On the Inevitability of the Rashomon Effect

Lucas Monteiro Paes*, Rodrigo Cruz†, Flavio P. Calmon*, and Mario Diaz†
*Harvard University, lucaspaes@g.harvard.edu, flavio@seas.harvard.edu
†Universidad Nacional Autónoma de México, rodri_cruz@comunidad.unam.mx, mario.diaz@sigma.iimas.unam.mx

***Abstract*—The *Rashomon effect* in machine learning (ML) occurs when multiple distinct models achieve similar average loss on a given learning task. The set of all models with expected loss smaller than ϵ is called the *Rashomon set*. The characterization of this set for a given learning task allows searching for models that satisfy additional constraints (e.g., interpretability, fairness) without compromising accuracy. Though folklore treats the Rashomon set as the collection of all indistinguishable "good" models, there are no established theoretical guarantees that models in this set are statistically indistinguishable. We fill this gap by proposing a hypothesis test framework to choose the best-performing model between two elements in the Rashomon set and derive lower and upper bounds for its probability of error. Specifically, we prove that for any $\epsilon > 0$ if the data set has less than $O\big([\epsilon \log(\epsilon/(1-\epsilon))]^{-1}\big)$ instances, models in the Rashomon set are statistically indistinguishable and the Rashomon effect is inevitable. Additionally, our bounds can guide data scientists to choose an ϵ that generates a Rashomon set so that any two models in it are indistinguishable.**

I. Introduction

The *Rashomon effect* describes the phenomenon where multiple machine learning (ML) models achieve similar performance on a given prediction task. This effect was first reported two decades ago by Breiman [1]. In a foresighted experiment, Breiman retrained a neural network one hundred times on three-dimensional data with a different parameter initialization for each run and observed 32 distinct local minima [1]. The Rashomon effect has since been reported across model classes, including ridge regression, decision trees, and deep neural networks [2], [3], and in several learning tasks, including computer vision, medical diagnostics, and natural language processing [4], [5]. At the heart of the Rashomon effect is the *Rashomon set*: the set of all models with average loss below ϵ on a given learning task [2], [3] (see Def. 1). We refer to ϵ as the *Rashomon parameter*.

The existence of non-trivial Rashomon sets poses both risks and opportunities. On the one hand, the arbitrary selection of a single model from the Rashomon set challenges the credibility of classifiers deployed in practice [4]. Models in the Rashomon set can produce wildly conflicting predictions on individual samples, a phenomenon known as *predictive multiplicity* [6]. On the other hand, the Rashomon effect can be leveraged to select models that satisfy additional properties without compromising accuracy, such as fairness [7], interpretability [8], and stability [9]. Even though the Rashomon set is often treated as the set of all statistically indistinguishable "good" models [3], [7], [8], there are currently no established theoretical methods for delineating this set. Such characterization was recently posed as one of the 10 Grand Challenges in interpretable ML [10].

In this paper, we aim to answer two questions. Firstly, "*Is the Rashomon effect inevitable?*". We show that, under certain assumptions, there is an information-theoretic limit on when two models can be distinguished in terms of their average accuracy (Prop. 1). This limit only depends on the cardinality of the test set and the loss function. As a consequence, when only finite test data is available, the Rashomon set for most model classes will be non-trivial: there will be a set of "good" models that cannot be distinguished through pairwise comparisons (Theorem 1). Thus, the Rashomon effect is indeed information-theoretically inevitable when only finite data is available.

The study of the Rashomon effect also requires careful choice of the Rashomon parameter ϵ. Currently, there is no established method for selecting ϵ when characterizing the Rashomon set of "good" models, with authors often choosing ϵ as an arbitrary value (e.g., loss within 1% of a reference model [7]). Therefore, our second question of interest is "*How should the Rashomon parameter ϵ be chosen?*" We provide bounds that allow the calculation of the maximum Rashomon parameter ϵ that ensures that two models in the Rashomon set are provably statistically indistinguishable (Theorems 2 and 3). Again, our results depend on the sample size and loss function used to evaluate model performance. Our bounds can help data scientists delineate the set of statistically indistinguishable "good" models and avoid an ad hoc choice of ϵ.

Related work. There are several potential causes for the Rashomon effect, including model underspecification [4], the use of training algorithms that converge to local minima and rely on randomization (e.g., stochastic gradient descent) [9], and the computational hardness of identifying a global minimum in certain learning tasks [1]. We focus instead on the inevitability of the Rashomon effect when only a finite amount of data is available to evaluate and compare ML models.

Semenova *et al.* [2] studied the volume of the Rashomon set to analyze the existence of simple ML models. The work closest to ours is [3], which proposes methods for learning variable importance by exploring the Rashomon set and provides concentration results for metrics defined in the Rashomon set. The main difference is that we focus on the existence of the Rashomon effect itself and the choice of the Rashomon parameter. Our work aims to determine when the empirical losses of models in the Rashomon set are information-theoretically indistinguishable (i.e., the Rashomon effect is

This work is partially funded by the National Science Foundation under grants CAREER 1845852, FAI 2040880, and CIF 1900750.

inevitable) and to provide a method for selecting the Rashomon parameter based on this converse result.

A. Setup & notation

Let $X \in \mathcal{X}$, and $Y \in \mathcal{Y}$ be two random variables with joint distribution $P_{X,Y}$. Here, X represents a set of characteristics used as input for a ML model, and Y is the target predicted output. We denote a model hypothesis class by $\mathcal{H} \subset \{h \mid h : \mathcal{X} \to \mathcal{Y}\}$ (e.g., logistic regression or support vector machines). We assume we are given a set of n independent and identically distributed random vectors $S_n \triangleq \{(X_i, Y_i)\}_{i=1}^n$. In Section II, we select a sample size of $2n$ to avoid dependence between the empirical risk of two classifiers. For this particular case, given $h_1, h_2 \in \mathcal{H}$, we define $\mathbf{Z}(h_1, h_2, S_{2n}) \triangleq \{\ell(h_1(X_i), Y_i), \ell(h_2(X_{n+i}), Y_{n+i}))\}_{i=1}^n$. Throughout this paper, we focus on classification and mainly consider the 0-1 loss; i.e. $\ell(x, y) = \mathbb{1}_{\{h(x) \neq y\}}$. We denote the population loss of an element $h \in \mathcal{H}$ by $L(h) = \mathbb{E}[\ell(h(X), Y)]$, and its empirical loss by $\hat{L}_{S_n}(h) = \frac{1}{n}\sum_{i=1}^n \ell(h(X_i, Y_i))$. When S_n is clear from the context we denote $\hat{L}_{S_n}(h) \triangleq \hat{L}(h)$. We refer to a binomial random variable with n trials and probability of success p as $\text{Bin}(n, p)$. A beta random variable has a continuous probability distribution defined on $[0, 1]$, we denote its cumulative distribution function by:

$$I_x(a, b) = \frac{1}{B(a, b)} \int_0^x u^{a-1}(1-u)^{b-1}\, du, \tag{1}$$

where $a, b > 0$ and $B(a, b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$. We denote its quantile function as $\text{qBeta}(\,\cdot\,; a, b)$.

The Rashomon set is defined as the set of all models whose performance surpasses a certain threshold. This threshold, denoted by ϵ, is called the *Rashomon parameter*. Next, we define the Rashomon set and its empirical version.

Definition 1 (Rashomon Set, [2], [3])**.** The *true Rashomon set* and the *empirical Rashomon set* are defined, respectivelly, as:

$$\mathcal{R}(\epsilon, \mathcal{H}) \triangleq \{h \in \mathcal{H} : L(h) \leq \epsilon\}, \tag{2}$$

$$\hat{\mathcal{R}}(\epsilon, \mathcal{H}, S_n) \triangleq \left\{h \in \mathcal{H} : \hat{L}(h) \leq \epsilon\right\}. \tag{3}$$

When other parameters are clear from the context, we denote the Rashomon set by $\mathcal{R}(\epsilon)$ and its empirical version by $\hat{\mathcal{R}}(\epsilon)$.

B. Overview of Main Results

We start by studying pairwise comparisons between models. We use these comparisons to rigorously define the loosely used term "indistinguishable models" via a hypothesis testing framework (Section II). We also propose a more practical approach based on confidence intervals (Section III). We then show that, when only a finite amount of data is available to test performance, the Rashomon effect is inevitable (Section IV) – i.e., there exists a Rashomon set such that the empirical loss of any two models in this set are indistinguishable according to our definitions. Finally, we apply these results to select the Rashomon parameter such that the empirical performance of all models in the Rashomon set are provably indistinguishable when evaluated on a finite dataset (Section V). The proofs of theoretical results and code presented in this paper are available in *https://github.com/LucasMonteiroPaes/On-the-Inevitability-of-the-Rashomon-Effect*. Our **main contributions** are:

- We propose a hypothesis test framework for pairwise comparison of the empirical losses of two models evaluated over a test set with finite samples. We cast the problem of selecting the Rashmon parameter as determining the minimum ϵ for which this hypothesis test fails for any pair of models in the Rashomon set.
- We provide bounds for the probability of error in the hypothesis test framework. Remarkably, we show that it is necessary to have at least $O\big([\epsilon \log(\epsilon/(1-\epsilon))]^{-1}\big)$ data points in the test set in order to distinguish two models in the Rashomon set with size ϵ, making the Rashomon effect unavoidable (Theorem 1).
- Finally, we provide methods that can guide the choice of Rashomon parameter ϵ while ensuring that any two models in the Rashomon set are indistinguishable.

II. Classifier comparison via hypothesis testing

This section uses a hypothesis-testing approach to compare the performance of two classifiers under 0-1 loss. Recall that the *Rashomon effect* happens when at least a pair of different models achieve similar average performance, yet one cannot determine which classifier has a better (smaller) expected loss. The following hypothesis test establishes this decision problem.

Definition 2 (Hypothesis test for comparing empirical losses)**.** Given two models $h_1, h_2 \in \mathcal{H}$, hypothesis H_0 and H_1 are defined as

$$H_0 : L(h_1) \leq L(h_2) \quad \text{vs} \quad H_1 : L(h_1) > L(h_2).$$

The probability of error of a *decision function* $\psi : \mathbf{Z} \mapsto \{0, 1\}$ that accepts/rejects H_0 is given by

$$P_{\text{error}}(\psi) \triangleq \frac{1}{2}[\Pr(\psi(\mathbf{Z}) = 0 \mid H_1) + \Pr(\psi(\mathbf{Z}) = 1 \mid H_0)],$$

where $\mathbf{Z}$ represents $\mathbf{Z}(h_1, h_2, S_{2n})$ as defined in Section I-A.

We study the probability of error associated with this hypothesis test, i.e., $\min_\psi P_{\text{error}}(\psi)$, to understand the fundamental limitations of model selection. This quantity essentially summarizes the distinguishability between the two models. Intuitively, if the error probability is close to 0.5, any strategy is almost equivalent to a random guess – i.e., the models are statistically indistinguishable. On the other hand, if $\min_\psi P_{\text{error}}(\psi)$ is small, then there is a way to ascertain the best model. Definition 3 makes this intuition precise.

Definition 3 (α-testing indistinguishable)**.** Given $\alpha \in (0, 0.5)$ we say that two models h_1 and h_2 are *α-testing indistinguishable* if $P_{\text{error}}(\psi) \geq \alpha$ for every possible test ψ, i.e.:

$$\min_\psi P_{\text{error}} \geq \alpha. \tag{4}$$

Conversely, h_1 and h_2 are *α-testing distinguishable* if $P_{\text{error}}(\psi) < \alpha$ for some test ψ.

If $L(h_1)$ and $L(h_2)$ are close enough, it is challenging to choose the best model regardless of the decision function ψ, i.e., if $|L(h_1) - L(h_2)|$ is sufficiently small, we expect P_{error} to be large across all possible tests. Conversely, if $|L(h_1) - L(h_2)|$ is sufficiently large, we expect the optimal test to achieve a small P_{error}. The following result reflects this intuition.

Proposition 1 (Bounds for P_{error})**.** *If ℓ is the 0-1 loss and $h_1, h_2 \in \mathcal{H}$ are such that $0 < \epsilon_0 \leq L(h_1), L(h_2) \leq \epsilon$, then:*

$$\inf_{\psi} P_{\text{error}}(\psi) \geq \frac{1}{2}\left[1 - \sqrt{1 - \exp\left(-n(\epsilon - \epsilon_0)\log\left(\frac{\epsilon(1-\epsilon_0)}{\epsilon_0(1-\epsilon)}\right)\right)}\right]$$

$$\inf_{\psi} P_{\text{error}}(\psi) \leq \frac{1}{2}\left(1 - \frac{|L(h_1) - L(h_2)|^2}{8(1-\epsilon_0)\epsilon}\right)^n.$$

Remark 1. Proposition 1 depends on the existence of a lower bound for all losses (ϵ_0). This dependence arises from approximating the Total Variation distance by the KL-divergence, which "explodes" for $L(h) = 0$ – for more details, see the Appendix. Using the Hellinger distance bound on Total Variation [11] it is possible to remove the side effect of KL-divergence at the cost of a looser bound when ϵ_0 exists.

The upper bound in Proposition 1 depends on the true losses of h_1 and h_2, hence it is not computable in practice. However, this result is sufficient to prove that models in the Rashomon set are indistinguishable. To do so, we analyze an optimized version of the upper bound that does not depend on the losses – see Theorem 1.

The assumption that there exists a universal lower bound $\epsilon_0 > 0$ for the model loss is reasonable since we do not expect the loss to be zero except in certain deterministic classification tasks [12]. For instance, $\epsilon_0 > 0$ will hold if the model class $\mathcal{H}$ does not contain the model that generated the map $X \to Y$, or when even the maximum a posterior estimator does not achieve 0 loss (i.e., there is inherent uncertainty on Y given X). Nevertheless, computing the bound in Proposition 1 in practice does require knowledge of a lower bound ϵ_0 for the true loss of all models in the model class. Alas, this lower bound might be hard to find in many applications. For this reason, we introduce the confidence-interval approach for model distinguishability in the next section. Intuitively, two models are indistinguishable when the confidence intervals for their expected losses overlap. This approach is computationally straightforward and can be used to explicitly calculate the Rashomon parameter.

III. Classifier comparison via confidence intervals

We present next a decision function for performing the hypothesis test in Definition 2 based on comparing confidence intervals (CIs). Although comparing overlapping CIs may not be a **rigorous** test of equality of means [13], the simple and graphical nature of this approach makes it more likely to be used by practitioners – this confidence approach is informally used in the Rashomon effect literature [6]. Given that this method usually depends on the sample means, we analyze how close the empirical losses of two classifiers need to be to consider them indistinguishable from this practical perspective. In Section V, we will use the results from this section to calculate the maximum Rashomon parameter (ϵ) such that all models in the empirical Rashomon set $\hat{\mathcal{R}}(\epsilon)$ are δ-confidence indistinguishable.

In contrast to Section II, we assume that all predictors are tested against the same unique dataset $S_n = \{(X_i, Y_i)\}_{i=1}^n$ and analyze when methods to compute Confidence Intervals (CIs) prove futile to discern between elements of $\hat{\mathcal{R}}(\epsilon)$.

If the CIs for two unknown means do not overlap, these parameters are considered significantly different. However, if the CIs overlap, one cannot conclude that they are significantly similar (see, *e.g.*, [14]). Nevertheless, given that the latter situation leads to difficulties in decision making, we deem two predictors as indistinguishable from a CIs perspective if these two sets overlap.

Definition 4 (δ-confidence indistinguishable)**.** Given a threshold $\delta \in (0, 1)$, two models h_1 and h_2 with empirical losses $\hat{L}(h_1)$ and $\hat{L}(h_2)$, define confidence intervals $I_\delta(h_1)$ and $I_\delta(h_2)$ for the losses with confidence $1 - \delta$. We say that h_1 and h_2 are *δ-confidence indistinguishable* if:

$$I_\delta(h_1) \cap I_\delta(h_2) \neq \emptyset. \tag{5}$$

Conversely, h_1 and h_2 are *δ-confidence distinguishable* if $I_\delta(h_1) \cap I_\delta(h_2) = \emptyset$.

Remark 2. There exists multiple $1 - \delta$ CIs, and Definition 4 is agnostic to any specific interval construction. We outline two procedures for constructing CIs: a measure concentration approach – which has optimal order – and the Clopper-Pearson method – which provides an exact confidence interval for the 0-1 loss.

To obtain a confidence interval for **any** bounded loss function with confidence at least $1 - \delta$, we use Hoeffding's Inequality [15, Prop. 2.7]. Via this method, we obtain an interval $I_\delta(h)$ for $L(h)$. Specifically, this set is given by:

$$I_\delta(h) = \left[\hat{L}(h) - \sqrt{\frac{\log(2/\delta)}{2n}}, \hat{L}(h) + \sqrt{\frac{\log(2/\delta)}{2n}}\right].$$

Henceforth, we refer to this CI construction procedure as the *Hoeffding's Inequality method*. A consequence of the explicit nature of this approach is the following.

Proposition 2. *For any given loss function ℓ bounded by 1, under the Hoeffding's Inequality method, two models $h_1, h_2 \in \mathcal{H}$ are δ-indistinguishable if and only if*

$$|\hat{L}(h_1) - \hat{L}(h_2)| \leq 2\sqrt{\frac{\log(2/\delta)}{2n}}.$$

The Hoeffding's Inequality method and Proposition 2 hold for every bounded loss function. However, for the case where the loss of interest is the 0-1 loss, an optimal approach exists to create confidence intervals: the Clopper-Pearson (CP) interval [16]. This *exact* method is described in [17, Sec. 4.2.1] as follows: let X be a $\text{Bin}(n, p)$ random variable where p is unknown. If x is observed, then the Clopper-Pearson interval of confidence level $1 - \delta$ is

$$I_\delta = [L_\delta(x), U_\delta(x)],$$

where $L_\delta(x)$ and $U_\delta(x)$ are the solutions in θ to equations:

$$\Pr(\text{Bin}(n,\theta) \geq x) = \frac{\delta}{2} \quad \text{and} \quad \Pr(\text{Bin}(n,\theta) \leq x) = \frac{\delta}{2}.$$

Next, we specify when two hypotheses are δ-confidence indistinguishable under the CP method.

Proposition 3. *Under the CP method, the hypotheses h_1 and h_2 are δ-indistinguishable if and only if*

$$\begin{aligned}&\text{qBeta}(\delta/2; n\max \hat{L}(h_i), n - n\max \hat{L}(h_i) + 1)\\ \leq &\text{qBeta}(1-\delta/2; n\min \hat{L}(h_i) + 1, n - n\min \hat{L}(h_i)),\end{aligned}$$

where the maximum and minimum are taken over $i \in \{1,2\}$.

Proposition 2 and Proposition 3 reveal when two models are δ-confidence indistinguishable. In the next section, we use these results, together with the results from Section II, to show that the Rashomon effect is inevitable, *i.e.*, we show that there always exists a Rashomon parameter such that all models in the associated Rashomon set are (i) δ-confidence indistinguishable, or (ii) α-testing indistinguishable.

IV. Rashomon effect inevitability

In this section, we first show that the *Rashomon effect is inevitable* using the hypothesis test framework introduced in Section II: given finite samples for evaluating empirical loss, a *true* Rashomon set exists such that any two models in this set are α-testing indistinguishable. We then prove an analogous result using the CI approach: an *empirical* Rashomon set exists such that any two models in it are δ-confidence indistinguishable.

Testing approach. Our goal is to bound the minimum probability of error in the Rashomon set, i.e.,

$$\inf_{h_1,h_2 \in \mathcal{R}(\epsilon)} \inf_{\psi} P_{\text{error}}(\psi).$$

Theorem 1 shows that, when the Rashomon parameter ϵ is sufficiently small, the Rashomon effect is inevitable and, conversely, when the Rashomon parameter is large enough, there are distinguishable models in the Rashomon set.

Theorem 1 (Rashomon Effect Inevitability by Testing)**.** *Let $\epsilon > 0$ be the Rashomon parameter, and $\epsilon_0 > 0$ be a lower bound for the loss of all models in the Rashomon set – i.e., $\epsilon_0 \leq L(h) \quad \forall h \in \mathcal{R}(\epsilon)$. Fixed $\alpha \in (0, 0.5)$ we have:*

1) *If $(\epsilon - \epsilon_0)\log\Big(\frac{\epsilon(1-\epsilon_0)}{\epsilon_0(1-\epsilon)}\Big) \leq \frac{1}{n}\log\Big(\frac{1}{1-(1-2\alpha)^2}\Big)$, then any two models in $\mathcal{R}(\epsilon)$ are α-testing indistinguishable.*
2) *If $\frac{(\epsilon-\epsilon_0)^2}{(1-\epsilon_0)\epsilon} \geq 8\Big(1 - [2\alpha]^{1/n}\Big)$, then there exists two models in $\mathcal{R}(\epsilon)$ that are α-testing distinguishable.*

In Figure 1, we show the distinguishability regions from Theorem 1. In the red region, all models in the Rashomon set are indistinguishable. In the blue area, there are models in the Rashomon set that are distinguishable. Finally, in the gray region, our bounds do not capture (in)distinguishability.

Confidence approach. We show that there exists an *empirical* Rashomon set such that any two models in it are δ-confidence indistinguishable. Unlike Theorem 1, the confidence approach can be used in practice to find the largest ϵ such that all models in the empirical Rashomon set are indistinguishable (i.e., no "gray region" as in Fig. 1).

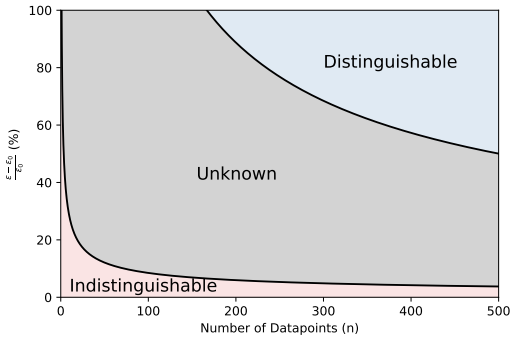

Fig. 1. Indistinguishable vs. distinguishable regions using the hypothesis test approach. The red region refers to the setup where models are indistinguishable, the blue region to the setup where models are distinguishable, and the gray refers to the region where model distinguishability is unknown. In this test we take $\epsilon_0 = 0.01$, $\epsilon = (1 + \text{y-axis})\epsilon_0$, and n was calculated using Theorem 1.

Theorem 2. *For any loss function ℓ bounded by 1, under Hoeffding's Inequality method, any two hypotheses in the empirical Rashomon Set ($\hat{\mathcal{R}}(\epsilon)$) are δ-indistinguishable if and only if ϵ satisfies*

$$\epsilon - \epsilon_0 \leq 2\sqrt{\frac{\log(2/\delta)}{2n}},$$

where $\epsilon_0 = \min_{h\in\mathcal{H}} \hat{L}(h)$.

Using Proposition 3, it is possible to decide if the CP intervals overlap, i.e., if models are δ-confidence indistinguishable using the CP interval. Theorem 3 shows that all models in $\hat{\mathcal{R}}(\epsilon)$ are δ-indistinguishable for sufficiently small ϵ.

Theorem 3. *Under the CP method, any two hypotheses in the empirical Rashomon Set are δ-indistinguishable if and only if:*

$$\text{qBeta}(\delta/2; n\epsilon, n - n\epsilon + 1) \leq \text{qBeta}(1-\delta/2; n\epsilon_0 + 1, n - n\epsilon_0).$$

where $\epsilon_0 = \min_{h\in\mathcal{H}} \hat{L}(h)$.

V. Rashomon parameter selection

Theorems 1, 2, and 3 reveal that the Rashomon effect is inevitable when models are compared using a finite dataset to evaluate their loss. In this section, we illustrate how these results can be applied to select a Rashomon parameter ϵ in order to ensure that all models in the Rashomon set are, in fact, statistically indistinguishable.

Hypothesis test framework. Let $X_i \in \mathbb{R}^d$ and $Y_i = \text{BSC}_p\big((\text{sgn}(\theta^T X_i))\big)$ where BSC is a binary symmetric channel with parameter $p = 0.01$, and sgn is the sign function. Suppose we are fitting a model of the form $\hat{Y} = \text{sgn}(\hat{\theta}^T X)$ to our data. The best we can do is take $\hat{\theta} = \theta$. Therefore, the smallest population 0-1 loss possible is $p = 0.01$.

In Figure 2, we vary the dataset size n and the lower bound for the probability of error α to find values the maximum

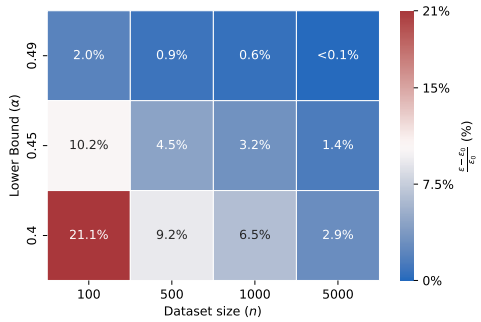

Fig. 2. Dataset size vs. lower bound trade-off for ϵ selection using the test approach. For each dataset size n (x-axis) and lower bound α (y-axis), the maximum Rashomon parameter $\epsilon = (1+\gamma)\epsilon_0$ with γ displayed in the z-axis.

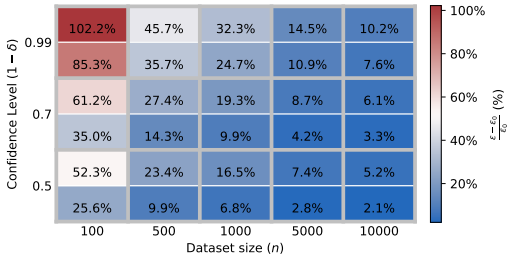

Fig. 3. Dataset size vs. Confidence trade-off for ϵ selection using the confidence approach. For each dataset size n x-axis and lower bound δ y-axis, we find the maximum Rashomon parameter $\epsilon = (1+\gamma)\epsilon_0$ with two values for γ in the z-axis: **top** of the cell for the Hoeffding's inequality and **botton** of the cell for CP.

Rashomon parameter ϵ – reached by our bounds – such that any two models in the Rashomon set are indistinguishable. Accordingly to Theorem 1 (1) with $\epsilon_0 = p = 0.01$, finding such ϵ is equivalent to solve the following optimization problem:

$$\epsilon = \sup\left\{\epsilon < 0.5 \mid (\epsilon - \epsilon_0)\log\left(\frac{\epsilon(1-\epsilon_0)}{\epsilon_0(1-\epsilon)}\right) \leq \frac{2(1-2\alpha)^2}{n}\right\}.$$

Confidence interval approach. Given a test set and the corresponding empirical risk minimizer, we find the maximum Rashomon parameter ϵ that makes any two models in $\hat{\mathcal{R}}(\epsilon)$ indistinguishable under the procedures described in Section III.

For the Hoeffding's inequality method, the maximum is given by:

$$\epsilon = \epsilon_0 + 2\sqrt{\frac{\log(2/\delta)}{2n}}.$$

For the CP approach, we appeal to Lemma 3 and Theorem 3, to conclude that if $\omega \in \mathbb{R}$ is the solution to:

$$\text{qBeta}(\delta/2; \omega, n-\omega+1) = \text{qBeta}(1-\delta/2; n\epsilon_0+1, n-n\epsilon_0),$$

then the maximum Rashomon parameter for which any two models in $\hat{\mathcal{R}}(\epsilon)$ are indistinguishable under the CP method is

$$\epsilon = \frac{\lceil \omega \rceil}{n}.$$

Figure 3 shows the values for the maximum Rashomon parameter that makes any two models in the empirical Rashomon set δ-confidence indistinguishable for different values of confidence (δ) and dataset size (n).

Our results suggest that, for datasets of size close to $n = 5$k (*e.g.*, COMPAS [18]), if the Rashomon parameter is smaller than 107.4% of ϵ_0, the Hoeffding's inequality method will deem every predictor in empirical Rashomon set δ-confidence indistinguishable even for small confidences $\delta \leq 0.5$. On the other hand, under the CP method, only if ϵ is smaller than 102.8% the elements of $\hat{\mathcal{R}}(\epsilon)$ would be δ-confidence indistinguishable. However, it is common in the literature to only treat 1% loss tolerance as indistinguishable models (see [7] and [5]). This ad hoc choice of ϵ is likely severely underestimating the size of the Rashomon set.

VI. Concluding Remarks and Limitations

When machine learning models are used in applications of individual consequences, such as medicine and lending, the Rashomon effect can have both negative and positive consequences. The existence of multiple competing models for a given learning task may challenge the credibility of data-driven decisions supported by the output of a single model selected based on empirical (average) performance (e.g., as in predictive multiplicity [6]). Alternatively, the existence of a large number of models in the Rashomon set creates opportunities for model selection in terms of criteria beyond accuracy, such as fairness and interpretability [3].

In this paper, we propose an information-theoretic approach to pairwise model comparison. We proved that, given a finite dataset, the Rashomon effect is inevitable – i.e., models in a true Rashomon set are indistinguishable. We also introduce a more practical model comparison approach by using confidence intervals. With this, we show that models in an empirical Rashomon set are indistinguishable. Finally, we leverage our results to design methods to select the Rashomon parameter, allowing data scientists to explore the Rashomon set with theoretical guarantees that it is composed only of indistinguishable models.

Our bounds require the existence of an $\epsilon_0 > 0$ that lower-bounds the loss of all models in the Rashomon set. This assumption is, in part, a result of using KL-divergence to bound total variation in the derivation of our results. However, as discussed earlier, the existence of such ϵ_0 is not unreasonable: If the model class $\mathcal{H}$ (that only contains deterministic maps) does not include the model that generated the map $X \to Y$ (this map may be random), then such ϵ_0 exists. Also, we only show that the Rashomon set is information-theoretically inevitable when the loss being used is the 0-1 loss (Theorem 1). An interesting direction of future work is to generalize this result for all bounded loss functions.

REFERENCES

[1] L. Breiman, "Statistical modeling: The two cultures," *Statistical Science*, vol. 16, no. 3, pp. 199–231, 2001.

[2] L. Semenova, C. Rudin, and R. Parr, "On the existence of simpler machine learning models," in *2022 ACM Conference on Fairness, Accountability, and Transparency*. ACM, 2022, pp. 1827–1858.

[3] A. Fisher, C. Rudin, and F. Dominici, "All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously," *Journal of Machine Learning Research - JMLR*, vol. 20, pp. 1–81, 2019.

[4] A. D'Amour, K. Heller, D. Moldovan, B. Adlam, B. Alipanahi, A. Beutel, C. Chen, J. Deaton, J. Eisenstein, M. D. Hoffman *et al.*, "Underspecification presents challenges for credibility in modern machine learning," *Journal of Machine Learning Research*, 2020.

[5] H. Hsu and F. Calmon, "Rashomon capacity: A metric for predictive multiplicity in classification," in *Advances in Neural Information Processing Systems*, 2022.

[6] C. Marx, F. Calmon, and B. Ustun, "Predictive multiplicity in classification," in *Proceedings of the 37th International Conference on Machine Learning*. PMLR, 2020, p. 6765–6774.

[7] A. Coston, A. Rambachan, and A. Chouldechova, "Characterizing fairness over the set of good models under selective labels," in *International Conference on Machine Learning*. PMLR, 2021, pp. 2144–2155.

[8] R. Xin, C. Zhong, Z. Chen, T. Takagi, M. Seltzer, and C. Rudin, "Exploring the whole rashomon set of sparse decision trees," in *Advances in Neural Information Processing Systems*, 2022.

[9] E. Black, K. Leino, and M. Fredrikson, "Selective ensembles for consistent predictions," in *The Tenth International Conference on Learning Representations (ICLR)*, 2022.

[10] C. Rudin, C. Chen, Z. Chen, H. Huang, and L. Semenova, "Interpretable machine learning: Fundamental principles and 10 grand challenges," *Statistics Surveys*, vol. 16, 2022.

[11] Y. Polyanskiy and Y. Wu, "Lecture notes on information theory," *Lecture Notes for ECE563 (UIUC) and*, vol. 6, no. 2012-2016, p. 7, 2014.

[12] O. Shamir, "Distribution-specific hardness of learning neural networks," *The Journal of Machine Learning Research*, vol. 19, no. 1, pp. 1135–1163, 2018.

[13] N. Schenker and J. F. Gentleman, "On judging the significance of differences by examining the overlap between confidence intervals," *The American Statistician*, vol. 55, no. 3, pp. 182–186, 2001.

[14] A. Knezevic, "Overlapping confidence intervals and statistical significance," *StatNews: Cornell University Statistical Consulting Unit*, vol. 73, no. 1, 2008.

[15] P. Massart, *Concentration inequalities and model selection: Ecole d'Eté de Probabilités de Saint-Flour XXXIII-2003*. Springer, 2007.

[16] C. J. Clopper and E. S. Pearson, "The use of confidence or fiducial limits illustrated in the case of the binomial," *Biometrika*, vol. 26, no. 4, pp. 404–413, 1934.

[17] L. D. Brown, T. T. Cai, and A. DasGupta, "Interval estimation for a binomial proportion," *Statistical science*, vol. 16, no. 2, pp. 101–133, 2001.

[18] J. Angwin, J. Larson, S. Mattu, and L. Kirchner, "Machine bias," *ProPublica*, 2016.

[19] J. Duchi, "Lecture notes for statistics 311/electrical engineering 377," *URL: https://stanford.edu/class/stats311/Lectures/full_notes.pdf*, vol. 2, p. 23, 2016.

[20] J. Bretagnolle and C. Huber, "Estimation des densités: risque minimax," *Séminaire de probabilités de Strasbourg*, vol. 12, pp. 342–363, 1978.

[21] N. L. Johnson, A. W. Kemp, and S. Kotz, *Univariate discrete distributions*. John Wiley & Sons, 2005, vol. 444.

APPENDIX A
INDISTINGUISHABILITY BY TEST

Lemma 1. *Assume that $0 < \epsilon_0 < \epsilon < 1$ are given. If $p, q \in [0,1]$ are such that $\epsilon_0 \leq p \leq q \leq \epsilon$, then,*

$$D_{\text{KL}}(\text{Ber}(p) \otimes \text{Ber}(q) \| \text{Ber}(q) \otimes \text{Ber}(p)) \leq (\epsilon - \epsilon_0) \log\left(1 + \frac{\epsilon - \epsilon_0}{\epsilon_0(1-\epsilon)}\right). \quad (6)$$

Proof. For ease of notation, let D denote the divergence on the LHS of (6). It could be verified that

$$D = (q-p)\log\left(1 + \frac{q-p}{p\bar{q}}\right).$$

From this equality and the assumption $\epsilon_0 \leq p \leq q \leq \epsilon$, we can verify that

$$D \leq (\epsilon - \epsilon_0)\log\left(1 + \frac{\epsilon - \epsilon_0}{\epsilon_0(1-\epsilon)}\right),$$

as required. □

Lemma 2. *Assume that $0 < \epsilon_0 < \epsilon < 1$ are given. If $p, q \in [0,1]$ are such that $\epsilon_0 \leq p \leq q \leq \epsilon$, then,*

$$H^2(\text{Ber}(p) \otimes \text{Ber}(q) \| \text{Ber}(q) \otimes \text{Ber}(p)) \geq \frac{(p-q)^2}{4\epsilon(1-\epsilon_0)}. \quad (7)$$

Proof. From the definition of the Hellinger Distance, it follows immediately that

$$H^2(\text{Ber}(p) \otimes \text{Ber}(q) \| \text{Ber}(q) \otimes \text{Ber}(p)) = \left(\sqrt{p\bar{q}} - \sqrt{\bar{p}q}\right)^2 = \frac{(p-q)^2}{\left(\sqrt{p\bar{q}} + \sqrt{\bar{p}q}\right)^2}.$$

Invoking that $\epsilon_0 \leq p \leq q \leq \epsilon$ completes the proof. □

A. Proof of Proposition 1

Proof. From [19, Prop. 2.17], we have

$$\inf_{\psi} P_{\text{error}}(\psi) = \frac{1}{2}\inf_{\psi} \Pr(\psi(\mathbf{Z}) = 0)|H_1) + \Pr(\psi(\mathbf{Z}) = 1)|H_0) = \frac{1}{2}(1 - \text{TV}[\Pr(\cdot \,|\, H_0), \Pr(\cdot \,|\, H_1)]). \quad (8)$$

Let $p \triangleq \min\{L(h_1), L(h_2)\}$ and $q \triangleq \max\{L(h_1), L(h_2)\}$. Given that $\mathbf{Z} \sim \bigotimes_{i=1}^{n}[\text{Bern}(L(h_1)) \otimes \text{Bern}(L(h_2))]$, then

$$\Pr(\cdot \,|\, H_0) = \bigotimes_{i=1}^{n}[\text{Bern}(p) \otimes \text{Bern}(q)],$$

$$\Pr(\cdot \,|\, H_1) = \bigotimes_{i=1}^{n}[\text{Bern}(q) \otimes \text{Bern}(p)].$$

Recall that $\frac{1}{2}H^2(P,Q) \leq \text{TV}(P,Q)$ [11, Eq. 6.9] and also that $\text{TV}(P,Q) \leq \sqrt{1 - e^{-D_{\text{KL}}(P\|Q)}}$ [20]. The proposition follows from these inequalities and the tensorization properties of the KL-divergence and the Hellinger distance [11, Ch. 2 & 6] together with Lemmas 1-2. □

B. Proof of Theorem 1

Proof. Define $\epsilon_0 \triangleq \min_{h \in \mathcal{H}} L(h)$. If $h_1, h_2 \in \mathcal{R}(\epsilon)$, then $h_1, h_2 \in [\epsilon_0, \epsilon]$. If $(\epsilon - \epsilon_0)\log\left(1 + \frac{\epsilon - \epsilon_0}{\epsilon_0(1-\epsilon)}\right) \leq \frac{1}{n}\log\left(\frac{1}{1-(1-2\alpha)^2}\right)$, because of Proposition 1, $\inf_{\psi} P_{\text{error}}(\psi) \geq \alpha$. On the other hand, let $h_*, h \in \mathcal{R}(\epsilon)$ be classifiers such that $L(h_*) = \epsilon_0$ and $L(h^*) = \epsilon$. If $\frac{(\epsilon-\epsilon_0)^2}{(1-\epsilon_0)\epsilon} \geq 8\left(1 - [2\beta]^{1/n}\right)$, then h_* and h^* are β-testing distinguishable due to Proposition 1. □

APPENDIX B
INDISTINGUISHABILITY BY CONFIDENCE

A. Proof of Proposition 2

Let $h_1, h_2 \in \mathcal{H}$. Recall that Hoeffding's Inequality [15, Prop. 2.7] ensures that, for every $i \in \{1,2\}$,

$$\Pr\left(|L(h_i) - \hat{L}(h_i)| \leq \sqrt{\frac{\log 2/\alpha}{2n}}\right) \geq 1 - \alpha. \quad (9)$$

Therefore, $I_i = \left\{x : |x - \hat{L}(h_i)| \leq \sqrt{\frac{\log 2/\alpha}{2n}}\right\}$ is the interval of confidence level at least $1-\alpha$ that the Hoeffding Inequality method creates for $L(h_i)$. If $|\hat{L}(h_1) - \hat{L}(h_2)| \leq 2\sqrt{\frac{\log(2/\alpha)}{2n}}$, then

$$\min_{i \in \{1,2\}} \hat{L}(h_i) + \sqrt{\frac{\log 2/\alpha}{2n}} \geq \max_{i \in \{1,2\}} \hat{L}(h_i) - \sqrt{\frac{\log 2/\alpha}{2n}}. \quad (10)$$

Thus, $I_1 \cap I_2 \neq \varnothing$. On the other hand, if $|\hat{L}(h_1) - \hat{L}(h_2)| > 2\sqrt{\frac{\log(2/\alpha)}{2n}}$, the opposite of equation 10 holds; i.e. $I_1 \cap I_2 = \varnothing$.

B. Lemma 3

Lemma 3. *The functions $U_\delta(x)$ and $L_\delta(x)$ are increasing in $x \in \mathbb{Z}^+$ for every $\delta \in (0,1)$.*

Proof. Due to the Binomial-Beta relation [21, Eq. 3.18]:

$$\mathbb{P}(\text{Bin}(n,p) \geq k) = I_p(k, n-k+1), \quad (11)$$

$L_\delta(x) = \textit{qBeta}(\delta/2; x, n-x+1)$ and $U_\delta(x) = \textit{qBeta}(1-\delta/2; x+1, n-x)$. Let $\delta \in (0,1)$ and $x_1 < x_2$ be positive integers. Define $q_1 \triangleq \textit{qBeta}(\delta/2; x_1, n-x_1+1)$ and $q_2 \triangleq \textit{qBeta}(\delta/2; x_2, n-x_2+1)$. Suppose that $q_1 \geq q_2$, since the regularized beta function is increasing,

$$I_{q_1}(x_1, n-x_1+1) \geq I_{q_2}(x_1, n-x_1+1). \quad (12)$$

Additionally, given that $x_1 < x_2$,

$$\mathbb{P}(\text{Bin}(n,q_2) \geq x_1) > \mathbb{P}(\text{Bin}(n,q_2) \geq x_2). \quad (13)$$

From (12)-(13), the Beta-Binomial relation (11) and the definition of q_1, q_2 it follows that

$$\delta/2 = I_{q_1}(x_1, n-x_1+1) > I_{q_2}(x_2, n-x_2+1) = \delta/2$$

which is clearly absurd. Therefore, $q_1 < q_2$. Under similar arguments, $U_\delta(x)$ is increasing in $x \in \mathbb{Z}^+$. □

C. Proof of Proposition 3

Suppose that h_1 and h_2 are δ-testing indistinguishable. Then $I_\delta(\hat{L}(h_1)) \cap I_\delta(\hat{L}(h_2)) \neq \varnothing$. In light of Lemma 3, this means that

$$L_\delta(n \max\{\hat{L}(h_1), \hat{L}(h_2)\}) \leq U_\delta(n \min\{\hat{L}(h_1), \hat{L}(h_2)\}).$$

Thus,

$$\begin{aligned} &\textit{qBeta}(\delta/2; n \max \hat{L}(h_i), n - n \max \hat{L}(h_i) + 1) \\ \leq &\textit{qBeta}(1 - \delta/2; n \min \hat{L}(h_i) + 1, n - n \min \hat{L}(h_i)). \end{aligned}$$

On the other hand, if

$$\begin{aligned} &\textit{qBeta}(\delta/2; n \max \hat{L}(h_i), n - n \max \hat{L}(h_i) + 1) \\ > &\textit{qBeta}(1 - \delta/2; n \min \hat{L}(h_i) + 1, n - n \min \hat{L}(h_i)), \end{aligned}$$

then

$$L_\delta(n \max\{\hat{L}(h_1), \hat{L}(h_2)\}) > U_\delta(n \min\{\hat{L}(h_1), \hat{L}(h_2)\}).$$

As a consequence,

$$I_\delta(\hat{L}(h_1)) \cap I_\delta(\hat{L}(h_2)) = \varnothing.$$

D. Proof of Theorem 2

Let $h_1, h_2 \in \hat{\mathcal{R}}(\epsilon)$. Notice that

$$|\hat{L}(h_1) - \hat{L}(h_2)| \leq \epsilon - \epsilon_0.$$

If $\epsilon \leq 2\sqrt{\frac{\log(2/\alpha)}{2n}}$, then, from Proposition 2, it follows that h_1 and h_2 are α-indistinguishable under the Hoeffding Inequality method. Now, assume that $\epsilon - \epsilon_0 > 2\sqrt{\frac{\log(2/\alpha)}{2n}}$. The hypotheses that achieve empirical risk ϵ and ϵ_0 respectively are not α-indistinguishable.

E. Proof of Theorem 3

Assume that $L_\alpha(n\epsilon) \leq U_\alpha(n\epsilon_0)$. For any $h_1, h_2 \in \hat{\mathcal{R}}(\epsilon)$, it follows that

$$L_\alpha(n \max \hat{L}(h_i)) \leq L_\alpha(n\epsilon), \tag{14}$$

$$U_\alpha(n\hat{L}(h^*)) \leq U_\alpha(n \min \hat{L}(h_i)). \tag{15}$$

Therefore, $L_\alpha(n \max \hat{L}(h_i)) \leq U_\alpha(n \min \hat{L}(h_i))$; i.e. h_1 and h_2 are α-indistinguishable under the CP method. On the other hand, suppose $L_\alpha(n\epsilon) > U_\alpha(n\epsilon_0)$. Note that the hypotheses that achieve empirical risk ϵ and ϵ_0 respectively are not α-indistinguishable.