: for instance, if group A has a higher positive prediction rate, flipping the prediction of an individual in group A from 1 to 0 reduces the disparity by $\frac{1}{|A|}$, flipping the prediction of an individual in group B from 0 to 1 reduces the disparity by $\frac{1}{|B|}$, and other flips would increase disparity. In this case, the optimal knapsack solution is to flip the predictions of the k_A lowest-weight individuals with $f_{\text{opt}}(x_i) = 1$ from group A and the k_B lowest-weight individuals with $f_{\text{opt}}(x_i) = 0$ from group B. We can then find the optimal values of k_A and k_B (that minimize disparity while satisfying the constraint on accuracy) through a linear-time, incremental search, as shown in Appendix A.1, Algorithm 1, and thus the run time is dominated by the $O(N \log N)$ sorting of items by weight.

4.1.2 Optimizing for error rate balance. We present an efficient, $O(N \log N)$ fractional knapsack approach to find the model that minimizes FPR or TPR disparity over the Rashomon set $R_N(\epsilon)$, to within $O(\frac{1}{N})$ of the optimal disparity, as described in detail in Appendix A.2, Algorithm 2. Again, the goal of this algorithm is to find the lowest-cost individual predictions to flip to reduce disparity, until we either use up the entire error tolerance ϵ or completely remove the disparity.

In this case, however, there are more than two distinct values of v_i (how much flipping an individual reduces disparity) so the PPR solution described above does not work. Instead we use an approximation, the fractional knapsack solution, which flips individuals' predictions (setting $\theta_i = 1$) in descending order of the ratio of their value v_i (the amount they reduce the model's disparity) to their weight w_i (the amount they increase the model's error). This continues until an individual will not "fit" in the knapsack since maximum weight (i.e., error threshold) is reached. Then, a "fraction" of this individual is added to the knapsack. In our case, we cannot flip a fraction of an individual— thus, rather than adding the fractional element, we show that it would reduce disparity by an amount $\theta_i v_i$ that is $O(\frac{1}{N})$. Since the fractional knapsack solution $\sum \theta_i v_i$ is an upper bound on the 0-1 knapsack solution, we know that our solution (excluding the fractional element) reduces disparity to within $O(\frac{1}{N})$ of the optimal disparity.

4.2 Sampling models uniformly at random from the Rashomon set

We now turn to showing how we can sample models uniformly from the Rashomon set, which shows us what typical models from the Rashomon set look like. While we could just sample random flip vectors and keep the ones that are in the Rashomon set, this approach will be ineffective: as we discuss in Appendix B, the vast majority of flip vectors will not be in the Rashomon set. Instead, we propose a new approach based on Gibbs sampling [15] to sample models uniformly from the Rashomon set. This approach, described in Appendix B, Algorithm 3, is computationally efficient, requiring $O(N)$ time per sample.

The key idea of Gibbs sampling is to exploit knowledge of *conditional distributions* even when the full distribution is unknown. In our setting, while we do not know the joint distribution of flip probabilities θ for all records in the dataset, we can easily compute the chance that a data record d_i will flip ($\theta_i = 1$) conditional on which other data records are flipped (θ_{-i}). We show in Appendix B that there are only two possibilities: if the flip vector $\theta_{i=1}$ (with $\theta_i = 1$ and all other flips the same as θ_{-i}) is in the Rashomon set, then there is a 50/50 chance that $\theta_i = 1$, and otherwise we know $\theta_i = 0$. We can then redraw θ_i with the corresponding probability (either 0.5 or 0) of being 1. Given this simple and computationally efficient conditional sampling step, our Gibbs sampling approach starts with the zero vector θ_0, which is guaranteed to be in the Rashomon set, and iteratively samples θ_i (given the current values of θ_{-i}) for all N data elements. To ensure uncorrelated samples from the joint distribution, we take one sample every 10 iterations (where one iteration includes resampling all N elements of θ in randomly permuted order), after an initial burn-in period of 500 iterations. For each dataset and each value of ϵ considered, we run 10,000 iterations of Gibbs sampling, resulting in 950 samples.

4.3 Experiments on real data

We now describe our experimental design for comparing randomly sampled and optimally fair models within Rashomon sets on real data, showing the importance of intentional fairness. Throughout this paper, we present experimental results on three real-world datasets that are commonly used as benchmarks in the fair machine learning literature: German Credit ("German"), Adult, and Heritage Health ("Health"). Details of all three datasets are described in Appendix D.

As noted above, the Bayes-optimal probabilities p_i are unknown for these real-world datasets, but can be well-estimated using sufficient training data. Since we wish to compare the methods over all N data records ($N = 1,000$ for German, $N = 46,443$ for Adult, and $N = 184,308$ for Health), we performed 5-fold cross-validation to estimate these probabilities. For each held-out 20% of the data, we trained a model $\hat{f}_{\text{opt}}(x)$ using the remaining 80% of the data to approximate the Bayes-optimal model $f_{\text{opt}}(x)$, and used its predicted probabilities $\hat{p}_i$ to estimate the Bayes-optimal probabilities p_i for that fold. More precisely, we trained logistic regression models on each dataset that matched typical (maximal) accuracies reported in the wider literature. To check the robustness of our results to the choice of model used for estimation of p_i, we re-ran all experiments using the estimated probabilities $\hat{p}_i$ from XGBoost models learned using 5-fold cross-validation (Appendix I), and found no notable differences.

To test the difference between randomly sampling from the Rashomon set and directly optimizing for the fairest model within the set on real data, for each dataset and each ϵ value, we compared the model found by optimizing the desired fairness metric (PPR, TPR, or FPR) over the Rashomon set $R_N(\epsilon)$, as described in Section 4.1, to the distributions of models found by (i) uniform random sampling over all models (flip vectors) $\theta \in R_N(\epsilon)$, as described in Section 4.2, and (ii) a simple baseline approach, sampling penalized logistic regression models (and corresponding flip vectors θ) from $R_N(\epsilon)$, as described in Section 4.3.1 below. For each distribution of samples, we report the mean and 95% interquantile range, i.e., the 2.5 and 97.5 percentiles of the distribution.

We compare these approaches using three fairness criteria: statistical parity, or balanced positive prediction rate (PPR), balanced false positive rate (FPR), and balanced true positive rate (TPR). Disparities with respect to all three criteria were measured between the protected class ($A_i = a$) and non-protected class ($A_i \neq a$), using the sensitive attribute value for each dataset described in Appendix D. All three measures of disparity for a given flip vector θ were computed using the (estimated) Bayes-optimal probabilities p_i and corresponding weights w_i, rather than the observed outcomes y_i, as described in Appendix A. Results for PPR disparity are shown in Figure 1, and results for FPR and TPR disparity are shown in Appendix F, Figures 5 and 6.

Finally, to demonstrate that our results are robust to using different training sets (drawn from the same distribution) to learn the Bayes-optimal classifier, and that our optimal and sampled models generalize to previously unseen data, we perform an additional experiment in Appendix E, using separate partitions of the Adult dataset to estimate the Bayes-optimal probabilities, to learn parameter values for random sampling and class-specific decision thresholds for optimization, and to evaluate disparity respectively. Results are shown in Appendix E, Figure 4 and Appendix I, Figure 17.

4.3.1 Baseline approach: sampling linear models from the Rashomon set. As a simple baseline for comparison, which might be representative of how a company would typically choose a predictive model for deployment, we assume that a binary classifier is learned from a separate, large training dataset, where the model class is chosen *a priori* and therefore the set of possible flip vectors θ is restricted to members of that class. In particular, we assume that an L_2-penalized logistic regression model is learned. For consistency (since our experiments use all N data records), we use k-fold cross-validation and compute all metrics using predictions (for a given data record d_i) made using a model learned from the other $k - 1$

folds (excluding the fold that contains d_i). Moreover, since a company would typically explore the space of parameter values and choose a model with high accuracy, we learn penalized logistic regression models with different sources of random variation, evaluate their accuracy, and keep those models which are in the Rashomon set. More precisely, to sample over the Rashomon set of L_2-penalized logistic regression models, for a given dataset and value of $\epsilon \in \{0.001, 0.002, \ldots, 0.02\}$, we sample 1,000 models, where for each model we randomly sample the number of cross-validation folds $k \in \{2, 3, \ldots, 10\}$, the logistic regression solver, and the amount of L_2 penalization $C \in \{0.001, 0.01, 0.1, 1.0, 10, 100\}$, and then fit the penalized logistic regression model using scikit-learn. Given the model's predictions, we compute the flip vector θ and include the sampled model in the Rashomon set if $\frac{\theta \cdot W_N}{N} \leq \epsilon$.

4.4 Takeaways for policy and practice

- **A randomly sampled model within the Rashomon set is nowhere near as fair as the fairest model at any given ϵ.** As we can see from the gap between the blue and green lines in Figure 1, searching intentionally for the fairest model within the Rashomon set leads to much fairer models at the same ϵ than randomly sampling within the set. This shows us that a random model from the Rashomon set— one selected on the basis of accuracy alone— will have an extremely low chance of being the fairest, or even one of the fairer, models within the set. This in turn underscores the necessity of explicitly searching for fairer models before deployment– i.e., an LDA search.
- **In practice, it is often possible to completely eradicate disparities by searching within the Rashomon set for quite small ϵ.** As we can see in Figure 1, for German Credit and Health datasets, a model exists that completely eradicates demographic disparity in the dataset for $\epsilon < 0.005$, i.e., half a percentage point of accuracy loss. While the Adult dataset requires very slightly more than 2% accuracy loss to fully eradicate the disparity, this is still a small enough gap considered to be acceptable based on case studies of LDA searches [8].
- **Using repeated random sampling as a search strategy– i.e., looking across many models selected on the basis of accuracy and searching for the fairest among them—can give mixed results.** While our theoretical setup does not map onto how LDA searches would be done in practice— since we search through all the possible mappings of input to output for a dataset instead of generating actual parametric models—loosely, repeated random sampling corresponds to an LDA search that does not directly use protected attribute information until after all the models are trained, i.e., only as a step to evaluate models and choose among them post-hoc. Our optimal search method, on the other hand, corresponds to an LDA search method that uses some direct minimization of disparities across demographic groups during the model creation process, whether that be in hyperparameter tuning, optimization, or other parts of the pipeline. There is disagreement in the legal literature as to whether and to what extent interventions for disparity reduction across demographic groups that use protected class information are legally permissible [16, 19, 21]. Our experimental results show that we gain a lot by being able to directly intervene using protected attributes—however, repeated random sampling without direct use of protected attributes can in some cases be an effective technique as well, even if not as effective as direct intervention. In particular, we see that the German Credit results allow for a large reduction in disparity. At $\epsilon = .02$ (i.e., 2% error tolerance from the optimal model), the total PPR disparity could be reduced by 46% compared to the Bayes-optimal model by reaching the 2.5 percentile of the PPR disparity distribution, which could be achieved in practice by taking the fairest (lowest PPR disparity) of 40 samples from the Rashomon set. In addition, note that while random sampling over the *entire* Rashomon set of all possible mappings of x to y is not particularly effective at reducing disparity in the Health dataset, only searching within linear models is more effective–this is promising given that in practice, LDA searches are typically done within various model classes and not across all possible mappings. Divergence in the effectiveness of different random sampling methods based on model class is an interesting phenomenon that we look forward to studying in future work.

5 Understanding Individual Flip Probabilities

In this section, we present our results showing how to compute expected *flip probabilities* for every record d_i across all models in the Rashomon set, i.e., the chance that a given individual will experience a change in prediction from the Bayes-optimal model in a randomly sampled model in the Rashomon set. Knowing flip probabilities allows us to explore the *arbitrariness* that arises from the Rashomon set: many authors have pointed to the phenomenon of predictive multiplicity [23], where an individual can have different outcomes among different models in a Rashomon set, as a form of inequity through arbitrariness [2, 23]. By seeing the flip probabilities of any individual in the Rashomon set, we can see who is more and less susceptible to potentially arbitrary changes in outcome—and as we discuss in Section 5.3, group-level disparities across who is likely to experience a change in prediction.

5.1 Preliminaries and assumptions for our large-sample theoretical results

Throughout Sections 5 and 6, we present various theoretical results, and the corresponding takeaways for policy and practice, about individual flip probabilities, Rashomon set size, and use of error tolerance, in the *large-sample limit* where the number of data records $N \to \infty$. **For full statements and**

proofs of all theorems, see Appendix C. In this subsection, we present the notation needed to understand the theoretical results, along with the key assumptions that these results depend on.

As in Section 3, we assume data records $d_i = (x_i, y_i)$ drawn i.i.d. from distribution D, with corresponding Bayes-optimal probabilities $p_i = \Pr(y = 1 \mid x = x_i)$, and weights $w_i = |2p_i - 1|$. Let $\langle d_1, d_2, \ldots \rangle$ denote an infinite sequence of data records drawn i.i.d. from D, and let D_N denote records $\langle d_1, d_2, \ldots, d_N \rangle$, with corresponding Bayes-optimal probabilities $P_N = \langle p_1, p_2, \ldots, p_N \rangle$ and weights $W_N = \langle w_1, w_2, \ldots, w_N \rangle$. Moreover, let W be the distribution of weights for data records drawn i.i.d. from D, $w_i \sim W$ for all i, with probability density function (pdf) $f(w)$.

Let $R_N(\epsilon)$ denote the Rashomon set of models for error tolerance ϵ defined over data records $\langle d_1, \ldots, d_N \rangle$. We represent each model in $R_N(\epsilon)$ by a length-N binary *flip vector* $\theta \in \{0,1\}^N$, where $\theta_i = 1$ if $f(x_i) \neq f_{\text{opt}}(x_i)$, $\theta_i = 0$ if $f(x_i) = f_{\text{opt}}(x_i)$, and the Bayes-optimal classification $f_{\text{opt}}(x_i) = \mathbf{1}\{p_i > 0.5\}$. As shown in Section 3, a flip vector $\theta \in R_N(\epsilon)$ if and only if $\frac{\theta \cdot W_N}{N} \leq \epsilon$.

Key assumptions underlying the theoretical results below are threefold: (1) the number of data records N is large; (2) the distribution of weights $f(w)$ is continuous and positive on the interval [0,1]; and (3) ϵ is sufficiently small, less than half of the average weight. We observe that these assumptions are reasonable for all three datasets considered: (1) N is large enough (ranging from $N = 1,000$ for German Credit to $N = 184,308$ for Health) for the finite-sample results to be very close to their large-sample limits; (2) there is enough variability in the weights w_i to assume that they are drawn from a continuous, positive distribution; and (3) average weights for all three datasets range from 0.50 (German Credit) to 0.74 (Health), while the ϵ values we consider for our Rashomon sets are at most 0.02. Nevertheless, the assumptions might be violated for very small datasets (insufficient N); low-dimensional datasets with discrete-valued predictor variables (insufficient variability in w_i); or datasets where the prediction is extremely uncertain, $p_i \approx 0.5$ and thus $w_i \approx 0$, for many data records (average weight too small for the range of ϵ considered).

5.2 Individual flip probabilities

In order to reason about the arbitrariness of individual predictions, we define the *flip probability* $q_{N,i}$ for a given data record d_i, $i \in \{1, \ldots, N\}$, as the proportion of models in the Rashomon set $R_N(\epsilon)$ for which the model prediction $f(x_i)$ differs from the Bayes-optimal prediction $f_{\text{opt}}(x_i) = \mathbf{1}\{p_i > 0.5\}$, or equivalently, the proportion of flip vectors for which $\theta_i = 1$:

$$q_{N,i} = \frac{|\theta \in R_N(\epsilon) : \theta_i = 1|}{|R_N(\epsilon)|}.$$

As $N \to \infty$ for a given weight distribution W and error tolerance ϵ, flip probabilities become pairwise independent (Appendix C, Lemma C.5), and the flip probability $q_i = \lim_{N \to \infty} q_{N,i}$ depends only on the weight w_i. Thus we define the *asymptotic flip probability* function $q(w)$ as the flip probability q_i corresponding to a data record d_i with weight $w_i = w$. We then prove the following theorem (Appendix C, Theorem C.12):

Theorem 5.1 (Asymptotic flip probabilities). *Given the preliminaries and assumptions above, as $N \to \infty$, the flip probability corresponding to a data record with weight $w_i = w$ converges to*

$$q(w) = \frac{1}{1 + \exp(C(\epsilon)\, w)},$$

where $C(\epsilon) = g^{-1}(\epsilon)$ and $g(C) = \int_0^1 \frac{w f(w)}{1+\exp(Cw)}\, dw$.

As a consequence of this theorem, for a Rashomon set $R_N(\epsilon)$ with N large, we can obtain the flip probabilities for each individual, which we outline in Appendix C, Theorem C.12. Computing these flip probabilities provides us with multiple pieces of valuable information about the Rashomon set. First, we can use the flip probabilities to *exactly* (in the large-sample limit) and *efficiently* compute the *average* over the entire Rashomon set of any metric (e.g., PPR, FPR, or TPR disparity) which can be decomposed as a linear function of the individual predictions, as shown in Appendix G. This can help us better understand, without the need for computationally expensive random sampling, how much fairness we expect for a model drawn randomly from the Rashomon set, i.e., whether or not we will arrive at a reasonably fair model by optimizing solely for accuracy and not considering fairness. Second, the flip probabilities $q_{N,i}$ are related to the size of the Rashomon set (Appendix C, Lemma C.4), and thus, as we show in Section 6.1, the asymptotic size of the Rashomon set as $N \to \infty$ can be computed from the quantity $C(\epsilon)$ defined in Theorem 5.1.

Most importantly, however, understanding flip probabilities helps us reason about arbitrariness of prediction in the Rashomon set as we can see who is likely to be more and less susceptible to potentially arbitrary changes in outcome. More precisely, the flip probability $q_{N,i}$ for a given individual is the probability that their prediction will differ from that of the Bayes-optimal model, across all models in the Rashomon set, and we note that $0 < q_{N,i} < \frac{1}{2}$ for all $i \in \{1, \ldots, N\}$. Therefore, individuals with $q_{N,i} \approx 0$ have consistent predictions across the Rashomon set, while individuals with $q_{N,i} \not\approx 0$ may receive either classification depending on which model happens to be drawn, i.e., their prediction is *arbitrary*. While we expect individuals with low-confidence Bayes-optimal probabilities $p_i \approx \frac{1}{2}$ to receive arbitrary predictions, and individuals with high-confidence probabilities $p_i \approx 0$ or $p_i \approx 1$ to receive consistent predictions, the question remains: how far from the decision boundary $p_i = \frac{1}{2}$ must an individual be for their predictions to be consistent? We see in Figure 2(left) that the answer to this question differs across datasets and varies with ϵ.

In addition, understanding individual flip probabilities within the Rashomon set can shed light on another source of inequity: certain demographic subgroups may have systematically higher flip probabilities than others, meaning that

they are more likely to be exposed to arbitrary, inconsistent predictions, with potentially less reliable explanations for the outcomes they receive [5].[2] We note that this is a separate form of group-level unfairness from the typical measures of statistical parity and error rate balance, since two groups may have equal positive prediction rates but very different flip probabilities (see Appendix G for an example).

5.3 Experiments on real data

As in previous sections, we perform our experiments on the German Credit, Adult, and Health datasets. We first calculate the flip probabilities for all individuals in each dataset for varying values of ϵ, and use these flip probabilities to perform four experiments.

First, we graph the overall (population average) flip probability for all three datasets for models sampled uniformly at random from the Rashomon set $R_N(\epsilon)$ as a function of ϵ, compared to sampling linear models from the Rashomon set (Section 4.3.1) and the models that optimize PPR, FPR, and TPR over the Rashomon set (Section 4.1). These graphs are shown in Appendix G, Figure 7.

Second, we use the flip probabilities to calculate the average fairness of the Rashomon set as a function of ϵ for all three datasets. We display the output in Appendix G, Figure 8. These differ from the results in Section 4.3 since these are the average fairness of models across the *entire* Rashomon set, not only from a sample of models, but we note the close correspondence between the sampled and entire-Rashomon-set results.

Third, in Figure 2(left) we turn to displaying empirical results about arbitrariness within the Rashomon set: we show how the chance of an individual experiencing a flip in their predictions in the Rashomon set (as a function of how close their Bayes-optimal probability p_i is to the threshold of 0.5) differs across different datasets and values of error tolerance ϵ. To do this, we compute the value of $C(\epsilon)$ for each dataset and ϵ, and then compute the flip probability $q(w) = \frac{1}{1+\exp(C\lvert 2p-1\rvert)}$ for a fine grid of p values.

Fourth, we show the disparities in average flip probability in the three datasets between protected and non-protected groups as a function of ϵ, suggesting that some groups have systematically higher exposure than others to arbitrary, inconsistent decisions. We compare uniform sampling to the models that optimize PPR, FPR, and TPR over the Rashomon set (as described in Section 4.1). Graphs for the German, Adult, and Health datasets are shown in Figure 2(right) and Appendix G, Figure 9.

5.4 Takeaways for policy and practice

- **Even a small error tolerance leads to a lot of individual flips.** We observe that, for uniform sampling, the overall flip probability tends to be substantially higher than the error tolerance ϵ. In Figure 2(center), for $\epsilon = .02$, we see that 12%, of predictions are flipped on average for the German dataset. In Appendix G, Figure 7, we see that this trend continues, with Adult and Health having 7% and 5% respectively. The overall flip probability for models that optimize fairness tends to be higher than the overall flip probability for uniform sampling, for lower ϵ values where the optimization method is not able to remove all of the disparity. Once the disparity is removed, the flip probability for optimal models levels off, while the flip probability for uniform sampling continues to increase with ϵ.
- **Increasing error tolerance ϵ not only increases the *number* of flips that occur, but *which individuals* are likely to get flipped: more "certain" cases get flipped with higher ϵ.** As ϵ increases, individuals with true probability p_i further and further away from a 50/50 coin toss, i.e., more "certain" cases of a positive or negative outcome get flipped. For example, as we see in Figure 2(left), in the German Credit dataset at low ϵ (red line) everyone who has a predicted Bayes probability below 0.4 or above 0.6 has near-zero chance of experiencing a flip in prediction, but at higher ϵ (brown line), we see that individuals with p_i between 0.1 and 0.9 have a non-negligible chance of getting flipped. Some prior work has suggested the normative view that individuals with higher certainty in their outcome should be flipped less often [2]—to the extent that this is true in certain contexts, it may be important to balance the flip probability over a threshold of certainty with the need to reduce outcome-based unfairness. We also see **large differences in flip probabilities between datasets**: for the same ϵ value, an individual with a given true probability p_i is much less likely to be flipped for German Credit as compared to Adult or Health.
- **Asymmetries in the underlying model—e.g. uneven distributions of predicted probabilities across groups—lead to disparities in flip probabilities across demographic groups.** As we can see from Figure 2(right) and Appendix G, Figure 9, all three datasets have disparities in their average flip probabilities, though it is not always the disadvantaged group with a higher flip probability. Since the flip probability for uniform random sampling is a function of an individual's weight within the data–i.e., their distance from the threshold probability of 0.5–the individuals who are flipped more often are those for whom the Bayes-optimal model is less certain of its prediction. In the Adult dataset, it is the advantaged group that has a higher density of true probabilities p_i around 0.5, meaning that they are more likely to get flipped. In the German Credit and Health datasets, the disadvantaged group has a higher density of $p_i \approx 0.5$, and thus a higher flip probability.
- **We observe across datasets that optimizing for fairness and uniform sampling lead to large differences in who is flipped**: while both optimization and uniform

[2] As a caveat, these flip probabilities will not necessarily translate to who is most likely to get flipped in any given search for a less discriminatory algorithm (LDA), as LDA searches will typically restrict the class of models prior to searching, and thus will not exactly match random sampling from within the Rashomon set. However, it does let us understand who is most likely to get flipped in the largest possible Rashomon set $R_N(\epsilon)$.

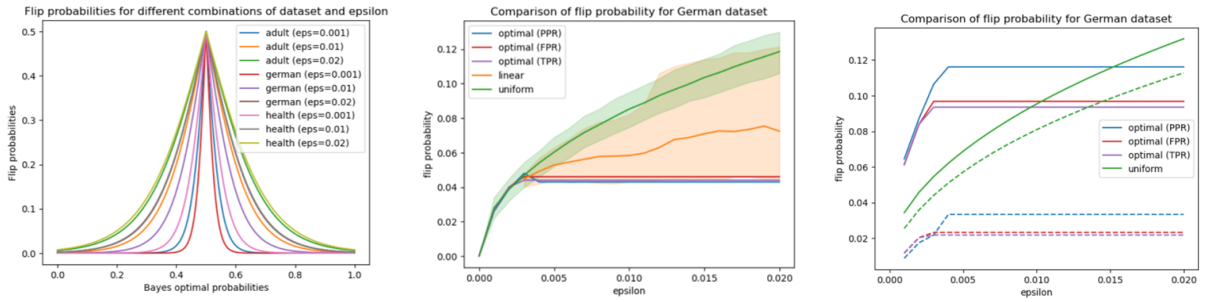

Figure 2: Left: Flip probability $q_{N,i}$ as a function of the Bayes-optimal probability p_i— in other words, how likely is an individual i to experience a change of prediction among models in the Rashomon set as a function of their true probability that $y_i = 1$? We show results for the German Credit, Adult, and Health datasets for $\epsilon \in \{0.001, 0.01, 0.02\}$, and see that there is large variation in flip probability distribution both as a function of dataset and ϵ. Center: Overall (population average) flip probability as a function of error tolerance ϵ for the German Credit dataset, for uniformly sampled models, linear models, and optimally fair models from the Rashomon set. For results for Adult and Health datasets, see Appendix G, Figure 7. Right: Group average flip probability, comparison between protected group (solid lines) and non-protected group (dashed lines), for the German Credit dataset, as a function of the error tolerance ϵ. Comparison of methods for optimizing PPR, FPR, and TPR (Section 4.1) and uniform random sampling (Section 4.2), over the Rashomon set $R_N(\epsilon)$. For results for Adult and Health datasets, see Appendix G, Figure 9.

sampling approaches tend to flip individuals who are near the decision boundary and thus have lower weights w_i, the optimization approaches also tend to flip individuals who are from the group that is less represented in the dataset and thus have higher values v_i, because flipping one person's prediction has a larger impact on the group average for the group that is smaller in size. In our datasets, the disadvantaged group (women for German and Adult, individuals over the age of 60 for Health) is also less represented. Thus, if the disadvantaged group is already flipped more than the advantaged group on average, because they tend to be closer to the decision boundary (as is the case for German and Health), optimizing for fairness will further exacerbate this disparity in who is receiving arbitrary and inconsistent predictions. For example, for German Credit, at $\epsilon = 0.004$ (the point at which PPR disparity is eliminated), the flip proportion for uniform random sampling is balanced (6.2% for women vs. 5.1% for men), but the model that optimizes PPR demonstrates a substantial disparity in who is flipped (11.6% for women vs. 3.3% for men). On the other hand, if the disadvantaged group is flipped substantially less than the advantaged group on average, because they tend to be farther from the decision boundary (as is the case for Adult), optimizing for fairness will instead mitigate this disparity.

6 Rashomon Set Size and Error Tolerance

In this section, we present results on the size of the Rashomon set and the distribution of how much of the error tolerance ϵ is used in the models of the Rashomon set. From these results, we suggest another set of takeaways— that when a company sets out to do a search for a less discriminatory algorithm (LDA), they should choose the highest error tolerance possible. However, especially when relying on repeated random sampling as an LDA search method, they should make sure they are comfortable with having a model that uses all of the error tolerance provided.

6.1 Rashomon set size

We derive an analytical expression for the asymptotic size of the Rashomon set, $|R_N(\epsilon)|$, as a function of the error tolerance ϵ, as the number of data records N that the Rashomon set is defined over goes to ∞. We note that $|R_N(\epsilon)|$ also depends on the distribution of weights $f(w)$ and thus is dataset-dependent. We provide the theorem here, with proof in Appendix C, Theorem C.13.

Theorem 6.1 (Asymptotic size of Rashomon set). *Given the preliminaries and assumptions in Section 5.1 above, let $R_N(\epsilon)$ denote the Rashomon set of models for error tolerance ϵ defined over data records $\langle d_1, \ldots, d_N \rangle$. Then*

$$\lim_{N\to\infty} \frac{\log |R_N(\epsilon)|}{N} = \log B(\epsilon),$$

where $B(\epsilon) = \exp\left(\int_0^\epsilon C(x)dx\right)$, $C(\epsilon) = g^{-1}(\epsilon)$, and $g(C) = \int_0^1 \frac{wf(w)}{1+\exp(Cw)}\, dw$.

In other words, for large N, the size of the Rashomon set $|R_N(\epsilon)|$ converges (in the sense above) to $B(\epsilon)^N$. Thus the size of the Rashomon set grows exponentially in N, the number of elements in the dataset, but the base of the exponential function B is an increasing function of ϵ. For $\epsilon = 0$ and $f(w)$ continuous, $|R_N(\epsilon)| = 1$ regardless of N, so $B = 1$. For sufficiently large ϵ, all 2^N flip vectors are in the Rashomon set, so $B = 2$. But the rate at which B increases from 1 to 2 with ϵ

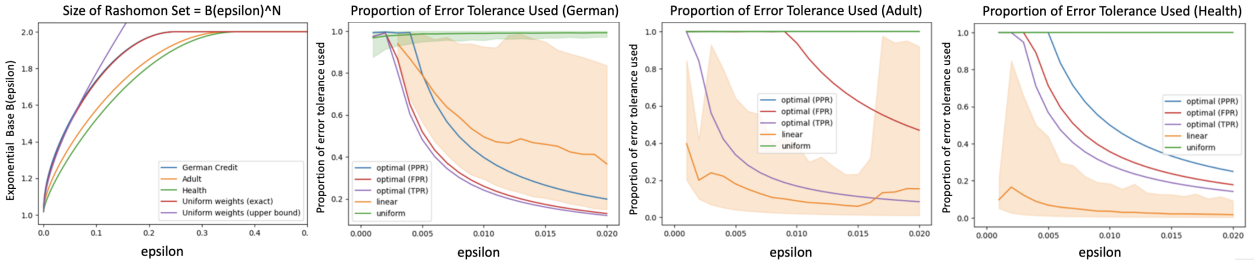

Figure 3: Left: Rashomon set size as a function of ϵ for Adult, German Credit, and Health datasets, and for uniformly distributed weights. Note that the German Credit and uniform weights curves coincide. The size of the Rashomon set is $|R_N(\epsilon)| = B(\epsilon)^N$, where the exponential base B (plotted here) ranges between 1 (for $\epsilon = 0$) and 2 (for large ϵ). We also separately plot $|R_N(\epsilon)|$ for each dataset in Appendix H, Figure 10. Right three figures: Proportion of error tolerance used, $\frac{\theta \cdot W_N}{N\epsilon}$, for the German, Adult, and Health datasets, as a function of the error tolerance ϵ. Comparison of methods for optimizing PPR (Section 4.1.1), optimizing FPR (Section 4.1.2), optimizing TPR (Section 4.1.2), uniform random sampling (Section 4.2), and sampling linear models (Section 4.3.1) over the Rashomon set $R_N(\epsilon)$.

will vary between datasets, depending on the distribution of weights $f(w)$, as we show in Figure 3(left). We give details on how to calculate $B(\epsilon)$, and therefore the size of the Rashomon set $B(\epsilon)^N$, in Appendix C, Theorem C.13. We also derive an exact value and an upper bound for $B(\epsilon)$ when the distribution of weights within the data records is uniform (Appendix C, Corollary C.14).

As we discuss in our takeaways, although a company has no control over N, it does have control over ϵ. As ϵ determines the base of the exponent B, this means that the size of the Rashomon set $|R_N(\epsilon)| = B(\epsilon)^N$ grows extremely quickly in ϵ as well.

6.2 Usage of error within the Rashomon set

We now show that as the number of data records N over which the Rashomon set $R_N(\epsilon)$ is defined goes to infinity (i.e., as the dataset grows large), as long as the error tolerance ϵ is sufficiently small (less than half of the average weight w_i), the models in the Rashomon set will use almost all of the error tolerance. That is, the average accuracy of a model in the Rashomon set will converge to the accuracy of the Bayes-optimal model minus ϵ.

Let $\overline{acc}(R_N(\epsilon))$ denote the average accuracy of models in $R_N(\epsilon)$, and let $acc_N(\theta_0)$ denote the accuracy of the Bayes-optimal classifier $f(x_i) = \mathbf{1}\{p_i > 0.5\}$ for data records $\langle d_1, \ldots, d_N \rangle$. The *average error tolerance used* is the difference $acc_N(\theta_0) - \overline{acc}(R_N(\epsilon))$, and must be less than or equal to ϵ. We now formally state the main result below, with proof in Appendix C (Theorem C.9):

Theorem 6.2 (Asymptotic use of the entire error tolerance). *Given the preliminaries and assumptions in Section 5.1 and the definitions above, let $R_N(\epsilon)$ denote the Rashomon set of models for error tolerance ϵ defined over data records $\langle d_1, \ldots, d_N \rangle$.*

Then as $N \to \infty$, the average error tolerance used by models in the Rashomon set converges to ϵ:

$$\lim_{N \to \infty} \left(acc_N(\theta_0) - \overline{acc}(R_N(\epsilon)) \right) = \epsilon.$$

This result implies that, for large N, there is a clear tradeoff between having a larger space of models to search over (since the size of the Rashomon set grows very rapidly with increasing ϵ) and the accuracy of the models one might find with this search, since the vast majority of models in the Rashomon set have accuracy very close to the Bayes-optimal accuracy minus ϵ. While we are not typically interested in the Rashomon set for very large values of ϵ where the assumption that ϵ is less than half of the average weight would not hold, we note that in such cases the entire error tolerance would not be used. Instead, as N becomes large, all flip probabilities would converge to 0.5, all or almost all of the 2^N possible flip vectors would be in the Rashomon set, and the average amount of error tolerance used would converge to half the average weight, which is less than ϵ.

6.3 Experiments on real data

6.3.1 Rashomon set size experiments. Given that the size of the Rashomon set $|R_N(\epsilon)|$ can be written as $B(\epsilon)^N$, where the exponential base B increases from 1 to 2 for increasing ϵ, we plot the values of B as a function of ϵ for the German Credit, Adult, and Health datasets in Figure 3(left). As noted above, we also derived both the exact value and the upper bound of B for uniformly distributed weights (Appendix C, Corollary C.14), and we plot these in Figure 3(left) for comparison. For small ϵ, the upper bound for uniformly distributed weights, $B(\epsilon) = \exp(\pi\sqrt{\epsilon/3})$, coincides closely with the exact values. We also plot the Rashomon set size $|R_N(\epsilon)|$ separately for the German Credit, Adult, and Health datasets in Appendix H, Figure 10. While we do not yet have a way of computing the (reduced) Rashomon set size when restricting our search to the space of linear models (L_2-penalized logistic regression) as described in Section 4.3.1, we can nevertheless examine what fraction

of the sampled linear models are in the Rashomon set as a function of ϵ. This is shown for the German Credit, Adult, and Health Datasets in Appendix H, Figure 11.

6.3.2 Use of error tolerance experiments. Given that, as $N \to \infty$, we expect the entire error tolerance ϵ to be used by models in the Rashomon set, we examine whether this holds for the three real-world datasets as well. In Figure 3(right), we plot the average proportion of the error tolerance used, $\frac{\theta \cdot W_N}{N\epsilon}$, for 950 flip vectors sampled uniformly at random from the Rashomon set $R_N(\epsilon)$, as described in Section 4.2, for each $\epsilon \in \{0.001, 0.002, \ldots, 0.02\}$. We also plot the 95% interquantile range for proportion of the error tolerance used, using the 2.5 and 97.5 percentiles of this distribution. For comparison, we also plot for each dataset in Figure 3(right) the proportion of the error tolerance used when (i) optimizing PPR, TPR, FPR, and over the Rashomon set, as in Sections 4.1.1 and 4.1.2, and (ii) searching over the set of linear models (L_2-penalized logistic regression) in the Rashomon set, as in Section 4.3.1.

6.4 Takeaways for policy and practice

- **Increasing ϵ drastically increases the size of the Rashomon set, especially for smaller ϵ: thus, companies searching for LDAs may want to set as large of an ϵ as possible to maximize the number of potential LDAs.** For example, in the German Credit dataset ($N = 1,000$), increasing ϵ from 0.005 to 0.02 moves the exponential base from 1.16 to 1.32 (Figure 3 (left)), increasing the Rashomon set size from 5×10^{65} to 6×10^{119} (Appendix H, Figure 10).
- **Especially when using random sampling to search for fairer models, our results suggest that most models found will use the entire error tolerance.** In Figure 3(right), we see that, as expected from Theorem 6.2, the average proportion of the error tolerance used by uniform random sampling over the entire Rashomon set is very close to 1 for all three datasets, and for the larger datasets (Adult and Health), even the 2.5 percentile of the distribution is virtually indistinguishable from 1. Similarly, for the optimization approaches, all of the error tolerance is used until the entire disparity is mitigated; then the proportion of error tolerance used decreases as $\frac{1}{\epsilon}$ for larger ϵ. Thus, while using a higher error tolerance ϵ could increase a company's opportunity to find fairer models within the Rashomon set, the company should be ready to use a model within the outer limits of that tolerance.
- **Caveats: Searching within particular model classes may *not* use up all of the error tolerance.** As we see from Figure 3(right), when restricting the search to linear models, the (non-exhaustive) set of linear models we found in the Rashomon set did not use up all of the error tolerance. In fact, the average proportion of the error tolerance used by randomly sampled linear models is much smaller than 1: about 40%, 15%, and 2% for German Credit, Adult, and Health datasets respectively.

7 Discussion

Before concluding, we discuss in more detail how our modeling set-up relates to practical searches that companies may make to look for less discriminatory models.

Models as Mappings. As explained in Section 3, we think of models in the Rashomon set as mappings from input features to binary decisions in $\{0, 1\}$. As this is an exhaustive set of all possible mappings that satisfy the accuracy constraint, this paper explores behavior of the *largest possible Rashomon set*. It may be the case that some of these models may not be reachable by typical training methods, such as using stochastic gradient descent to search for linear or deep models. However, we note that the model defined by a flip vector θ is reachable by a randomized classifier that deviates from the Bayes-optimal model (e.g., by randomizing labels as a function of the Bayes-optimal probability, rather than using a hard threshold at 0.5). To put it more practically, while some of the models in $R_N(\epsilon)$ may not be easy to find in a real-world search for LDAs if a company is limiting their search to a particular model class, e.g. linear models, this limitation is not a necessary choice. A company could search over many possible model classes, including flexible models that fit arbitrarily complex functions, and could also deviate from a learned model (such as their estimate of the Bayes-optimal classifier) to achieve fairness goals, e.g., by randomizing predictions or changing decision thresholds—by doing so, they could have the entire Rashomon set $R_N(\epsilon)$ at their disposal.

Relatedly, we clarify that the models discussed in this paper, and the techniques we present to find fairer models, generalize to previously unseen data, as we discuss in more detail in Section 4.3 and Appendix E. In particular, the models that we consider as alternatives in this paper represent systematic deviations from the Bayes-optimal classifier. Given that the Bayes-optimal classifier is estimated from labeled data and that its probabilistic predictions can be used to make classification decisions for previously unseen examples, a rule which defines how a given classifier deviates from Bayes-optimal will also generalize. As we discuss in Appendix E, our optimization approaches are equivalent to defining an optimal prediction threshold for each class, and our uniform sampling approach is equivalent to randomizing labels as a function of the Bayes-optimal probability. For sampling, the model disagrees with the Bayes-optimal prediction for each new data record with probability $q_i = \frac{1}{1+\exp(Cw_i)}$, as shown in Theorem 5.1. In each case, the parameters (prediction thresholds for optimization and the constant C for sampling) can be learned from one (unlabeled) partition of the data and then applied to optimize fairness over the Rashomon set, or to uniformly sample from the Rashomon set, for a different data partition. We demonstrate that our sampling and optimization methods generalize well to previously unseen data in Appendix E.

Bayes-optimal model. Finally, we note that while we assume access to the Bayes-optimal classifier for our theoretical results, in practice, our findings do not rely on having the exact Bayes-optimal model. Concretely, we believe that it is reasonable for a company to start from their estimated "most accurate" classifier, whether learned via a flexible classification approach or even within a specific model class, and then use our optimization approaches to identify class-specific decision thresholds that optimize fairness while obtaining accuracy within ϵ of their original model. Since we define our Rashomon set in terms of deviation from a baseline model, it is important to find as accurate as possible a baseline model when using this method of Rashomon set exploration in practice. We show in Appendix E that assuming access to a reasonably-sized training set, our results do not change substantially using estimates of the Bayes-optimal model learned from different partitions of the data. Nevertheless, if there is not enough labeled data to estimate a good model, then the Rashomon set and fairer models found in it may not be accurate or useful.

Overall, the modeling choices we make serve the main goal of this work, which is to more deeply understand the fundamental properties of the Rashomon set and the importance of intentionally searching for fairer models. In particular, they allow us to see what is *possible* to achieve within the Rashomon set with maximum flexibility, allowing us to see how much we can strive to accomplish. At the same time, while our results may deviate somewhat from what people observe in practice (e.g., if they search only through limited model classes), we believe our work can provide companies with a fairness goal to strive for, and suggest specific approaches that might help them approach that goal.

8 Conclusion

We introduce key results that help us to understand the largest possible Rashomon set, from the average fairness of models within the Rashomon set, to the probability of individuals having their prediction changed across all models in the set, and the size of the Rashomon set. These results lead us to several takeaways: (1) it is critical to search for fair models within the Rashomon set (to be *intentional* about fairness); (2) the arbitrariness of prediction within the Rashomon set changes drastically depending on the dataset and the error tolerance ϵ; and (3) companies should think carefully about setting ϵ when searching for fairer models within the Rashomon set, balancing flexibility of the search with accuracy of the resulting models. We hope this work shows the importance of searching for fair models within the Rashomon set, and sheds light on how to balance fairness gains with risks of arbitrariness.

Acknowledgments

This work was partially supported by the NSF Program on Fairness in Artificial Intelligence in Collaboration with Amazon, grant IIS-2040898.

References

[1] Barry Becker and Ronny Kohavi. 1996. Adult Dataset. https://archive.ics.uci.edu/dataset/2/adult
[2] Emily Black and Matt Fredrikson. 2021. Leave-One-out Unfairness. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (Virtual Event, Canada) *(FAccT '21)*. Association for Computing Machinery, New York, NY, USA, 285–295. doi:10.1145/3442188.3445894
[3] Emily Black, John Logan Koepke, Pauline Kim, Solon Barocas, and Mingwei Hsu. forthcoming 2024. Less Discriminatory Algorithms. *Geo. L. J.* 113 (forthcoming 2024).
[4] Emily Black, Klas Leino, and Matt Fredrikson. 2022. Selective Ensembles for Consistent Predictions. In *International Conference on Learning Representations*.
[5] Emily Black, Manish Raghavan, and Solon Barocas. 2022. Model Multiplicity: Opportunities, Concerns, and Solutions. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency* (Seoul, Republic of Korea) *(FAccT '22)*. Association for Computing Machinery, New York, NY, USA, 850–863. doi:10.1145/3531146.3533149
[6] Emily Black, Zifan Wang, and Matt Fredrikson. 2022. Consistent Counterfactuals for Deep Models. In *International Conference on Learning Representations*.
[7] Leo Breiman. 2001. Statistical modeling: The two cultures (with comments and a rejoinder by the author). 199–231 pages.
[8] Relman Colfax. 2022. Fair Lending Monitorship of Upstart Network's Lending Model: Third Report of the Independent Monitor. https://www.relmanlaw.com/media/cases/1333_PUBLIC%20Upstart%20Monitorship%203rd%20Report%20FINAL.pdf.
[9] A Feder Cooper, Katherine Lee, Madiha Zahrah Choksi, Solon Barocas, Christopher De Sa, James Grimmelmann, Jon Kleinberg, Siddhartha Sen, and Baobao Zhang. 2024. Arbitrariness and social prediction: The confounding role of variance in fair classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38. 22004–22012.
[10] Amanda Coston, Ashesh Rambachan, and Alexandra Chouldechova. 2021. Characterizing Fairness Over the Set of Good Models Under Selective Labels. 2144–2155 pages. https://proceedings.mlr.press/v139/coston21a.html
[11] Kathleen Creel and Deborah Hellman. 2022. The algorithmic leviathan: Arbitrariness, fairness, and opportunity in algorithmic decision-making systems. *Canadian Journal of Philosophy* 52, 1 (2022), 26–43.
[12] Alexander D'Amour, Katherine Heller, Dan Moldovan, Ben Adlam, Babak Alipanahi, Alex Beutel, Christina Chen, Jonathan Deaton, Jacob Eisenstein, Matthew D Hoffman, et al. 2020. Underspecification presents challenges for credibility in modern machine learning.
[13] Jiayun Dong and Cynthia Rudin. 2019. Variable importance clouds: A way to explore variable importance for the set of good models.
[14] Leonardo Ferreira. 2017. German Credit Risk - With Target. https://www.kaggle.com/datasets/kabure/german-credit-data-with-risk
[15] Stuart Geman and Donald Geman. 1984. Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 6, 6 (1984), 721–741. doi:10.1109/TPAMI.1984.4767596
[16] Talia B Gillis. 2021. The input fallacy. *Minn. L. Rev.* 106 (2021), 1175.
[17] Talia B Gillis, Vitaly Meursault, and Berk Ustun. 2024. Operationalizing the Search for Less Discriminatory Alternatives in Fair Lending. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*. 377–387.
[18] Juan Felipe Gomez, Caio Machado, Lucas Monteiro Paes, and Flavio Calmon. 2024. Algorithmic Arbitrariness in Content Moderation. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*. 2234–2253.
[19] Daniel E Ho and Alice Xiang. 2020. Affirmative algorithms: The legal grounds for fairness as awareness. 134 pages.
[20] Hans Hofmann. 1994. German Credit Dataset. https://archive.ics.uci.edu/dataset/144/statlog+german+credit+data
[21] Pauline T Kim. 2022. Race-aware algorithms: Fairness, nondiscrimination and affirmative action. *Cal. L. Rev.* 110 (2022), 1539.
[22] Benjamin Laufer, Manisch Raghavan, and Solon Barocas. 2024. Fundamental Limits in the Search for Less Discriminatory Algorithms–and How to Avoid Them. *arXiv preprint arXiv:2412.18138* (2024).
[23] Charles T. Marx, Flávio du Pin Calmon, and Berk Ustun. 2019. Predictive Multiplicity in Classification. http://arxiv.org/abs/1909.06677
[24] Kota Mata, Kentaro Kanamori, and Hiroki Arimura. 2022. Computing the collection of good models for rule lists. *arXiv preprint arXiv:2204.11285* (2022).
[25] Richard Merkin. 2012. Heritage Health Prize. https://www.kaggle.com/c/hhp/data
[26] Martin Pawelczyk, Klaus Broelemann, and Gjergji. Kasneci. 2020. On Counterfactual Explanations under Predictive Multiplicity.

[27] Randy Axelrod Phil Brierley, David Vogel. 2012. Heritage Provider Network Health Prize Round 1 Milestone Prize How We Did It – Team Market Makers. https://foreverdata.org/1015/content/milestone1-2.pdf
[28] Relman Colfax. 2024. Fair Lending Monitorship of Upstart Network's Lending Model: Fourth and Final Report of the Independent Monitor. https://www.relmanlaw.com/media/news/1512_Upstart%20Final%20Report.pdf.
[29] Kit T Rodolfa, Hemank Lamba, and Rayid Ghani. 2021. Empirical observation of negligible fairness–accuracy trade-offs in machine learning for public policy. 896–904 pages.
[30] Aaron Roth, Alexander Tolbert, and Scott Weinstein. 2023. Reconciling Individual Probability Forecasts. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*. 101–110.
[31] Cynthia Rudin, Chudi Zhong, Lesia Semenova, Margo Seltzer, Ronald Parr, Jiachang Liu, Srikar Katta, Jon Donnelly, Harry Chen, and Zachery Boner. 2024. Amazing Things Come From Having Many Good Models. arXiv:2407.04846 [cs.LG] https://arxiv.org/abs/2407.04846
[32] Lesia Semenova, Cynthia Rudin, and Ronald Parr. 2019. A study in Rashomon curves and volumes: A new perspective on generalization and model simplicity in machine learning.
[33] Lesia Semenova, Cynthia Rudin, and Ronald Parr. 2022. On the existence of simpler machine learning models. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. 1827–1858.
[34] Jamelle Watson-Daniels, David C Parkes, and Berk Ustun. 2023. Predictive multiplicity in probabilistic classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 37. 10306–10314.
[35] Rui Xin, Chudi Zhong, Zhi Chen, Takuya Takagi, Margo Seltzer, and Cynthia Rudin. 2022. Exploring the whole rashomon set of sparse decision trees. *Advances in neural information processing systems* 35 (2022), 14071–14084.

A Optimizing Fairness over the Rashomon Set

In this section, we propose efficient algorithms to find (i) the exact fairest model in the Rashomon set as measured by positive prediction rate (PPR) disparity, and (ii) a model that is guaranteed to have false positive rate (FPR) or true positive rate (TPR) disparity within $O(\frac{1}{N})$ of the fairest model in the Rashomon set.

A.1 Optimizing for statistical parity

In this sub-section, we propose an efficient knapsack solution (Algorithm 1), to find the exact fairest model that minimizes disparities in positive prediction rate (PPR) over the Rashomon set $R_N(\epsilon)$, in $O(N \log N)$ time.

Our first step is to derive expressions for the FPR and TPR disparities corresponding to a given flip vector θ. To do so, let P_A and P_B be the vectors of Bayes-optimal probabilities p_i for subgroups A (the *protected class*, data records d_i with sensitive attribute value $A_i = a$) and B (the *non-protected class*, data records d_i with sensitive attribute value $A_i \neq a$) respectively. Let $F^{\text{opt}} = \langle F_A^{\text{opt}}, F_B^{\text{opt}} \rangle$ denote the vector of Bayes-optimal binary predictions $f_{\text{opt}}(x_i)$, and let $F = \langle F_A, F_B \rangle$ denote the vector of binary predictions $f(x_i)$ corresponding to flip vector $\theta = \langle \theta_A, \theta_B \rangle$. We note that $F_A = F_A^{\text{opt}} \odot (1-\theta_A) + (1-F_A^{\text{opt}}) \odot \theta_A$, and $F_B = F_B^{\text{opt}} \odot (1-\theta_B) + (1-F_B^{\text{opt}}) \odot \theta_B$, where $\odot$ denotes element-wise product.

We can then define the positive prediction rate disparity as:

$$
\begin{aligned}
\text{disparity}_{PPR} &= |\mathbf{E}[f(x_i) \mid d_i \in A] - \mathbf{E}[f(x_i) \mid d_i \in B]| \\
&= |\Pr(f(x_i) = 1 \mid d_i \in A) - \Pr(f(x_i) = 1 \mid d_i \in B)| \\
&= \left| \frac{||F_A||_1}{|A|} - \frac{||F_B||_1}{|B|} \right| \\
&= \left| \frac{F_A \cdot \mathbf{1}}{|A|} - \frac{F_B \cdot \mathbf{1}}{|B|} \right| \\
&= \left| \frac{\left(F_A^{\text{opt}} \odot (1-\theta) + (1-F_A^{\text{opt}}) \odot \theta\right) \cdot \mathbf{1}}{|A|} \right. \\
&\qquad \left. - \frac{\left(F_B^{\text{opt}} \odot (1-\theta) + (1-F_B^{\text{opt}}) \odot \theta\right) \cdot \mathbf{1}}{|B|} \right| \\
&= \left| \frac{F_A^{\text{opt}} \cdot (1-\theta) + (1-F_A^{\text{opt}}) \cdot \theta}{|A|} \right. \\
&\qquad \left. - \frac{F_B^{\text{opt}} \cdot (1-\theta) + (1-F_B^{\text{opt}}) \cdot \theta}{|B|} \right|
\end{aligned}
$$

As noted in Section 4.1, we can express the minimization of PPR disparity over flip vectors θ, subject to the constraint that θ is in the Rashomon set $R_N(\epsilon)$, as a *knapsack problem*, where each data record d_i has a weight $w_i = |2p_i - 1|$ and a value v_i, and $\theta_i = 1$ corresponds to the inclusion of element i in the knapsack, adding w_i to the total weight and v_i to the total value. The 0-1 knapsack problem is then the constrained optimization with capacity $N\epsilon$: $\max \sum_i \theta_i v_i$ subject to $\theta_i \in \{0, 1\}$ and $\sum_i \theta_i w_i \leq N\epsilon$.

We now consider the expression for v_i, the change in PPR disparity when the prediction $f(x_i)$ is flipped (i.e., when θ_i is changed from 0 to 1). Assume without loss of generality that subgroup A has higher PPR, $\frac{||F_A||_1}{|A|} > \frac{||F_B||_1}{|B|}$. Then we see from the expression for PPR disparity above that flipping a prediction in group A from 1 to 0, or flipping a prediction in group B from 0 to 1, reduces disparity by $\frac{1}{|A|}$ or $\frac{1}{|B|}$ respectively, while other flips increase disparity. To see this, for a data record $d_i \in A$ with $F_i^{\text{opt}} = 1$,

$$
\begin{aligned}
v_i = &\left. \frac{F_A^{\text{opt}} \cdot (1-\theta) + (1-F_A^{\text{opt}}) \cdot \theta}{|A|} \right|_{\theta=0, F_A^{\text{opt}}=1} \\
&- \left. \frac{F_A^{\text{opt}} \cdot (1-\theta) + (1-F_A^{\text{opt}}) \cdot \theta}{|A|} \right|_{\theta=1, F_A^{\text{opt}}=1} = \frac{1}{|A|}. \qquad (1)
\end{aligned}
$$

We note that v_i for other cases can be calculated similarly. Thus we can write the value of element i for the knapsack problem as $v_i = \frac{1}{|A|}$ if $d_i \in A$ and $F_i^{\text{opt}} = 1$, $v_i = \frac{1}{|B|}$ if $d_i \in B$ and $F_i^{\text{opt}} = 0$, and $v_i = 0$ otherwise.

We now make the key observation that enables our efficient knapsack algorithm: there are only two distinct values $v_i > 0$, $\frac{1}{|A|}$ and $\frac{1}{|B|}$ for group A and B respectively. Thus, the optimal knapsack solution will consist of the k_A lowest-weight items from group A and the k_B lowest-weight items from group B, for some k_A and k_B.

Optimal values of k_A and k_B (i.e., those values that most reduce the disparity) could be calculated by a $O(N^2)$ brute-force search across all combinations of k_A and k_B that fit the capacity. However, through incremental search, that is, by keeping track of the optimal k_B for a given k_A, one can incrementally update k_B for $k_A - 1$ by adding the remaining lowest weight B items until the capacity is full, resulting in an incremental linear $O(N)$ search. Thus the run time is dominated by the $O(N \log N)$ sorting of items by weight. We present the algorithm below.

A.2 Optimizing for error rate balance

In this sub-section, we propose an efficient, $O(N \log N)$ fractional knapsack solution (Algorithm 2), to find the model that minimizes disparities in false positive rate (FPR) or true positive rate (TPR) over the Rashomon set $R_N(\epsilon)$, to within $O(\frac{1}{N})$ of the optimal disparity, in $O(N \log N)$ time.

Our first step is to derive expressions for the FPR and TPR disparities corresponding to a given flip vector θ. To do so, as in Appendix A.1, let P_A and P_B be the vectors of Bayes-optimal probabilities p_i for subgroups A (the *protected class*, data records d_i with sensitive attribute value $A_i = a$) and B (the *non-protected class*, data records d_i with sensitive attribute value $A_i \neq a$) respectively. Let $F^{\text{opt}} = \langle F_A^{\text{opt}}, F_B^{\text{opt}} \rangle$ denote the vector of Bayes-optimal binary predictions $f_{\text{opt}}(x_i)$, and let

Algorithm 1 0-1 Knapsack Algorithm for minimizing *PPR* disparity

Given: data records $D_N = (x_i, y_i)|_{i=1}^N$, Bayes-optimal (true) probabilities $P_N = p_i|_{i=1}^N$, capacity $N\epsilon$
Output: final disparity fd, flip vector θ

Calculate the value v_i of each record (either $\frac{1}{|A|}$, $\frac{1}{|B|}$, or 0) as described in the text above
Calculate the weight of each record i, $w_i = |2p_i - 1|$
Calculate the initial disparity in the data, id

if $id = 0$ or $\epsilon = 0$ **then**
  return $\theta = 0, fd = id$
end if

Calculate weights W_A, sorted in ascending order, along with their indices I_A for records in A with $v_i > 0$
Calculate weights W_B, sorted in ascending order, along with their indices I_B for records in B with $v_i > 0$

Calculate the maximum number of items in A with $v_i > 0$, $maxA$, that fit the capacity, adding items in ascending order of weight
Calculate the maximum number of items in B with $v_i > 0$, $maxB$, that fit the capacity along with the $maxA$ items, adding items in ascending order of weight

Initialize the best value ($bestval$) to $\frac{maxA}{|A|} + \frac{maxB}{|B|}$
Initialize the best number of items in A (k_A) to $maxA$ and the best number of items in B (k_B) to $maxB$

for a from $maxA - 1$ to 0 **do**
  Remove the highest-weight item in A with $v_i > 0$, and add items in B with $v_i > 0$ until the capacity is filled, adding items in ascending order of weight. Let b be the total number of items in B with $v_i > 0$ that have been added.
  if $\frac{a}{|A|} + \frac{b}{|B|} > bestval$ **then**
    Set $k_A = a, k_B = b, bestval = \frac{a}{|A|} + \frac{b}{|B|}$
  end if
end for

Calculate optimal flip vector θ and final disparity fd, setting $\theta_i = 1$ for the k_A lowest-weight items in A with $v_i > 0$ and the k_B lowest-weight items in B with $v_i > 0$.
return θ, fd

$F = \langle F_A, F_B \rangle$ denote the vector of binary predictions $f(x_i)$ corresponding to flip vector $\theta = \langle \theta_A, \theta_B \rangle$. We note that $F_A = F_A^{\text{opt}} \odot (1 - \theta_A) + (1 - F_A^{\text{opt}}) \odot \theta_A$, and $F_B = F_B^{\text{opt}} \odot (1 - \theta_B) + (1 - F_B^{\text{opt}}) \odot \theta_B$, where $\odot$ denotes element-wise product.

We can then define the false positive rate disparity and true positive rate disparity as:

$$\begin{aligned}\text{disparity}_{FPR} &= |\mathbf{E}[f(x_i) \mid d_i \in A, y_i = 0] - \mathbf{E}[f(x_i) \mid d_i \in B, y_i = 0]| \\ &= |\Pr(f(x_i) = 1 \mid d_i \in A, y_i = 0) - \Pr(f(x_i) = 1 \mid d_i \in B, y_i = 0)| \\ &= \left| \frac{\Pr(f(x_i) = 1, y_i = 0 \mid d_i \in A)}{\Pr(y_i = 0 \mid d_i \in A)} - \frac{\Pr(f(x_i) = 1, y_i = 0 \mid d_i \in B)}{\Pr(y_i = 0 \mid d_i \in B)} \right| \\ &= \left| \frac{(1 - P_A) \cdot F_A}{||1 - P_A||_1} - \frac{(1 - P_B) \cdot F_B}{||1 - P_B||_1} \right|.\end{aligned}$$

$$\begin{aligned}\text{disparity}_{TPR} &= |\mathbf{E}[f(x_i) \mid d_i \in A, y_i = 1] - \mathbf{E}[f(x_i) \mid d_i \in B, y_i = 1]| \\ &= |\Pr(f(x_i) = 1 \mid d_i \in A, y_i = 1) - \Pr(f(x_i) = 1 \mid d_i \in B, y_i = 1)| \\ &= \left| \frac{\Pr(f(x_i) = 1, y_i = 1 \mid d_i \in A)}{\Pr(y_i = 1 \mid d_i \in A)} - \frac{\Pr(f(x_i) = 1, y_i = 1 \mid d_i \in B)}{\Pr(y_i = 1 \mid d_i \in B)} \right| \\ &= \left| \frac{P_A \cdot F_A}{||P_A||_1} - \frac{P_B \cdot F_B}{||P_B||_1} \right|.\end{aligned}$$

We now compute the values v_i (the change in disparity when the prediction $f(x_i)$ is flipped, i.e., when θ_i is changed from 0 to 1) for FPR and TPR respectively.

For FPR, assume without loss of generality that subgroup A has higher FPR, $\frac{(1-P_A)\cdot F_A}{||1-P_A||_1} > \frac{(1-P_B)\cdot F_B}{||1-P_B||_1}$. Then flipping a prediction in group A from 1 to 0, or flipping a prediction in group B from 0 to 1, reduces the disparity by $\frac{1-p_i}{||1-P_A||_1}$ or $\frac{1-p_i}{||1-P_B||_1}$ respectively, while other flips increase disparity. Thus we can write the value of element i for the knapsack problem as $v_i = \frac{1-p_i}{||1-P_A||_1}$ if $d_i \in A$ and $F_i^{\text{opt}} = 1, v_i = \frac{1-p_i}{||1-P_B||_1}$ if $d_i \in B$ and $F_i^{\text{opt}} = 0$, and $v_i = 0$ otherwise.

For TPR, assume without loss of generality that subgroup A has higher TPR, $\frac{P_A \cdot F_A}{||P_A||_1} > \frac{P_B \cdot F_B}{||P_B||_1}$. Then flipping a prediction in group A from 1 to 0, or flipping a prediction in group B from 0 to 1, reduces the disparity by $\frac{p_i}{||P_A||_1}$ or $\frac{p_i}{||P_B||_1}$ respectively, while other flips increase disparity. Thus we can write the value of element i for the knapsack problem as $v_i = \frac{p_i}{||P_A||_1}$ if $d_i \in A$ and $F_i^{\text{opt}} = 1, v_i = \frac{p_i}{||P_B||_1}$ if $d_i \in B$ and $F_i^{\text{opt}} = 0$, and $v_i = 0$ otherwise.

To minimize FPR or TPR disparity over the Rashomon set $R_N(\epsilon)$, we note that elements have more than two distinct values, so we cannot apply the solution for PPR above. Instead, we approximate the 0-1 knapsack problem with the fractional knapsack problem: max $\sum_i \theta_i v_i$ subject to $\theta_i \in [0, 1]$ and $\sum_i \theta_i w_i \leq N\epsilon$. The standard solution to the fractional knapsack, which requires $O(N \log N)$ time, adds elements to the knapsack (setting $\theta_i = 1$) in descending order of their ratio $\frac{v_i}{w_i}$ until no further elements can be (fully) added, then adds a fraction of the next element ($0 < \theta_i < 1$) to fill the remaining capacity. Rather than adding the fractional element, we show that it would reduce disparity by an amount $\theta_i v_i$ that is $O(\frac{1}{N})$.

To see this, we note for FPR disparity, for an individual in group A, that $\theta_i < 1$, and $v_i = \frac{1-p_i}{||1-P_A||_1} < \frac{1}{||1-P_A||_1} = \frac{1}{N \Pr(y_i=0, d_i \in A)} = O(\frac{1}{N})$. Therefore $\theta_i v_i = O(\frac{1}{N})$. The other cases, for TPR disparity and for group B, proceed similarly.

Finally, since the fractional knapsack solution $\sum \theta_i v_i$ is an upper bound on the 0-1 knapsack solution, we know that our solution (excluding the fractional element) reduces disparity to within $O(\frac{1}{N})$ of the optimal disparity.

Thus we propose an $O(N \log N)$ fractional knapsack algorithm to find the final disparity and flip vector. The algorithm is a linear scan of values and weights, sorted by the ratio of their value to their weight. Thus the run time is dominated by the $O(N \log N)$ sorting of items by their ratio of value to weight. We present the algorithm below.

Algorithm 2 Fractional Knapsack Algorithm for minimizing *FPR* or *TPR* disparity

Given: data records $D_N = (x_i, y_i)|_{i=1}^N$, Bayes-optimal (true) probabilities $P_N = p_i|_{i=1}^N$, capacity $N\epsilon$
Output: final disparity fd, flip vector θ

Calculate the value v_i of each record as described in the text above
Calculate the weight of each record i, $w_i = |2p_i - 1|$
Calculate the initial disparity in the data, id

if $id = 0$ or $\epsilon = 0$ **then**
  return $\theta = 0, fd = id, fracVal = 0$
end if

Calculate weights W and values V, along with their indices I, of records with $v_i > 0$, sorted by $\frac{v_i}{w_i}$ in descending order

Initialize record index variable i, total weight $totWei$, and total value $totVal$ to 0

while $totVal < id$ **and** $totWei < Capacity$ **and** $i < len(I)$ **do**
  # Attempt to add next element i with $v_i > 0$; note that elements are added in descending order of $\frac{v_i}{w_i}$.
  if $totWei + w_i \leq Capacity$ **then**
    Add weight of element i to total weight $totWei$
    Add value of element i to total value $totVal$
    Increment i by 1
  else
    Calculate the fractional value of element i that would fill the entire capacity, $fracVal = \left(\frac{Capacity - totWei}{w_i}\right) v_i$
    break
  end if
end while

Calculate flip vector θ, setting $\theta_i = 1$ for all elements i added to the knapsack, excluding the fractional element.
Calculate final disparity $fd = id - totVal$
Note that $fracVal$ is an upper bound on the difference between fd and the true optimal disparity.
return $fd, \theta, fracVal$

B Gibbs sampling algorithm for uniform sampling from the Rashomon set

Algorithm 3 Sampling Flip Vectors Uniformly at Random from the Rashomon Set

Given: error tolerance ϵ; number of data records N; weight vector $W_N = \langle w_1, \ldots, w_N \rangle$. Note that $w_i = |2p_i - 1|$, where p_i is the Bayes-optimal probability $\Pr(y = 1 \mid x = x_i)$ for data record d_i, $i \in \{1, \ldots, N\}$.
Initialize Θ as an empty list
Initialize $\theta = \theta_0$, where θ_0 is the length-N binary flip vector consisting of all zeros.
for $t = 1$ to T **do**
  for i in random permutation of $\{1, \ldots, N\}$ **do**
    Calculate current amount of error tolerance used, $E_{\text{current}} = \frac{\theta \cdot W_N}{N}$
    Calculate $\Pr(\theta_i = 1 \mid \theta_{-i})$:
    if $E_{\text{current}} + (1 - \theta_i)\frac{w_i}{N} \leq \epsilon$ **then**
      $\Pr(\theta_i = 1 \mid \theta_{-i}) = \frac{1}{2}$
    else
      $\Pr(\theta_i = 1 \mid \theta_{-i}) = 0$
    end if
    Sample θ_i from Bernoulli($\Pr(\theta_i = 1 \mid \theta_{-i})$)
  end for
  if $t > B$ and $(t - B) \bmod K == 0$ **then**
    Append θ to Θ
  end if
end for
return Θ

The definition of the Rashomon set $R_N(\epsilon)$ in Section 3 suggests that a simple *rejection sampling* approach could be used to draw models uniformly at random from the Rashomon set. That is, one could draw a binary flip vector $\theta \in \{0, 1\}^N$ uniformly at random from the set of all 2^N possible flip vectors by drawing $\theta_i \sim \text{Bernoulli}(0.5)$ for all $i \in \{1 \ldots N\}$, and then keep only those vectors θ that are in the Rashomon set, i.e., with $\frac{\theta \cdot W_N}{N} \leq \epsilon$. The problem with this simple approach is that, as N increases, the probability that θ is in the Rashomon set goes to 0. This can be seen from Appendix C, Theorem C.13: for ϵ less than half the average weight, $B(\epsilon) < 2$, and $\lim_{N\to\infty} \frac{|R_N(\epsilon)|}{2^N} = \lim_{N\to\infty} \frac{B(\epsilon)^N}{2^N} = 0$.

Thus we propose an alternative approach based on *Gibbs sampling* [15]. The key idea is to sequentially sample one element θ_i of the flip vector at a time, conditional on all the other elements θ_{-i}. While we do not have a closed form for the joint distribution of $\langle \theta_1, \ldots, \theta_N \rangle$ for $\theta \in R_N(\epsilon)$, computing the *conditional* distribution of θ_i given θ_{-i} is straightforward. Let $\theta_{i=0} = \langle \theta_1, \ldots, \theta_{i-1}, 0, \theta_{i+1}, \ldots, \theta_N \rangle$ and $\theta_{i=1} = \langle \theta_1, \ldots, \theta_{i-1}, 1, \theta_{i+1}, \ldots, \theta_N \rangle$. Then we know that $\frac{\theta_{i=0} \cdot W_N}{N} \leq \frac{\theta \cdot W_N}{N} \leq \frac{\theta_{i=1} \cdot W_N}{N}$. This implies that, if $\theta \in R_N(\epsilon)$ and $\theta_i = 1$, then $\theta_{i=0}$ and $\theta_{i=1}$ are both in the Rashomon set, so $\Pr(\theta_i = 1 \mid \theta_{-i}) = \frac{1}{2}$. If $\theta \in R_N(\epsilon)$ and $\theta_i = 0$, then $\theta_{i=0} \in R_N(\epsilon)$, but we must check whether $\theta_{i=1} \in R_N(\epsilon)$, i.e.,

whether $\frac{\theta \cdot W_N + w_i}{N} \leq \epsilon$. If so, then $\Pr(\theta_i = 1 \mid \theta_{-i}) = \frac{1}{2}$, and if not, then $\Pr(\theta_i = 1 \mid \theta_{-i}) = 0$.

Given this simple and computationally efficient conditional sampling step, our Gibbs sampling approach starts with the zero vector θ_0, which is guaranteed to be in the Rashomon set, and iteratively samples $\theta_i \sim \text{Bernoulli}(p)$, where $p = \Pr(\theta_i = 1 \mid \theta_{-i})$ as described above, for each $i \in \{1, \ldots, N\}$. To ensure uncorrelated samples from the joint distribution, we take one sample every 10 iterations (where one iteration includes resampling all N elements of θ in randomly permuted order), after an initial burn-in period of 500 iterations. For each dataset and each value of ϵ considered, we run 10,000 iterations of Gibbs sampling, resulting in 950 samples.

Algorithm 3 presents the pseudocode for our Gibbs sampling approach, enabling us to sample length-N binary flip vectors uniformly at random from the Rashomon set $R_N(\epsilon)$. The sampling algorithm follows the idea [15], in which the Markov chain is the sequence of flip vectors $\theta^{(0)}, \theta^{(1)}, \ldots, \theta^{(T)}$ generated as the algorithm progresses. Each $\theta^{(t)}$ is a point $\theta \in \{0,1\}^N$ that does not violate the Rashomon set constraint $\frac{\theta \cdot W_N}{N} \leq \epsilon$, and thus $\theta^{(t)} \in R_N(\epsilon)$.

Specifically, we start by initializing the flip vectors Θ as an empty list. We then initialize flip vector $\theta = \theta_0$, the length-N binary vector of zeros, which is guaranteed to be in the Rashomon set since $\frac{\theta_0 \cdot W_N}{N} = 0$.

Throughout $T = 10,000$ iterations, where each iteration involves resampling each of the N elements of θ in randomly permuted order, we keep track of the current amount of error tolerance used, $E_{\text{current}} = \frac{\theta \cdot W_N}{N}$, which can be done through incremental updates of E_{current} whenever an element θ_i is modified. To resample element θ_i, we first compute $\Pr(\theta_i = 1 \mid \theta_{-i})$, the probability of $\theta_i = 1$ conditional on the current values of the other elements of θ, and then draw $\theta_i \sim \text{Bernoulli}(\Pr(\theta_i = 1 \mid \theta_{-i}))$. To compute $\Pr(\theta_i = 1 \mid \theta_{-i})$, we note that all flip vectors in the Rashomon set must be equally likely to be drawn. Thus, if $\theta_{i=0}$ and $\theta_{i=1}$ are both in the Rashomon set, we know $\Pr(\theta_i = 1 \mid \theta_{-i}) = \frac{1}{2}$, while if $\theta_{i=0}$ is in the Rashomon set and $\theta_{i=1}$ is not, then $\Pr(\theta_i = 1 \mid \theta_{-i}) = 0$. We note that $\theta_{i=0}$ will always be in the Rashomon set, as described in the main text. To check if $\theta_{i=1}$ is in the Rashomon set, we must evaluate whether $\frac{\theta_{i=1} \cdot W_N}{N} = E_{\text{current}} + (1 - \theta_i)\frac{w_i}{N} \leq \epsilon$.

To ensure that each sampled flip vector θ is drawn independently from the joint distribution of $\langle \theta_1, \ldots, \theta_N \rangle$, we begin recording θ only after the number of iterations exceeds the burn-in period $B = 500$, and thereafter sample one value of θ every $K = 10$ iterations (i.e., at iterations 510, 520, …). When the algorithm terminates, all recorded flip vectors are appended to Θ, resulting in the final list of sampled vectors.

We see that each iteration requires stepping through the $O(N)$ data records. For each data record, we perform an $O(1)$ check (whether or not the flip vector with $\theta_i = 1$ is in the Rashomon set; note that we keep track of the current value of $\frac{\theta \cdot W_N}{N}$ throughout for computational efficiency) and an $O(1)$ resampling of θ_i from either Bernoulli(0.5) or Bernoulli(0). Since the number of iterations is a fixed constant, this means that the overall runtime of the algorithm is $O(N)$.

C Proofs of Theorems

In this section, we formally derive the theoretical results in the main paper, Sections 5 and 6. Note that the order in which we derive these results is different than the order in which they are presented in the main paper, as many of the results build on each other.

Let $\langle d_1, d_2, \ldots \rangle$ denote an infinite sequence of data records drawn i.i.d. from distribution D, and let D_N denote the subsequence $\langle d_1, d_2, \ldots, d_N \rangle$. Assume that each $d_i = (x_i, y_i)$ where x_i represents a set of predictor variables and y_i is a binary outcome variable. Let p_i denote the Bayes-optimal probability, $p_i = \Pr(y = 1 \mid x = x_i)$, and let w_i denote the corresponding weight, $w_i = |2p_i - 1|$. We define the vectors $P_N = \langle p_1, p_2, \ldots, p_N \rangle$, and $W_N = \langle w_1, w_2, \ldots, w_N \rangle$. Moreover, let P and W be the distributions of Bayes-optimal probabilities and weights respectively, for data records drawn i.i.d. from D.

Let $R_N(D_N, \epsilon)$ denote the largest possible Rashomon set of models for data records $\langle d_1, \ldots, d_N \rangle$. Since R_N can be computed using only the weights W_N, we can also write $R_N(P_N, \epsilon)$, $R_N(W_N, \epsilon)$, or simply $R_N(\epsilon)$ when the context (specifically, the weight vector W_N) is clear. Each distinct model in $R_N(\epsilon)$ represents a different binary classification of the data records $\langle d_1, \ldots, d_N \rangle$, $\langle f(x_1), \ldots, f(x_N) \rangle$, where $f(x_i) \in \{0, 1\}$ for all $i \in \{1, \ldots, N\}$, and thus there are at most 2^N models in $R_N(\epsilon)$. Note that the classifier can be probabilistic, i.e., two data records with identical x_i could have different $f(x_i)$ values. The *Bayes-optimal classifier* (the classifier with the lowest expected 0/1 loss, or equivalently, the highest expected accuracy) is a deterministic function of the Bayes-optimal probabilities p_i: $f_{\text{opt}}(x_i) = 1$ if $p_i > 0.5$, and $f_{\text{opt}}(x_i) = 0$ otherwise. We represent each model in $R_N(\epsilon)$ by a binary *flip vector* $\theta \in \{0,1\}^N$, where $\theta_i = 1$ if $f(x_i) \neq f_{\text{opt}}(x_i)$, and $\theta_i = 0$ if $f(x_i) = f_{\text{opt}}(x_i)$.

Definition C.1 (Accuracy of a model defined by a flip vector θ). Let $D_N = \langle d_1, d_2, \ldots, d_N \rangle$ be data records drawn i.i.d. from distribution D with corresponding Bayes-optimal probabilities $P_N = \langle p_1, p_2, \ldots, p_N \rangle$ and corresponding weights $W_N = \langle w_1, w_2, \ldots, w_N \rangle$, where $w_i = |2p_i - 1|$. The accuracy of a model with flip vector θ is

$$\begin{aligned}
acc(\theta) &= \frac{1}{N} \sum_{i=1\ldots N} (p_i f(x_i) + (1 - p_i)(1 - f(x_i))) \\
&= acc(\theta_0) + \frac{1}{N} \sum_{i=1\ldots N} \theta_i \left(((1 - p_i) - p_i) \mathbf{1}\{p_i > 0.5\}) \right. \\
&\qquad \left. + (p_i - (1 - p_i)) \mathbf{1}\{p_i \leq 0.5\}) \right) \\
&= acc(\theta_0) - \frac{1}{N} \sum_{i=1\ldots N} \theta_i |2p_i - 1| \\
&= acc(\theta_0) - \frac{\theta \cdot W_N}{N},
\end{aligned}$$

where $acc(\theta_0)$ is the accuracy of the Bayes-optimal classifier (and corresponding flip vector θ_0 consisting of all zeros).

Definition C.2 (Rashomon set). The Rashomon set of models $R_N(\epsilon)$ for error tolerance ϵ defined over data records $\langle d_1, \ldots, d_N \rangle$ is defined as the set of all models with corresponding flip vectors $\theta \in \{0,1\}^N$ such that $acc(\theta) \geq acc(\theta_0) - \epsilon$. Therefore, from Definition C.1 above, $R_N(\epsilon) = \{\theta \in \{0,1\}^N : \frac{\theta \cdot W_N}{N} \leq \epsilon\}$.

Definition C.3 (Flip probability). Let $R_N(\epsilon)$ denote the Rashomon set of models for error tolerance ϵ defined over data records $\langle d_1, \ldots, d_N \rangle$. For a given data record d_i, $i \in \{1, \ldots, N\}$, the *flip probability* $q_{N,i}$ is defined as the proportion of models in the Rashomon set for which the model prediction $f(x_i)$ differs from the Bayes-optimal prediction $f_{\text{opt}}(x_i) = \mathbf{1}\{p_i > 0.5\}$, or equivalently, the proportion of flip vectors for which $\theta_i = 1$:

$$q_{N,i} = \frac{|\theta \in R_N(\epsilon) : \theta_i = 1|}{|R_N(\epsilon)|}.$$

Lemma C.4 (Relationship between flip probability, weight, and Rashomon set size).

$$q_{N,i} = \frac{\left|R_{N,-i}\left(\frac{N\epsilon - w_i}{N-1}\right)\right|}{\left|R_{N,-i}\left(\frac{N\epsilon - w_i}{N-1}\right)\right| + \left|R_{N,-i}\left(\frac{N\epsilon}{N-1}\right)\right|},$$

where $R_{N,-i}(\epsilon)$ is the Rashomon set of models for error tolerance ϵ defined over the $N-1$ data records $\langle d_1, \ldots, d_{i-1}, d_{i+1}, \ldots, d_N \rangle$.

Proof. We can rewrite the criterion for membership in the Rashomon set, $\frac{\theta \cdot W_N}{N} \leq \epsilon$, as $\theta_{-i} \cdot W_{N,-i} + \theta_i w_i \leq N\epsilon$, where θ_{-i} and $W_{N,-i}$ are the flip vector omitting element i and the weight vector omitting element i respectively. The numerator of the above expression, and the first term of the denominator, represent the flip vectors θ for which $\theta_i = 1$. To satisfy $\frac{\theta \cdot W_N}{N} \leq \epsilon$ for these flip vectors, we must have $\theta_{-i} \cdot W_{N,-i} \leq N\epsilon - w_i$, or $\frac{\theta_{-i} \cdot W_{N,-i}}{N-1} \leq \frac{N\epsilon - w_i}{N-1}$. The second term of the denominator represents the flip vectors θ for which $\theta_i = 0$. To satisfy $\frac{\theta \cdot W_N}{N} \leq \epsilon$ for these flip vectors, we must have $\theta_{-i} \cdot W_{N,-i} \leq N\epsilon$, or $\frac{\theta_{-i} \cdot W_{N,-i}}{N-1} \leq \frac{N\epsilon}{N-1}$. □

Lemma C.5 (Asymptotic Pairwise Independence of Flip Probabilities). *Let $R_N(\epsilon)$ denote the Rashomon set of models for error tolerance ϵ defined over data records $\langle d_1, \ldots, d_N \rangle$. For any $i, j \in \{1, \ldots, N\}$, $i \neq j$, as $N \to \infty$, the flip probability q_j becomes independent of θ_i. Specifically:*

$$\lim_{N\to\infty} (q_{N,j} \mid \theta_i = 1) = \lim_{N\to\infty} (q_{N,j} \mid \theta_i = 0).$$

Proof. We consider two cases:

Case 1: $w_j = 0$.

When $w_j = 0$, flipping element j has no impact on total error. Therefore $q_{N,j} = \frac{1}{2}$ regardless of θ_i, so $\lim_{N\to\infty}(q_{N,j} | \theta_i = 1) = \lim_{N\to\infty}(q_{N,j} \mid \theta_i = 0) = \frac{1}{2}$.

Case 2: $w_j \neq 0$.

We compute the asymptotic odds ratio $\lim_{N\to\infty} \frac{q_{N,j}}{1-q_{N,j}}$ for both $\theta_i = 0$ and $\theta_i = 1$, and show that these two quantities are equal.

(1) For $\theta_i = 0$, using identical logic to Lemma C.4 above, the odds ratio is:

$$\frac{q_{N.j}}{1 - q_{N,j}} = \frac{\left|R_{N,-i,-j}\left(\frac{N\epsilon - w_j}{N-2}\right)\right|}{\left|R_{N,-i,-j}\left(\frac{N\epsilon}{N-2}\right)\right|},$$

where $\left|R_{N,-i,-j}(\epsilon)\right|$ is the size of the Rashomon set with error tolerance ϵ over the $N-2$ elements of $\langle d_1, \ldots, d_N \rangle$ excluding d_i and d_j.

Next, we define $\log B(\epsilon) = \lim_{N\to\infty} \frac{\log |R_N(\epsilon)|}{N}$, and note that, since $R_N(\epsilon)$ has minimum size 1 (for $\epsilon = 0$) and maximum size 2^N (for large ϵ), $B(\epsilon) \in [1, 2]$ for all $0 \leq \epsilon \leq 1$. We can also write $\log B(\epsilon) = \lim_{N\to\infty} \frac{\log |R_{N,-i,-j}(\epsilon)|}{N-2}$. Taking the logarithm of the expression above and letting $N \to \infty$, we obtain:

$$\begin{aligned}
&\lim_{N\to\infty} \log\left(\frac{q_{N.j}}{1 - q_{N,j}}\right) \\
&= \lim_{N\to\infty} \left(\log\left|R_{N,-i,-j}\left(\frac{N\epsilon - w_j}{N-2}\right)\right| - \log\left|R_{N,-i,-j}\left(\frac{N\epsilon}{N-2}\right)\right|\right) \\
&= \lim_{N\to\infty} (N-2)\left(\log B\left(\frac{N\epsilon - w_j}{N-2}\right) - \log B\left(\frac{N\epsilon}{N-2}\right)\right).
\end{aligned}$$

By the definition of the derivative of $\log B$, and noting that $-\frac{w_j}{N-2} \to 0$ as $N \to \infty$, we can write:

$$\begin{aligned}
\lim_{N\to\infty} \left(\log B\left(\frac{N\epsilon - w_j}{N-2}\right) - \log B\left(\frac{N\epsilon}{N-2}\right)\right) &= \lim_{N\to\infty} \\
-\frac{w_j}{N-2}\left(\frac{d \log B}{d\epsilon}\right)\Bigg|_{\frac{N\epsilon}{N-2}},&
\end{aligned}$$

and thus,

$$\begin{aligned}
\lim_{N\to\infty} \log\left(\frac{q_{N.j}}{1 - q_{N,j}}\right) &= \lim_{N\to\infty} -w_j \left(\frac{d \log B}{d\epsilon}\right)\Bigg|_{\frac{N\epsilon}{N-2}} \\
&= -w_j \left(\frac{d \log B}{d\epsilon}\right)\Bigg|_{\epsilon}.
\end{aligned}$$

(2) For $\theta_i = 1$, using identical logic to Lemma C.4 above, the odds ratio is:

$$\frac{q_{N.j}}{1 - q_{N,j}} = \frac{\left|R_{N,-i,-j}\left(\frac{N\epsilon - w_i - w_j}{N-2}\right)\right|}{\left|R_{N,-i,-j}\left(\frac{N\epsilon - w_i}{N-2}\right)\right|}.$$

Taking the logarithm and letting $N \to \infty$, we obtain:

$$\begin{aligned}
\lim_{N\to\infty} \log\left(\frac{q_{N.j}}{1 - q_{N,j}}\right) &= \lim_{N\to\infty} \left(\log\left|R_{N,-i,-j}\left(\frac{N\epsilon - w_i - w_j}{N-2}\right)\right|\right. \\
&\quad \left. - \log\left|R_{N,-i,-j}\left(\frac{N\epsilon - w_i}{N-2}\right)\right|\right) \\
&= \lim_{N\to\infty} (N-2)\left(\log B\left(\frac{N\epsilon - w_i - w_j}{N-2}\right)\right. \\
&\quad \left. - \log B\left(\frac{N\epsilon - w_i}{N-2}\right)\right).
\end{aligned}$$

By the definition of the derivative of $\log B$, and noting that $-\frac{w_j}{N-2} \to 0$ as $N \to \infty$, we can write:

$$\lim_{N\to\infty} \left(\log B\left(\frac{N\epsilon - w_i - w_j}{N-2}\right) - \log B\left(\frac{N\epsilon - w_i}{N-2}\right)\right) = \lim_{N\to\infty} -\frac{w_j}{N-2}\left(\frac{d\log B}{d\epsilon}\right)\Big|_{\frac{N\epsilon - w_i}{N-2}},$$

and thus,

$$\lim_{N\to\infty} \log\left(\frac{q_{N.j}}{1-q_{N,j}}\right) = \lim_{N\to\infty} -w_j\left(\frac{d\log B}{d\epsilon}\right)\Big|_{\frac{N\epsilon - w_i}{N-2}} = -w_j\left(\frac{d\log B}{d\epsilon}\right)\Big|_{\epsilon}.$$

Thus for both $\theta_i = 0$ and $\theta_i = 1$, $\lim_{N\to\infty} \log\left(\frac{q_{N.j}}{1-q_{N,j}}\right) = -w_j\left(\frac{d\log B}{d\epsilon}\right)\Big|_{\epsilon}$, and the proof is completed. □

Lemma C.6 (Functional form of flip probabilities). *Let $D_N = \langle d_1, d_2, \ldots, d_N\rangle$ be data records drawn i.i.d. from distribution D with corresponding Bayes-optimal probabilities $P_N = \langle p_1, p_2, \ldots, p_N\rangle$ and corresponding weights $W_N = \langle w_1, w_2, \ldots, w_N\rangle$, where $w_i = |2p_i - 1|$. Assume $w_i \sim W$ where distribution W has pdf $f(w) > 0$ for $w \in [0,1]$. Let $R_N(\epsilon)$ denote the Rashomon set of models for error tolerance ϵ defined over data records $\langle d_1, \ldots, d_N\rangle$. Consider the flip probabilities $q_{N,i}$ corresponding to Rashomon set $R_N(\epsilon)$, and define $q_i = \lim_{N\to\infty} q_{N,i}$. Then we can write:*

$$q_i = \frac{1}{1+\exp(Cw_i)},$$

where C is constant for a given ϵ and a given weight distribution W.

Proof. Consider any three data records d_i, d_j, and d_k such that $w_i + w_j = w_k$. We know that such triples exist as $N \to \infty$ because of the continuity and positivity of the weight distribution. Next we consider any pair of flip vectors $\theta, \theta' \in \{0,1\}^N$ such that $(\theta_i, \theta_j, \theta_k) = (1,1,0)$, $(\theta'_i, \theta'_j, \theta'_k) = (0,0,1)$, and $\theta_l = \theta'_l$ for all $l \in \{1,2,\ldots,N\}$, $l \notin \{i,j,k\}$. The total weight is the same for both vectors: $\theta \cdot W_N = w_i + w_j + W_{rest} = w_k + W_{rest} = \theta' \cdot W_N$, where $W_{rest} = \sum_{l\in\{1,2,\ldots,N\}, l\notin\{i,j,k\}} \theta_l w_l$, and thus either both flip vectors $\theta, \theta' \in R_N(\epsilon)$, or both flip vectors $\theta, \theta' \notin R_N(\epsilon)$. This means that, for flip vectors $\theta \in R_N(\epsilon)$, the probability that $(\theta_i, \theta_j, \theta_k) = (1,1,0)$ and the probability that $(\theta_i, \theta_j, \theta_k) = (0,0,1)$ are equal. For $N \to \infty$, pairwise independence (Lemma C.5) allows us to write these probabilities as $q_i(q_j)(1-q_k)$ and $(1-q_i)(1-q_j)q_k$ respectively, and thus $q_i(q_j)(1-q_k) = (1-q_i)(1-q_j)q_k$. We can then rearrange terms and take the logarithm:

$$\log\left(\frac{q_i}{1-q_i}\right) + \log\left(\frac{q_j}{1-q_j}\right) = \log\left(\frac{q_k}{1-q_k}\right).$$

This establishes that for any data elements d_i, d_j, and d_k with $w_i + w_j = w_k$, $h(w_i) + h(w_j) = h(w_k)$, where the function $h(w) = \log\left(\frac{q(w)}{1-q(w)}\right)$ The equation $h(w_i)+h(w_j) = h(w_i+w_j)$ is Cauchy's functional equation for additive functions.

Moreover, the function $h(w)$ is monotone for $w \in (0,1)$. To see this, we note that flip probability $q(w)$ is monotonically decreasing with w, since for every possible configuration θ_{-i}, higher weight w_i monotonically decreases (i.e., does not increase) the probability that $\theta_i = 1$ is in the Rashomon set, and does not change the probability that $\theta_i = 0$ is in the Rashomon set. Moreover, $\log\left(\frac{q}{1-q}\right)$ is increasing with q, so $h(w) = \log\left(\frac{q(w)}{1-q(w)}\right)$ is monotone.

Monotonicity of $h(w)$ is a sufficient condition for ensuring that the Cauchy functional equation does not have pathological (non-linear) solutions, and thus the only solutions are linear functions $h(w) = -Cw$, where C is a constant (for a given W and ϵ). Therefore, the log-odds of the flip probability is proportional to the weight:

$$\log\left(\frac{q_i}{1-q_i}\right) = -Cw_i.$$

Finally, the flip probability q_i can be expressed as:

$$q_i = \frac{1}{1+\exp(Cw_i)}.$$

which completes the proof. (Note that the value of C, as a function of ϵ, will be obtained in Corollary C.10 below.) □

Lemma C.7 (Asymptotic size of Rashomon set as a function of C). *Let $D_N = \langle d_1, d_2, \ldots, d_N\rangle$ be data records drawn i.i.d. from distribution D with corresponding Bayes-optimal probabilities $P_N = \langle p_1, p_2, \ldots, p_N\rangle$ and corresponding weights $W_N = \langle w_1, w_2, \ldots, w_N\rangle$, where $w_i = |2p_i - 1|$. Assume $w_i \sim W$ where distribution W has pdf $f(w) > 0$ for $w \in [0,1]$. Let $R_N(\epsilon)$ denote the Rashomon set of models for error tolerance ϵ defined over data records $\langle d_1, \ldots, d_N\rangle$. Let $C(\epsilon)$ denote the constant in the asymptotic flip probability, $q_i = \lim_{N\to\infty} q_{N,i} = \frac{1}{1+\exp(C(\epsilon)w_i)}$, for error tolerance ϵ. Then:*

$$\lim_{N\to\infty} \frac{\log|R_N(\epsilon)|}{N} = \log B(\epsilon),$$

where

$$B(\epsilon) = \exp\left(\int_0^{\epsilon} C(x)dx\right).$$

Proof. From Lemma C.4, we know

$$q_i = \lim_{N\to\infty} \frac{\left|R_{N,-i}\left(\frac{N\epsilon - w_i}{N-1}\right)\right|}{\left|R_{N,-i}\left(\frac{N\epsilon - w_i}{N-1}\right)\right| + \left|R_{N,-i}\left(\frac{N\epsilon}{N-1}\right)\right|},$$

where $R_{N,-i}(\epsilon)$ is the Rashomon set of models for error tolerance ϵ defined over the $N-1$ data records $\langle d_1, \ldots, d_{i-1}, d_{i+1}, \ldots, d_N\rangle$. And from Lemma C.6, we know that $q_i = \frac{1}{1+\exp(C(\epsilon)w_i)}$. Setting these quantities equal to each other, we have:

$$\lim_{N\to\infty} \frac{\left|R_{N,-i}\left(\frac{N\epsilon - w_i}{N-1}\right)\right|}{\left|R_{N,-i}\left(\frac{N\epsilon - w_i}{N-1}\right)\right| + \left|R_{N,-i}\left(\frac{N\epsilon}{N-1}\right)\right|} = \frac{1}{1+\exp(C(\epsilon)w_i)}.$$

Inverting both sides:

$$\lim_{N\to\infty}\left(1 + \frac{\left|R_{N,-i}\left(\frac{N\epsilon}{N-1}\right)\right|}{\left|R_{N,-i}\left(\frac{N\epsilon - w_i}{N-1}\right)\right|}\right) = 1+\exp(C(\epsilon)w_i).$$

Subtracting 1 and taking the logarithm of both sides:

$$\lim_{N\to\infty}\left(\log\left|R_{N,-i}\left(\frac{N\epsilon}{N-1}\right)\right| - \log\left|R_{N,-i}\left(\frac{N\epsilon - w_i}{N-1}\right)\right|\right) = C(\epsilon)w_i.$$

Dividing both sides by w_i:

$$\lim_{N\to\infty} \frac{\log\left|R_{N,-i}\left(\frac{N\epsilon}{N-1}\right)\right| - \log\left|R_{N,-i}\left(\frac{N\epsilon - w_i}{N-1}\right)\right|}{(N-1)\frac{w_i}{N-1}} = C(\epsilon).$$

By the definition of derivative, noting that $\frac{w_i}{N-1} \to 0$ as $N \to \infty$:

$$\lim_{N\to\infty}\left(\frac{1}{N-1}\right)\frac{d\log\left|R_{N,-i}(\epsilon)\right|}{d\epsilon}\bigg|_{\frac{N\epsilon}{N-1}} = C(\epsilon).$$

Equivalently, we can write:

$$\lim_{N\to\infty}\left(\frac{1}{N}\right)\frac{d\log|R_N(\epsilon)|}{d\epsilon}\bigg|_{\epsilon} = C(\epsilon).$$

Integrating both sides with respect to ϵ:

$$\lim_{N\to\infty}\left(\frac{1}{N}\right)\log|R_N(\epsilon)| = \int_0^{\epsilon} C(x)\,dx + \text{constant}$$

We know that the constant is 0 since, for $\epsilon = 0$, we have:

$$\lim_{N\to\infty}|R_N(\epsilon)| = 1.$$

Thus we have:

$$\lim_{N\to\infty}\frac{\log|R_N(\epsilon)|}{N} = \int_0^{\epsilon} C(x)\,dx.$$

Finally, defining $B(\epsilon) = \exp\left(\int_0^{\epsilon} C(x)dx\right)$, we can write:

$$\lim_{N\to\infty}\frac{\log|R_N(\epsilon)|}{N} = \log B(\epsilon).$$

□

Definition C.8 (Average accuracy of models in the Rashomon set). Let $D_N = \langle d_1, d_2, \ldots, d_N\rangle$ be data records drawn i.i.d. from distribution D with corresponding Bayes-optimal probabilities $P_N = \langle p_1, p_2, \ldots, p_N\rangle$ and corresponding weights $W_N = \langle w_1, w_2, \ldots, w_N\rangle$, where $w_i = |2p_i - 1|$. Let $R_N(\epsilon)$ denote the Rashomon set of models for error tolerance ϵ defined over data records $\langle d_1, \ldots, d_N\rangle$, and consider the corresponding flip probabilities $q_{N,i}$. Then the average accuracy of models $\theta \in R_N(\epsilon)$ can be written as:

$$\begin{aligned}
\overline{acc}(R_N(\epsilon)) &= \frac{1}{|R_N(\epsilon)|}\sum_{\theta\in R_N(\epsilon)} acc(\theta)\\
&= acc_N(\theta_0) - \frac{1}{|R_N(\epsilon)|}\sum_{\theta\in R_N(\epsilon)}\frac{\theta\cdot W_N}{N}\\
&= acc_N(\theta_0) - \frac{1}{|R_N(\epsilon)|}\sum_{\theta\in R_N(\epsilon)}\sum_{i=1\ldots N}\frac{\theta_i w_i}{N}\\
&= acc_N(\theta_0) - \frac{1}{N}\sum_{i=1\ldots N} w_i \frac{1}{|R_N(\epsilon)|}\sum_{\theta\in R_N(\epsilon)}\theta_i\\
&= acc_N(\theta_0) - \frac{1}{N}\sum_{i=1\ldots N} w_i q_{N,i},
\end{aligned}$$

where $acc_N(\theta_0)$ is the Bayes-optimal accuracy for data elements $\langle d_1, \ldots, d_N\rangle$.

Further, assume $w_i \sim W$ where distribution W has pdf $f(w) > 0$ for $w \in [0,1]$. Then we can write the asymptotic average accuracy as:

$$\lim_{N\to\infty}\overline{acc}(R_N(\epsilon)) = acc(\theta_0) - \int_0^1 w\,q(w)\,f(w)\,dw,$$

where $acc(\theta_0)$ is the asymptotic Bayes-optimal accuracy, $acc(\theta_0) = \lim_{N\to\infty} acc_N(\theta_0)$, and $q(w)$ denotes the asymptotic flip probability $q_i = \lim_{N\to\infty} q_{N,i}$ corresponding to an element i with weight w.

Theorem C.9 (Asymptotic use of the entire error tolerance). *Let $D_N = \langle d_1, d_2, \ldots, d_N\rangle$ be data records drawn i.i.d. from distribution D with corresponding Bayes-optimal probabilities $P_N = \langle p_1, p_2, \ldots, p_N\rangle$ and corresponding weights $W_N = \langle w_1, w_2, \ldots, w_N\rangle$, where $w_i = |2p_i - 1|$. Assume $w_i \sim W$ where distribution W has pdf $f(w) > 0$ for $w \in [0,1]$. Let $R_N(\epsilon)$ denote the Rashomon set of models for error tolerance ϵ defined over data records $\langle d_1, \ldots, d_N\rangle$, and assume that ϵ is less than half of the average weight, i.e., $\int_0^1 w\,f(w)\,dw > 2\epsilon$. Let $q(w)$ denote the asymptotic flip probability $q_i = \lim_{N\to\infty} q_{N,i}$ corresponding to an element i with weight w.*

Then as $N \to \infty$, the average error tolerance used by models in the Rashomon set converges to ϵ:

$$acc(\theta_0) - \lim_{N\to\infty}\overline{acc}(R_N(\epsilon)) = \int_0^1 w\,q(w)\,f(w)\,dw = \epsilon.$$

Proof. As every flip vector θ in the Rashomon set $R_N(\epsilon)$ must satisfy $\sum_{i=1}^N \frac{\theta_i w_i}{N} \le \epsilon$, we have:

$$\begin{aligned}
\frac{1}{|R_N(\epsilon)|}\sum_{\theta\in R_N(\epsilon)}\sum_{i=1\ldots N}\frac{\theta_i w_i}{N} &= \frac{1}{N}\sum_{i=1\ldots N} w_i \frac{1}{|R_N(\epsilon)|}\sum_{\theta\in R_N(\epsilon)}\theta_i\\
&= \frac{1}{N}\sum_{i=1\ldots N} w_i q_{N,i}\\
&\le \epsilon.
\end{aligned}$$

As $N \to \infty$, we can replace the summation with the integral:

$$\int_0^1 w\, q(w)\, f(w)\, dw \le \epsilon.$$

Since $q(w) = \frac{1}{1+\exp(C(\epsilon)\, w)}$ from Lemma C.6, we have:

$$\int_0^1 \frac{w}{1+\exp(C(\epsilon)\, w)}\, f(w)\, dw \le \epsilon.$$

Next, the assumption $\int_0^1 w\, f(w)\, dw > 2\epsilon$ ensures that $C(\epsilon) > 0$. To see this, we first note that the expected error $\int_0^1 w\, q(w)\, f(w)\, dw$ is a monotonically decreasing function of $C(\epsilon)$. Then if we assume $C(\epsilon) \le 0$, the flip probability becomes $q(w) \ge \frac{1}{2}$, leading to an average error of at least $\frac{1}{2}\int_0^1 w\, f(w)\, dw > \epsilon$, which contradicts the fact above that the expected error can be at most ϵ.

Now consider any $\delta > 0$ with $0 < \delta < \epsilon$. We apply Lemma C.7 to compute the asymptotic sizes of the Rashomon sets at ϵ and $\epsilon - \delta$ respectively:

$$\lim_{N\to\infty} \frac{\log |R_N(\epsilon)|}{N} = \int_0^{\epsilon} C(x)\, dx.$$

and

$$\lim_{N\to\infty} \frac{\log |R_N(\epsilon-\delta)|}{N} = \int_0^{\epsilon-\delta} C(x)\, dx.$$

Thus we can write

$$\lim_{N\to\infty} \frac{1}{N} \log \frac{|R_N(\epsilon)|}{|R_N(\epsilon-\delta)|} = \left(\int_0^{\epsilon} C(x)\, dx - \int_0^{\epsilon-\delta} C(x)\, dx\right) = \int_{\epsilon-\delta}^{\epsilon} C(x)\, dx.$$

Since $C(x) > 0$ for $x \le \epsilon$, we know that $\int_{\epsilon-\delta}^{\epsilon} C(x)\, dx > 0$, and $\exp\left(\int_{\epsilon-\delta}^{\epsilon} C(x)\, dx\right) > 1$. Therefore we have:

$$\lim_{N\to\infty} \frac{|R_N(\epsilon)|}{|R_N(\epsilon-\delta)|} = \lim_{N\to\infty} \left(\exp\left(\int_{\epsilon-\delta}^{\epsilon} C(x)\, dx\right)\right)^N \to \infty.$$

This implies that for large N, the number of flip vectors (models) with total error in the interval $[\epsilon-\delta, \epsilon]$ dominates the Rashomon set. The proportion of flip vectors with error less than $\epsilon - \delta$ becomes negligible. Since almost all flip vectors in $R_N(\epsilon)$ have errors between $\epsilon - \delta$ and ϵ, the asymptotic expected error $\int_0^1 w\, q(w)\, f(w)\, dw$ is greater than or equal to $\epsilon - \delta$. Because $\delta > 0$ is arbitrary and can be made as small as desired, we have:

$$\int_0^1 w\, q(w)\, f(w)\, dw \ge \epsilon - \delta \quad \text{for all } \delta > 0.$$

Combining with the initial inequality, we have:

$$\epsilon - \delta \le \int_0^1 w\, q(w)\, f(w)\, dw \le \epsilon \quad \text{for all } \delta > 0,$$

and thus

$$\int_0^1 w\, q(w)\, f(w)\, dw = \epsilon,$$

which completes the proof. □

Corollary C.10 (Value of C). *Let $D_N = \langle d_1, d_2, \ldots, d_N\rangle$ be data records drawn i.i.d. from distribution D with corresponding Bayes-optimal probabilities $P_N = \langle p_1, p_2, \ldots, p_N\rangle$ and corresponding weights $W_N = \langle w_1, w_2, \ldots, w_N\rangle$, where $w_i = |2p_i - 1|$. Assume $w_i \sim W$ where distribution W has pdf $f(w) > 0$ for $w \in [0,1]$. Let $R_N(\epsilon)$ denote the Rashomon set of models for error tolerance ϵ defined over data records $\langle d_1, \ldots, d_N\rangle$, and assume that ϵ is less than half of the average weight, i.e., $\int_0^1 w\, f(w)\, dw > 2\epsilon$.*

Then the value of $C(\epsilon)$ in Lemmas C.6 and C.7 can be written as $C(\epsilon) = g^{-1}(\epsilon)$, where $g(C) = \int_0^1 \frac{wf(w)}{1+\exp(Cw)}\, dw$.

Proof. From Theorem C.9, we have

$$\int_0^1 w\, q(w)\, f(w)\, dw = \epsilon,$$

and from Lemma C.6, we have

$$q(w) = \frac{1}{1+\exp(C(\epsilon)\, w)}.$$

Putting these together with the function g defined above, we have

$$g(C(\epsilon)) = \int_0^1 \frac{wf(w)}{1+\exp(C(\epsilon)\, w)}\, dw = \epsilon,$$

or equivalently, $C(\epsilon) = g^{-1}(\epsilon)$, and the proof is completed. □

Corollary C.11 (Value of C for uniformly distributed weights). *Let $D_N = \langle d_1, d_2, \ldots, d_N\rangle$ be data records drawn i.i.d. from distribution D with corresponding Bayes-optimal probabilities $P_N = \langle p_1, p_2, \ldots, p_N\rangle$ and corresponding weights $W_N = \langle w_1, w_2, \ldots, w_N\rangle$, where $w_i = |2p_i - 1|$. Assume $w_i \sim$ Uniform$[0,1]$. Let $R_N(\epsilon)$ denote the Rashomon set of models for error tolerance ϵ defined over data records $\langle d_1, \ldots, d_N\rangle$, and assume that ϵ is less than half of the average weight, i.e., $\int_0^1 w\, f(w)\, dw > 2\epsilon$.*

Then the value of $C(\epsilon)$ in Lemmas C.6 and C.7 can be written as $C(\epsilon) = g^{-1}(\epsilon)$, where:

$$g(C) = \int_0^1 \frac{w}{1+\exp(Cw)}\, dw = \frac{12\, Li_2(-e^{-C}) - 12C\log(e^{-C}+1) + \pi^2}{12C^2}.$$

Moreover, $C(\epsilon) < \frac{\pi}{\sqrt{12\epsilon}}$, and $C(\epsilon) \approx \frac{\pi}{\sqrt{12\epsilon}}$ for small ϵ.

Proof. Given $w_i \sim$ Uniform$[0,1]$, we know that the pdf $f(w) = 1$ for $w \in [0,1]$. Plugging this into the result of Corollary C.10, we obtain

$$\int_0^1 \frac{w}{1+\exp(Cw)}\, dw = \epsilon.$$

The integral can be computed as:

$$\int_0^1 \frac{w}{1+\exp(Cw)}\,dw = \frac{12\,\mathrm{Li}_2(-e^{-C}) - 12C\log(e^{-C}+1) + \pi^2}{12C^2},$$

where Li_2 is the dilogarithmic (Spence's) function. Since the first two terms of the rhs are negative for all C, $\frac{\pi^2}{12C^2} > \epsilon$, and thus $C < \frac{\pi}{\sqrt{12\epsilon}}$. As $\epsilon \to 0$, C becomes large, and the first two terms of the rhs go to 0 from below. Thus we have $\frac{\pi^2}{12C^2} \approx \epsilon$, and $C \approx \frac{\pi}{\sqrt{12\epsilon}}$. □

Theorem C.12 (Asymptotic flip probabilities). *Let $D_N = \langle d_1, d_2, \ldots, d_N\rangle$ be data records drawn i.i.d. from distribution D with corresponding Bayes-optimal probabilities $P_N = \langle p_1, p_2, \ldots, p_N\rangle$ and corresponding weights $W_N = \langle w_1, w_2, \ldots, w_N\rangle$, where $w_i = |2p_i - 1|$. Assume $w_i \sim W$ where distribution W has pdf $f(w) > 0$ for $w \in [0,1]$. Let $R_N(\epsilon)$ denote the Rashomon set of models for error tolerance ϵ defined over data records $\langle d_1, \ldots, d_N\rangle$, and assume that ϵ is less than half of the average weight, i.e., $\int_0^1 w\,f(w)\,dw > 2\epsilon$. Consider the flip probabilities $q_{N,i}$ corresponding to Rashomon set $R_N(\epsilon)$, and define $q_i = \lim_{N\to\infty} q_{N,i}$. Then*

$$q_i = \frac{1}{1+\exp(C(\epsilon)\,w_i)},$$

where $C(\epsilon) = g^{-1}(\epsilon)$ and $g(C) = \int_0^1 \frac{wf(w)}{1+\exp(Cw)}\,dw$.

Proof. The statement follows immediately from Lemma C.6, which gives the functional form of q_i, and Corollary C.10, which gives the expression for C. □

Remark. As a consequence of this theorem, for a Rashomon set $R_N(\epsilon)$ with N large, we can obtain the flip probabilities for each individual in two steps: (1) calculate the value of C for the given weight distribution W and error tolerance ϵ; and (2) compute $q_i = \frac{1}{1+\exp(Cw_i)}$ for each individual i. To calculate C if the pdf $f(w)$ of the weight distribution W is known, we solve the integral equation $g(C) = \int_0^1 \frac{wf(w)}{1+\exp(Cw)}\,dw = \epsilon$. Alternatively, given a finite dataset of size N, we estimate the true weight distribution W using the empirical distribution W_N, and thus solve the equation $\frac{1}{N}\sum_{i=1\ldots N} \frac{w_i}{1+\exp(Cw_i)} = \epsilon$. In either case, we note that the lhs decreases monotonically with C, allowing an efficient solution by binary search.

Theorem C.13 (Asymptotic size of Rashomon set). *Let $D_N = \langle d_1, d_2, \ldots, d_N\rangle$ be data records drawn i.i.d. from distribution D with corresponding Bayes-optimal probabilities $P_N = \langle p_1, p_2, \ldots, p_N\rangle$ and corresponding weights $W_N = \langle w_1, w_2, \ldots, w_N\rangle$, where $w_i = |2p_i - 1|$. Assume $w_i \sim W$ where distribution W has pdf $f(w) > 0$ for $w \in [0,1]$. Let $R_N(\epsilon)$ denote the Rashomon set of models for error tolerance ϵ defined over data records $\langle d_1, \ldots, d_N\rangle$, and assume that ϵ is less than half of the average weight, i.e., $\int_0^1 w\,f(w)\,dw > 2\epsilon$. Then*

$$\lim_{N\to\infty} \frac{\log|R_N(\epsilon)|}{N} = \log B(\epsilon),$$

where $B(\epsilon) = \exp\left(\int_0^\epsilon C(x)dx\right)$, $C(\epsilon) = g^{-1}(\epsilon)$, and $g(C) = \int_0^1 \frac{wf(w)}{1+\exp(Cw)}\,dw$.

Proof. The statement follows immediately from Lemma C.7, which gives the size of the Rashomon set in terms of C, and Corollary C.10, which gives the expression for C. □

Remark. To compute the exponential base $B(\epsilon)$, and therefore the Rashomon set size $|R_N(\epsilon)| = B(\epsilon)^N$, given a finite dataset of size N, we can calculate the value of $C(\epsilon)$ for a fine grid of ϵ values by solving the equation $\frac{1}{N}\sum_{i=1\ldots N} \frac{w_i}{1+\exp(Cw_i)} = \epsilon$. We then use numerical integration to estimate $B(\epsilon) = \exp\left(\int_0^\epsilon C(x)dx\right)$. Alternatively, for $N \to \infty$ with a known distribution of weights, $w_i \sim W$ with pdf $f(w)$, we instead solve the integral $\int_0^1 \frac{wf(w)}{1+\exp(Cw)}\,dw = \epsilon$ to obtain $C(\epsilon)$.

Corollary C.14 (Asymptotic size of Rashomon set for uniformly distributed weights). *Let $D_N = \langle d_1, d_2, \ldots, d_N\rangle$ be data records drawn i.i.d. from distribution D with corresponding Bayes-optimal probabilities $P_N = \langle p_1, p_2, \ldots, p_N\rangle$ and corresponding weights $W_N = \langle w_1, w_2, \ldots, w_N\rangle$, where $w_i = |2p_i - 1|$. Assume $w_i \sim$ Uniform$[0,1]$. Let $R_N(\epsilon)$ denote the Rashomon set of models for error tolerance ϵ defined over data records $\langle d_1, \ldots, d_N\rangle$, and assume that ϵ is less than half of the average weight, i.e., $\int_0^1 w\,f(w)\,dw > 2\epsilon$. Then*

$$\lim_{N\to\infty} \frac{\log|R_N(\epsilon)|}{N} = \log B(\epsilon),$$

where $B(\epsilon) = \exp\left(\int_0^\epsilon C(x)dx\right)$, $C(\epsilon) = g^{-1}(\epsilon)$, and $g(C) = \int_0^1 \frac{w}{1+\exp(Cw)}\,dw = \frac{12\,Li_2(-e^{-C}) - 12C\log(e^{-C}+1)+\pi^2}{12C^2}$.

Moreover, $B(\epsilon) < \exp\left(\pi\sqrt{\frac{\epsilon}{3}}\right)$, and $B(\epsilon) \approx \exp\left(\pi\sqrt{\frac{\epsilon}{3}}\right)$ for small ϵ.

Proof. The statement follows from Lemma C.7, which gives the size of the Rashomon set in terms of C, and Corollary C.11, which gives the exact and approximate (upper bound) expressions for C for uniform weights. Since $C(\epsilon) < \frac{\pi}{\sqrt{12\epsilon}}$, we know that $B(\epsilon) = \exp\left(\int_0^\epsilon C(x)dx\right) < \exp\left(\int_0^\epsilon \frac{\pi}{\sqrt{12x}}dx\right) = \exp\left(\pi\sqrt{\frac{\epsilon}{3}}\right)$. And since $C(\epsilon) \approx \frac{\pi}{\sqrt{12\epsilon}}$ for $\epsilon \to 0$, we know that $B(\epsilon) \approx \exp\left(\pi\sqrt{\frac{\epsilon}{3}}\right)$ for $\epsilon \to 0$. □

D Description of benchmark datasets

Throughout this paper, we present experimental results on three real-world datasets that are commonly used as benchmarks in the fair machine learning literature: German Credit ("German"), Adult, and Heritage Health ("Health").

We use a preprocessed version of the German Credit data [20] publicly available on Kaggle [14], which includes credit risk as an outcome variable. Numerical attributes in the dataset, namely *Age*, *Job*, *Credit Amount*, and *Duration*,

were discretized to a categorical attribute as follows: *Age* was discretized based on whether $Age \geq 25$ or otherwise; *Job* was discretized based on the number of jobs (no job, one job, or more than one job); *Credit Amount*, whose values range from 250 to 18,400, was discretized into five bins; and *Duration*, whose values range from 4 to 72, was discretized into five bins. Categorical and binary attributes were unchanged. Finally, attributes were 1-hot encoded.

We use a publicly available version of the Adult data [1], which includes income as an outcome variable. Numerical attributes in the dataset, namely *age*, *fnlwgt* (final weight), *education-num* (education level), *capital-gain*, *capital-loss*, and *hours-per-week*, were binarized using their median value as the threshold. Categorical and binary attributes were unchanged. Finally, attributes were 1-hot encoded.

We use a publicly available version of the Heritage Health data [25]. We use similar features as the winning team, Market Makers [27]. We generate the features using the SQL script in the Appendix of [27], which generates the majority of the variables in data set 1. Since the *ageMISS* feature corresponds to whether the age value is missing or not, all rows with $ageMISS = 1$ were removed, and the *ageMISS* feature was dropped. Additionally, the sensitive feature S was created with $S = 0$ when the age is between 0 and 59 ($age\_05 = 1$ or $age\_15 = 1$ or $age\_25 = 1$ or $age\_35 = 1$ or $age\_45 = 1$ or $age\_55 = 1$), and $S = 1$ otherwise. Numerical attributes were binarized using their median value as the threshold. Categorical and binary attributes were unchanged. Finally, attributes were 1-hot encoded.

In German Credit ($N = 1,000$), there are 690 men (labeled $gender = 1$ in the dataset) and 310 women ($gender = 0$). The outcome variable (high risk) is whether an individual is considered high-risk for a loan. Women ($gender = 0$) are the minority class (31.0% of the dataset) and are disadvantaged (35.2% likely to be considered high risk for a loan, vs. 27.7% for men).

In Adult ($N = 46,443$), there are 15,203 women (labeled $sex = 1$ in the dataset) and 31,240 men ($sex = 0$). The outcome variable (*income*) is whether a person has income over $50,000. Women ($sex = 1$) are the minority class (32.7% of the dataset) and are disadvantaged (11.2% likely to be predicted high income vs. 30.9% for men).

In Health ($N = 184,308$), there are 73,535 individuals over the age of 60 (labeled $S = 1$ in the dataset) and 110,773 other individuals ($S = 0$). The outcome variable (*DaysInHospital*) represents whether a person will spend any days in the hopsital that year. Individuals over the age of 60 ($S = 1$) are the minority class (39.9% of the dataset) and are disadvantaged ($DaysInHospital = 1$ 19.6% of the time, vs. 10.6% for others).

As noted in Section 4.3 above, we estimate the the Bayes-optimal probabilities p_i for all three datasets using 5-fold cross-validation, using two different approaches, logistic regression (main paper) and XGBoost (Appendix I). We report the cross-validated accuracy scores for each dataset using the approximate Bayes-optimal predictions $f_{\text{opt}}(x_i) = \mathbf{1}\{\hat{p}_i > 0.5\}$ and observed outcomes y_i. For logistic regression, accuracy was 73.7%, 84.4%, and 86.2% for German, Adult, and Health datasets respectively. For XGBoost, accuracy was 69.4%, 84.3%, and 88.8% for German, Adult, and Health datasets respectively.

E Experiments on generalization to previously unseen data

Given that our models (as characterized by flip vectors θ) are defined in terms of their labeling of the N training samples (as compared to the labels produced by the Bayes-optimal classifier), one might ask how flip vectors correspond to generalizable models that could be used to label previously unseen test data. We consider two natural approaches for generalization. First, for an arbitrary flip vector θ, we can consider its corresponding labels $\hat{y}_i = \mathbf{1}\{(p_i > 0.5 \text{ and } \theta_i = 0) \text{ or } (p_i \leq 0.5 \text{ and } \theta_i = 1)\}$ as training data, and define the corresponding 1-nearest neighbor classifier with ties in distance broken uniformly at random. This would imply that, if a given test sample has been seen one or more times in the training data, its predicted label is drawn from the same distribution as the training predictions, and if not, nearby points are used to assign the label. In either case, this approach potentially results in a randomized classifier. Alternatively, many flip vectors θ might be created by rules that generalize from training to test data. Given that the Bayes-optimal classifier is estimated from labeled data and that its probabilistic predictions can be used to make classification decisions for previously unseen examples, a rule which defines how a given classifier deviates from Bayes-optimal (e.g., by changing the classification threshold from 0.5 to a different value, or by randomizing labels as a function of the Bayes-optimal probability), will also generalize.

In particular, the algorithms we present for finding fairer models through optimal search and random sampling generalize since they effectively create rules for how to deviate from the Bayes-optimal predictor. For optimal search, separate prediction thresholds are created for each class. For sampling, the model disagrees with the Bayes-optimal prediction for each new data record with probability $q_i = \frac{1}{1+\exp(Cw_i)}$. For each value of the error tolerance ϵ, the prediction thresholds for optimization and the constant C for sampling can be learned from one (unlabeled) sample of the data and generalize to another.

To illustrate that our approach and results generalize in practice, we perform the following experiment. For six different trials, we split the Adult dataset into three partitions, using one to estimate the Bayes-optimal probabilities $\Pr(y = 1 \mid x)$, one to learn C values for random sampling and class-specific decision thresholds for PPR, FPR, and TPR optimization, and one to recreate PPR, FPR, and TPR disparity curves (as a function of ϵ) for our sampling method (green) and our optimal search methods (blue). Critically, note that we do not use training labels for the second and third partitions, only for learning the Bayes-optimal classifier. As shown in Figure 4 and Appendix I, Figure 17, for PPR and FPR, all six permutations of the three partitions produced disparity curves that were very close to each other and to the curves estimated from the entire dataset with no splitting: all curves' standard deviations and

RMSEs, averaged across epsilon values, were less than 0.008. The differences are larger for TPR, as expected for the small effective sample size.

Thus the above experiment demonstrates that the flip vectors θ learned by our optimal search and sampling approaches can generalize from one "training" partition of the data to a second "test" partition. This is not surprising given that the training and test partitions are both unlabeled, while labeled data is used only to estimate the Bayes-optimal model. The above experiment also demonstrates the robustness of our results to estimation of the Bayes-optimal model $\Pr(y = 1 \mid x)$. Model estimates across the three partitions of the data differed, yet we did not see substantial differences in disparity curves generated either by optimization or by sampling. Similarly, our conclusions remained consistent when using a different model class to estimate Bayes-optimal probabilities (Appendix I).

F Additional results for Section 4.3 (Intentional Fairness)

Figures 5 and 6 show the disparity in false positive rate (FPR) and true positive rate (TPR) respectively, as a function of the error tolerance ϵ. These figures compare the methods for optimizing FPR and TPR disparities over the Rashomon set $R_N(\epsilon)$ (Section 4.1) to uniform random sampling (Section 4.2) and sampling linear models (Section 4.3.1). We see that both sets of results are very similar to the results presented for PPR disparity in Figure 1.

G Additional results for Section 5.2 and 5.3 (Flip Probabilities)

As noted in Section 5.2, we can *exactly* (in the large-sample limit) and *efficiently* compute the *average* over the entire Rashomon set $R_N(\epsilon)$ of any metric (such as accuracy, PPR disparity, FPR disparity, or TPR disparity) which can be decomposed as a linear function, $h_i^0 + h_i^1 f(x_i)$, of the individual predictions $f(x_i)$ using the flip probabilities $q_{N,i}$. To see this, we can write:

$$\begin{aligned}
&\frac{1}{|R_N(\epsilon)|} \sum_{\theta \in R_N(\epsilon)} \sum_{i=1\ldots N} (h_i^0 + h_i^1 f(x_i)) \\
&\quad = \frac{1}{|R_N(\epsilon)|} \sum_{\theta \in R_N(\epsilon)} \sum_{i=1\ldots N} (h_i^0 + h_i^1 (\theta_i \mathbf{1}\{p_i \le 0.5\} \\
&+ (1-\theta_i)\mathbf{1}\{p_i > 0.5\})) \\
&\quad = \sum_{i=1\ldots N} h_i^0 + \sum_{i=1\ldots N} h_i^1 \frac{1}{|R_N(\epsilon)|} \sum_{\theta \in R_N(\epsilon)} (\theta_i \mathbf{1}\{p_i \le 0.5\} \\
&+ (1-\theta_i)\mathbf{1}\{p_i > 0.5\}) \\
&\quad = \sum_{i=1\ldots N} h_i^0 + \sum_{i=1\ldots N} h_i^1 \left(\mathbf{1}\{p_i > 0.5\} + (\mathbf{1}\{p_i \le 0.5\} \right. \\
&- \left. \mathbf{1}\{p_i > 0.5\}) \frac{\sum_{\theta \in R_N(\epsilon)} \theta_i}{|R_N(\epsilon)|}\right) \\
&\quad = \sum_{i=1\ldots N} h_i^0 + \sum_{i=1\ldots N} h_i^1 \left(\mathbf{1}\{p_i > 0.5\} + (\mathbf{1}\{p_i \le 0.5\} \right. \\
&- \left. \mathbf{1}\{p_i > 0.5\}) q_{N,i}\right).
\end{aligned}$$

Concretely, for accuracy we have $h_i^0 = \frac{1-p_i}{N}$ and $h_i^1 = \frac{2p_i - 1}{N}$. For PPR disparity, assuming wlog that subgroup A has greater PPR, we have $h_i^0 = 0$ and $h_i^1 = \frac{\mathbf{1}\{d_i \in A\}}{|A|} - \frac{\mathbf{1}\{d_i \in B\}}{|B|}$. For FPR disparity, assuming wlog that subgroup A has greater FPR, we have $h_i^0 = 0$ and $h_i^1 = \frac{(1-p_i)\mathbf{1}\{d_i \in A\}}{||1-P_A||_1} - \frac{(1-p_i)\mathbf{1}\{d_i \in B\}}{||1-P_B||_1}$. Finally, For TPR disparity, assuming wlog that subgroup A has greater TPR, we have $h_i^0 = 0$ and $h_i^1 = \frac{p_i \mathbf{1}\{d_i \in A\}}{||P_A||_1} - \frac{p_i \mathbf{1}\{d_i \in B\}}{||P_B||_1}$.

Second, as noted in Section 5.2, comparing the amount of arbitrariness (as defined by the average flip probability) across demographic groups provides a very different notion of group fairness compared to typical definitions including statistical parity and error rate balance. As a simple proof-of-concept example, imagine that we have two equally-sized subgroups A and B with Bayes-optimal probabilities $p_i = 0.6$ for all members of group A, while group B is evenly split between $p_i = 0.51$ and $p_i = 0.99$. The Bayes-optimal classifier would predict everyone as positive, leading to PPR = FPR = TPR = 1 for both groups and no observed disparities. Yet the average flip probability for models in the Rashomon set $R_N(\epsilon)$ would be greater for one group than another depending on the value of ϵ. For a small value of $\epsilon = .001$, group B would be 32% more likely to be flipped than group A, while for a larger value of $\epsilon = .02$, group B would be 14% *less* likely to be flipped.

We now present three figures discussed in Section 5.3. In Figure 7, we graph the overall (population average) flip probability for all three datasets for models sampled uniformly at random from the Rashomon set $R_N(\epsilon)$ as a function of ϵ, compared to sampling linear models from the Rashomon set (as described in Section 4.2) and the models that optimize PPR, FPR, and TPR over the Rashomon set (as described in Section 4.1).

In Figure 8, we use the flip probabilities to compute the average PPR, FPR, and TPR disparities of the entire Rashomon set $R_N(\epsilon)$ as a function of the error tolerance ϵ for the German, Adult, and Health datasets. While these quantities can also be approximated by sampling a large number of flip vectors uniformly at random from the Rashomon set and computing their sample averages, as described in Section 4.2, using the flip probabilities is both exact and much more computationally efficient. We see that the sample averages (orange curves) and entire-Rashomon-set averages (blue curves) match closely in Figure 8, but the orange curves include a small amount of random noise while the blue curves are smooth.

In Figure 9, we compute the group average flip probabilities for the protected and non-protected groups as a function of the error tolerance ϵ for the German, Adult, and Health datasets.

H Additional results for Section 6.3.1 (Rashomon set size experiments)

We plot the Rashomon set size $|R_N(\epsilon)|$ for the German Credit, Adult, and Health datasets in Figure 10. We observe that the Rashomon set sizes are very large and scale rapidly with ϵ, since $|R_N(\epsilon)| = B(\epsilon)^N$. For the maximum ϵ value we consider, $\epsilon = 0.02$, we have $B = 1.32$ for German Credit, $B = 1.22$ for Adult, and $B = 1.17$ for Health.

Additionally, while we do not yet have a way of computing the (reduced) Rashomon set size when restricting our search to the space of linear (L_2-penalized logistic regression)

models as described in Section 4.3.1 above, we can nevertheless examine what fraction of the sampled linear models are in the Rashomon set as a function of ϵ. This is shown (for $\epsilon \in \{0.001, 0.002, \ldots, 0.02\}$) for the German Credit, Adult, and Health datasets in Figure 11. We see that, for the Adult and Health datasets, most of the randomly sampled linear models are in the Rashomon set, even for low ϵ-values. For the German Credit data, a substantial fraction of linear models are not in the Rashomon set, even when ϵ is large.

I Robustness check: use of an alternate model to estimate Bayes-optimal probabilities

As noted above, the Bayes-optimal probabilities p_i are unknown for real-world datasets, but can be well-estimated using sufficient training data. In the main paper, we used logistic regression to estimate these probabilities. To check the robustness of our results to the choice of model used for estimation of p_i, we re-ran all experiments using the estimated probabilities $\hat{p}_i$ from XGBoost models learned using 5-fold cross-validation. Here we present results comparing our methods for optimizing PPR (Section 4.1.1), optimizing FPR (Section 4.1.2), optimizing TPR (Section 4.1.2), uniform random sampling (Section 4.2), and sampling linear models (Section 4.3.1) over the Rashomon set $R_N(\epsilon)$. Results for PPR disparity, FPR disparity, TPR disparity, overall flip probability, and proportion of error tolerance used, all using the XGBoost-generated probability estimates, are shown in Figures 12-16. These can be compared to the corresponding results for logistic regression-generated probability estimates for PPR disparity, FPR disparity, TPR disparity, overall flip probability, and proportion of error tolerance used in Figures 1, 5, 6, 7, and 3(right) respectively. The primary difference we observe is that none of the randomly sampled linear models were in the Rashomon set for the German and Health datasets. For the Adult dataset, we observed randomly sampled linear models in the Rashomon set for $\epsilon \geq 0.008$, as compared to $\epsilon \geq 0.001$ for the logistic regression-generated probability estimates. These differences are not surprising given that linear models might not be able to fit the more complex, non-linear relationships modeled by XGBoost. Otherwise, results are very similar to those using the logistic regression-generated probability estimates, supporting our conclusions and policy takeaways above.

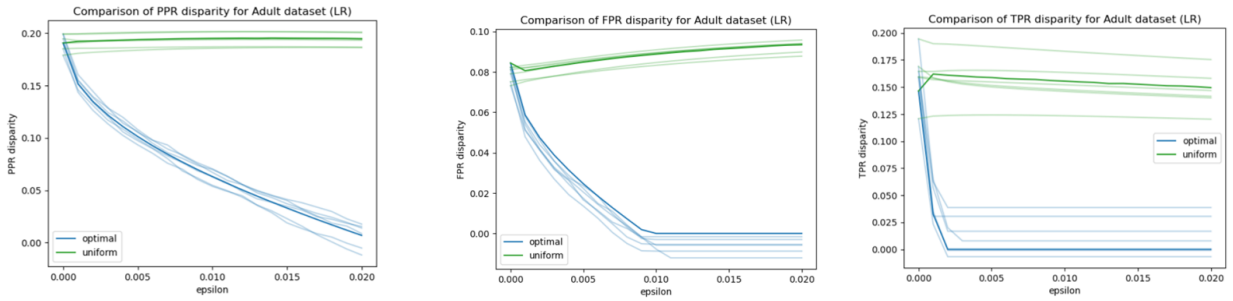

Figure 4: Generalization experiments. Disparity in positive prediction rate (left), false positive rate (center), and true positive rate (right) for the Adult dataset, as a function of the error tolerance ϵ. For optimization (Section 4.1) (blue) and uniform random sampling (Section 4.2) (green), the six dashed lines are formed using three separate partitions of the data for learning the Bayes-optimal model, learning C values for random sampling and class-specific decision thresholds for optimization, and forming disparity curves respectively; the first partition is labeled data and the second and third partitions are unlabeled. The solid lines represent our original results using the entire dataset.

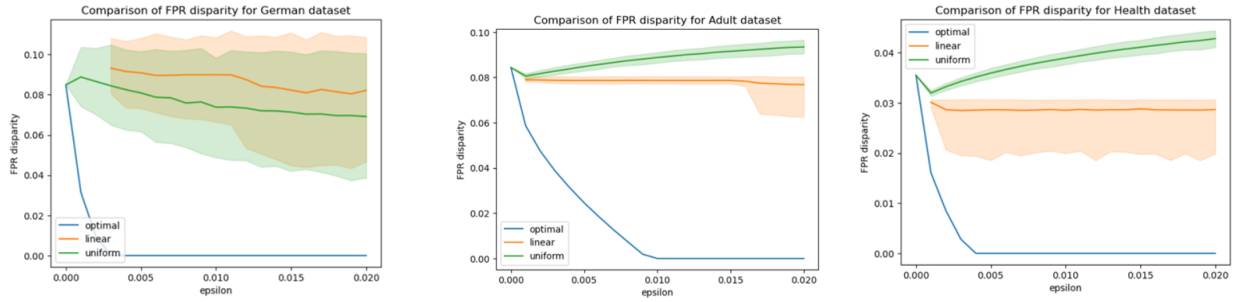

Figure 5: Disparity in false positive rate for the German, Adult, and Health datasets, as a function of the error tolerance ϵ. Comparison of methods for optimizing FPR (Section 4.1.2), uniform random sampling (Section 4.2), and sampling linear models (Section 4.3.1) over the Rashomon set $R_N(\epsilon)$.

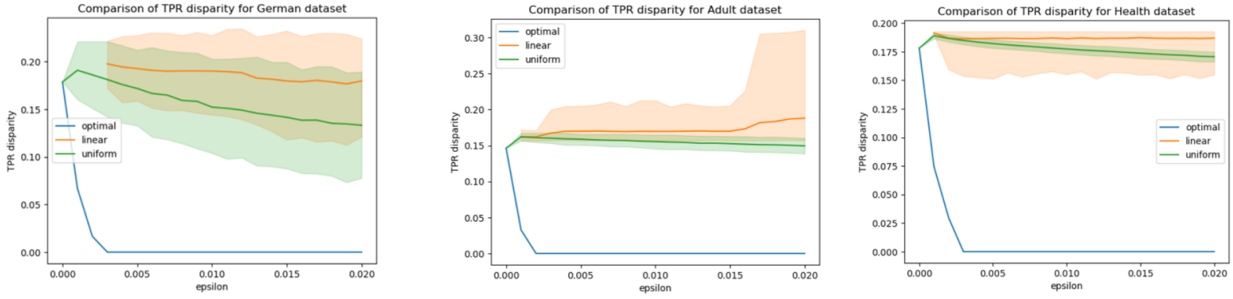

Figure 6: Disparity in true positive rate for the German, Adult, and Health datasets, as a function of the error tolerance ϵ. Comparison of methods for optimizing TPR (Section 4.1.2), uniform random sampling (Section 4.2), and sampling linear models (Section 4.3.1) over the Rashomon set $R_N(\epsilon)$.

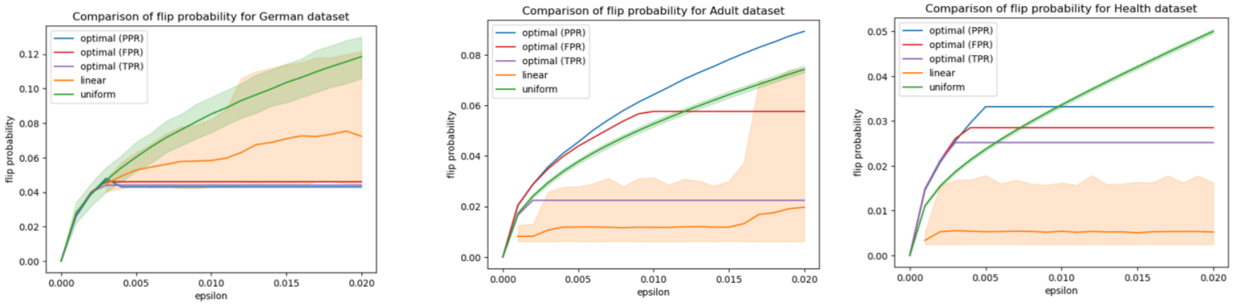

Figure 7: Overall (population average) flip probability for the German, Adult, and Health datasets, as a function of the error tolerance ϵ. Comparison of methods for optimizing PPR (Section 4.1.1), optimizing FPR (Section 4.1.2), optimizing TPR (Section 4.1.2), uniform random sampling (Section 4.2), and sampling linear models (Section 4.3.1) over the Rashomon set $R_N(\epsilon)$.

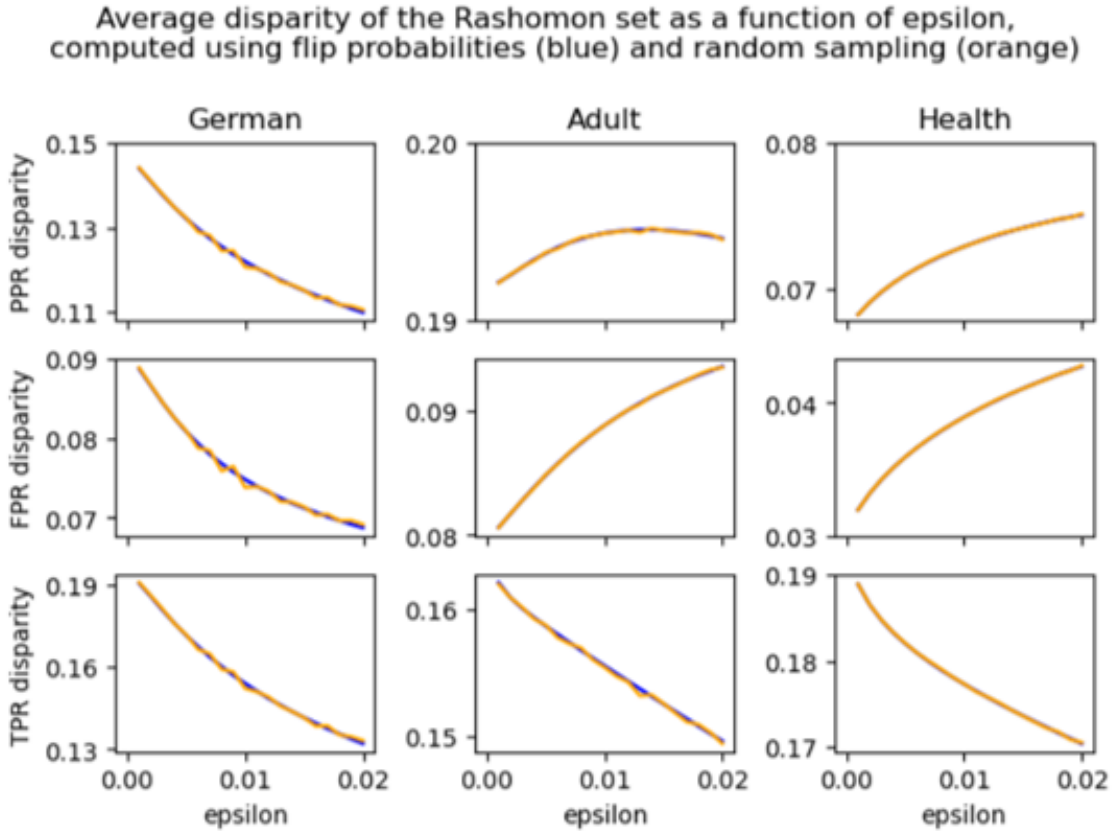

Figure 8: Comparison of calculated PPR, FPR, and TPR disparities as a function of ϵ for the German, Adult, and Health datasets. Blue curves: average disparity of the entire Rashomon set calculated using the flip probabilities, as described in Appendix G. Orange curves: average disparity of 950 flip vectors sampled uniformly at random from the Rashomon set, as described in Section 4.2

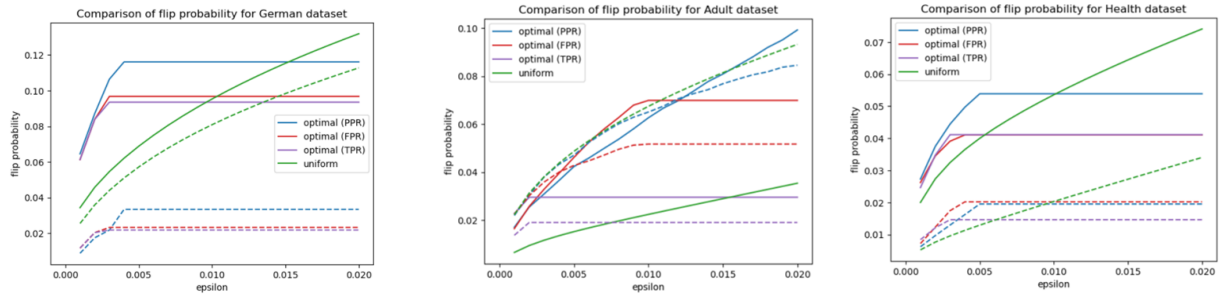

Figure 9: Group average flip probability, comparison between protected group (solid lines) and non-protected group (dashed lines), for the German, Adult, and Health datasets, as a function of the error tolerance ϵ. Comparison of methods for optimizing PPR (Section 4.1.1), optimizing FPR (Section 4.1.2), optimizing TPR (Section 4.1.2), and uniform random sampling (Section 4.2), over the Rashomon set $R_N(\epsilon)$.

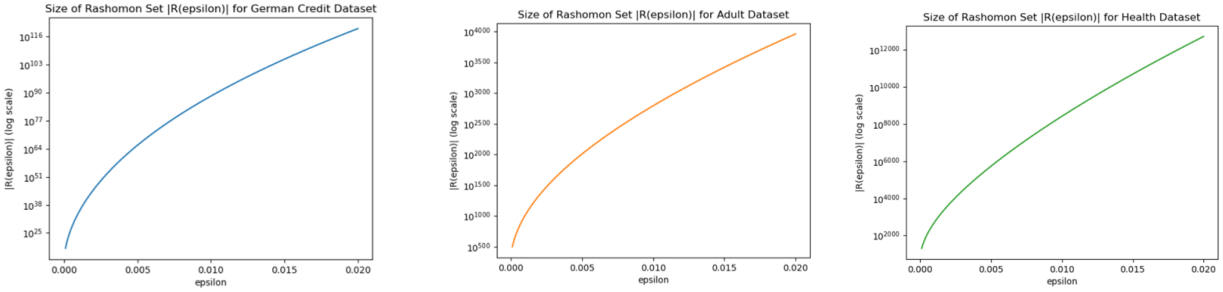

Figure 10: Rashomon set size $|R_N(\epsilon)|$ for the German Credit, Adult, and Health datasets. Note the logarithmic scale of the y-axis.

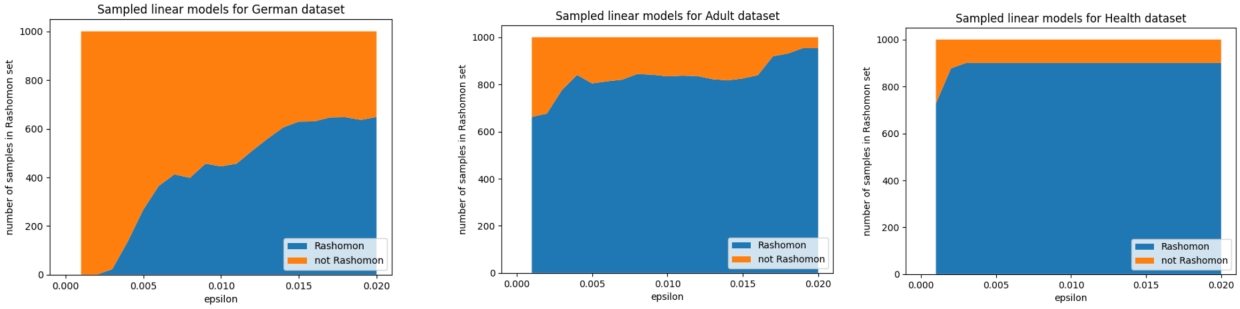

Figure 11: Proportion of randomly sampled linear models that are in the Rashomon set $R_N(\epsilon)$ as a function of the error tolerance ϵ, for the German Credit (left), Adult (center), and Health (right) datasets. The y-axis represents how many of the 1000 randomly sampled linear models are (blue) and are not (orange) in the Rashomon set.

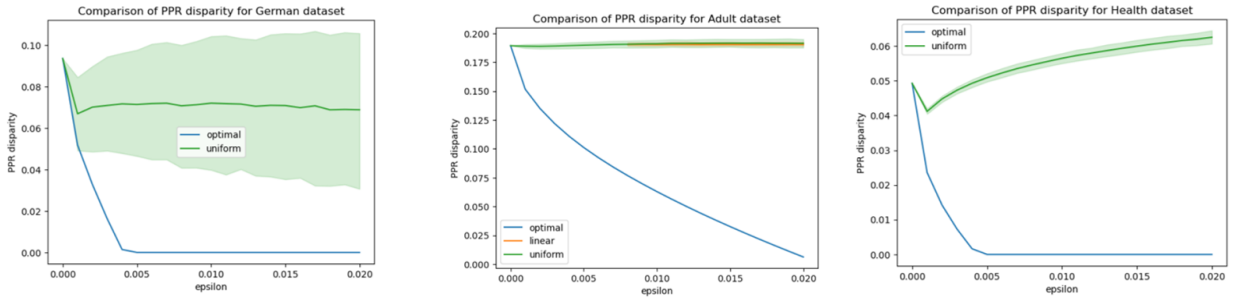

Figure 12: Robustness check using XGBoost instead of logistic regression to estimate Bayes-optimal probabilities. Disparity in positive prediction rate for the German, Adult, and Health datasets, as a function of the error tolerance ϵ. Comparison of methods for optimizing PPR (Section 4.1.1), uniform random sampling (Section 4.2), and sampling linear models (Section 4.3.1) over the Rashomon set $R_N(\epsilon)$. Note that no linear models were in the Rashomon set for German and Health datasets.

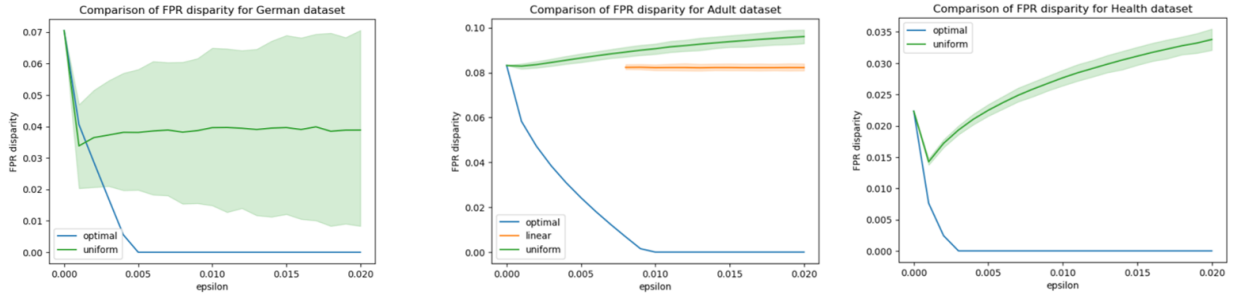

Figure 13: Robustness check using XGBoost instead of logistic regression to estimate Bayes-optimal probabilities. Disparity in false positive rate for the German, Adult, and Health datasets, as a function of the error tolerance ϵ. Comparison of methods for optimizing FPR (Section 4.1.2), uniform random sampling (Section 4.2), and sampling linear models (Section 4.3.1) over the Rashomon set $R_N(\epsilon)$. Note that no linear models were in the Rashomon set for German and Health datasets.

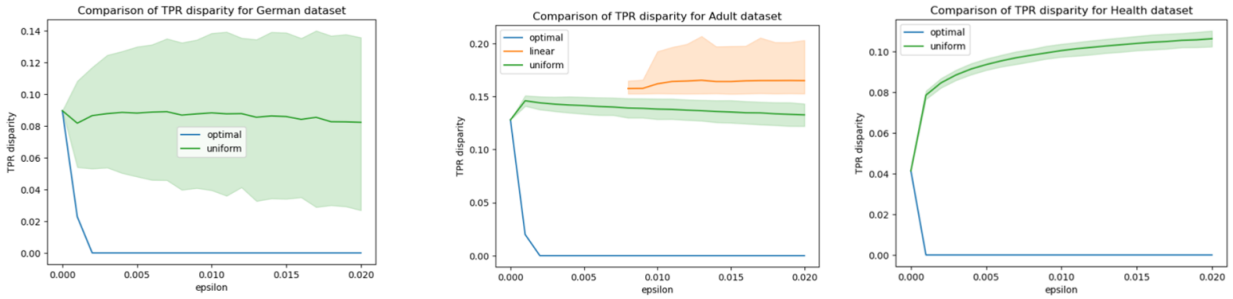

Figure 14: Robustness check using XGBoost instead of logistic regression to estimate Bayes-optimal probabilities. Disparity in true positive rate for the German, Adult, and Health datasets, as a function of the error tolerance ϵ. Comparison of methods for optimizing TPR (Section 4.1.2), uniform random sampling (Section 4.2), and sampling linear models (Section 4.3.1) over the Rashomon set $R_N(\epsilon)$. Note that no linear models were in the Rashomon set for German and Health datasets.

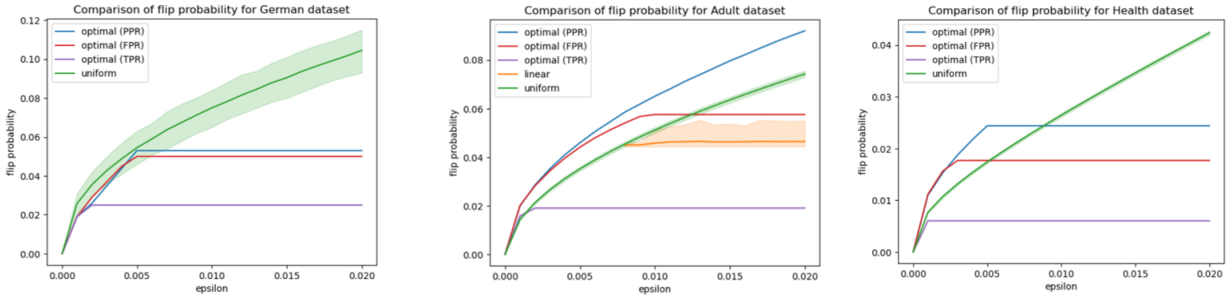

Figure 15: Robustness check using XGBoost instead of logistic regression to estimate Bayes-optimal probabilities. Overall (population average) flip probability for the German, Adult, and Health datasets, as a function of the error tolerance ϵ. Comparison of methods for optimizing PPR (Section 4.1.1), optimizing FPR (Section 4.1.2), optimizing TPR (Section 4.1.2), uniform random sampling (Section 4.2), and sampling linear models (Section 4.3.1) over the Rashomon set $R_N(\epsilon)$. Note that no linear models were in the Rashomon set for German and Health datasets.

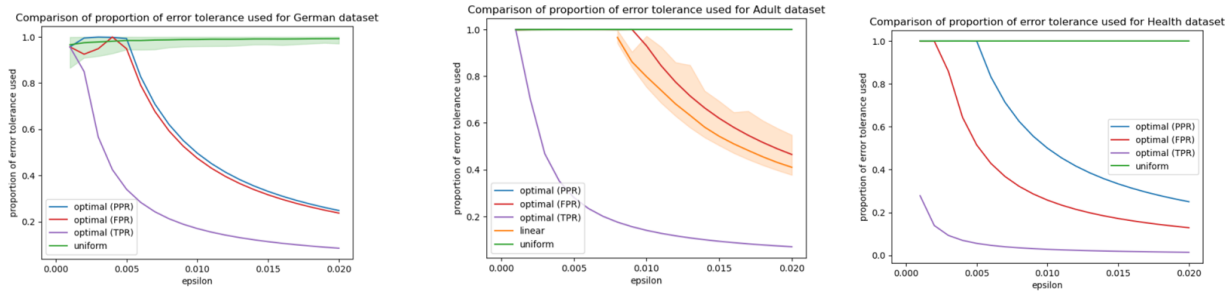

Figure 16: Robustness check using XGBoost instead of logistic regression to estimate Bayes-optimal probabilities. Proportion of error tolerance used, $\frac{\theta \cdot W_N}{N\epsilon}$, for the German, Adult, and Health datasets, as a function of the error tolerance ϵ. Comparison of methods for optimizing PPR (Section 4.1.1), optimizing FPR (Section 4.1.2), optimizing TPR (Section 4.1.2), uniform random sampling (Section 4.2), and sampling linear models (Section 4.3.1) over the Rashomon set $R_N(\epsilon)$. Note that no linear models were in the Rashomon set for German and Health datasets.

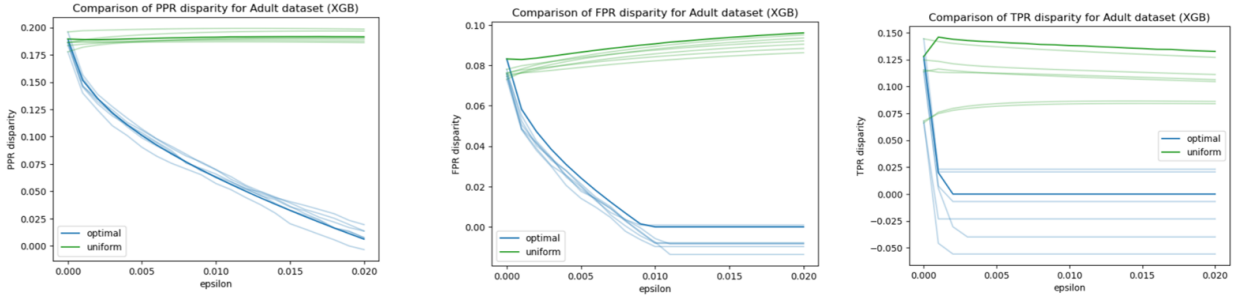

Figure 17: Robustness check using XGBoost instead of logistic regression to estimate Bayes-optimal probabilities. Disparity in positive prediction rate (left), false positive rate (center), and true positive rate (right) for the Adult dataset, as a function of the error tolerance ϵ. For optimization (Section 4.1) (blue) and uniform random sampling (Section 4.2) (green), the six dashed lines are formed using three separate partitions of the data for learning the Bayes-optimal model, learning C values for random sampling and class-specific decision thresholds for optimization, and forming disparity curves respectively; the first partition is labeled data and the second and third partitions are unlabeled. The solid lines represent our original results using the entire dataset.