Rashomon Capacity: A Metric for Predictive Multiplicity in Probabilistic Classification

Anonymous Author(s)

Abstract

Predictive multiplicity occurs when classification models with nearly indistinguishable average performances assign conflicting predictions to individual samples. When used for decision-making in applications of consequence (e.g., lending, education, criminal justice), models developed without regard for predictive multiplicity may result in unjustified and arbitrary decisions for specific individuals. We introduce a new measure of predictive multiplicity in probabilistic classification called Rashomon Capacity. Prior metrics for predictive multiplicity focus on classifiers that output thresholded (i.e., 0-1) predicted classes. In contrast, Rashomon Capacity applies to probabilistic classifiers, capturing more nuanced score variations for individual samples. We provide a rigorous derivation for Rashomon Capacity, argue its intuitive appeal, and demonstrate how to estimate it in practice. We show that Rashomon Capacity yields principled strategies for disclosing conflicting models to stakeholders. Our numerical experiments illustrate how Rashomon Capacity captures predictive multiplicity in various datasets and learning models, including neural networks. The tools introduced in this paper can help data scientists measure, report, and ultimately resolve predictive multiplicity prior to model deployment.

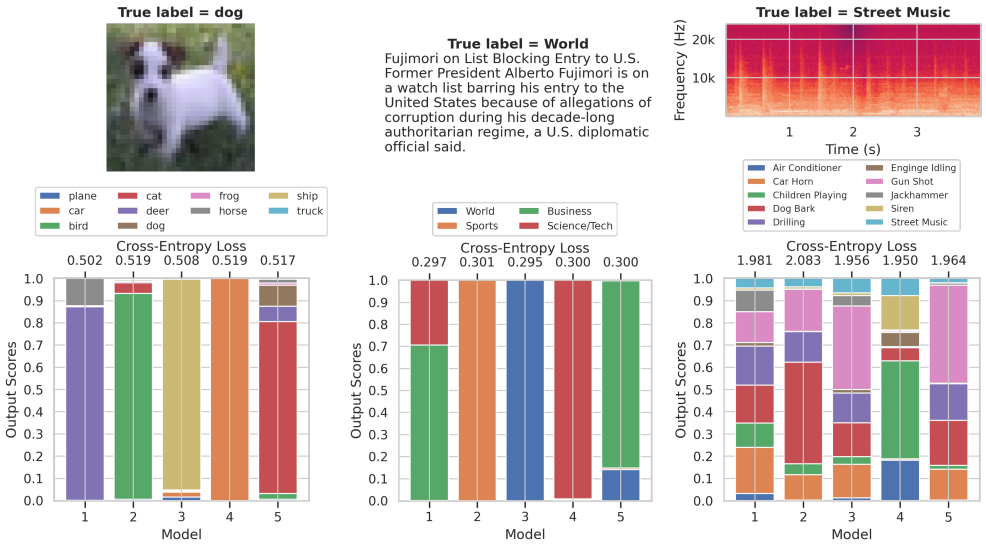
1 Introduction

Rashomon effect, introduced by Breiman [1], describes the phenomenon where a multitude of distinct predictive models achieve similar training or test loss. Breiman reported observing the Rashomon effect in several model classes, including linear regression, decision trees, and small neural networks. In a foresighted experiment, Breiman noted that, when retraining a neural network 100 times on three-dimensional data with different random initializations, he "*found 32 distinct minima, each of which gave a different picture, and having about equal test set error*" [1, Section 8]. The set of almost-equally performing models for a given learning problem is called the *Rashomon set* [2, 3].

The Rashomon effect has important—and potentially harmful—implications. We focus on a facet of the Rashomon effect in classification problems called *predictive multiplicity*. Predictive multiplicity occurs when competing models in the Rashomon set assign conflicting predictions to individual samples [4]. Fig. 1 presents an updated version of Breiman's neural network experiment and illustrates predictive multiplicity in three classification tasks with different data domains and neural network architectures. Here, models that achieve statistically-indistinguishable performance on a test set assign wildly different predictions to an input sample. If predictive multiplicity is not accounted for, the output for this sample may ultimately depend on arbitrary choices made during training (e.g., parameter initialization).

Predictive multiplicity captures the potential individual-level harm introduced by an arbitrary choice of a single model in the Rashomon set. When such a model is used to support automated decision-

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.



(a) CIFAR-10 dataset **(b)** AG News dataset **(c)** UrbanSound8k dataset.

Figure 1: The scores (bottom) of a sample (top) generated by competing models. Predictive multiplicity occurs on different data domains and learning models, including an image dataset (CIFAR-10 [8]) trained with VGG16 [9], a natural language dataset (AG News [10]) trained with a simple neural networks after tokenization, and an audio dataset (UrbanSound8k [11]) trained with LSTM [12].

making in sectors dominated by a few companies or Government—labeled *Algorithmic Leviathans* in [5, Section 3]—predictive multiplicity can lead to unjustified and systemic exclusion of individuals from critical opportunities. For example, an algorithm used for lending may deny a loan to a specific applicant. However, during model development, there may have been a competing model which performs equally well on average, yet would have approved the loan for this individual. As another example, Governments are increasingly turning to algorithms for grading exams that grant access to higher-level education (see, e.g., UK [6] and Brazil [7]). Here, again, accounting for predictive multiplicity is critical: an arbitrary choice of a single model in the Rashomon set may lead to an unwarranted restriction of educational opportunities to an individual student. In applications such as criminal justice and healthcare, models that do not account for predictive multiplicity are also at risk of causing arbitrary individual-level harm.

We introduce new methods for measuring, reporting, and resolving predictive multiplicity for probabilistic classifiers[1]. First, we postulate several properties that a predictive multiplicity metric must satisfy to simplify its interpretation by stakeholders. We then provide a new predictive multiplicity metric called *Rashomon Capacity*. Unlike prior metrics based on thresholded (i.e., 0 or 1) predictions, Rashomon Capacity captures more nuanced variations in scores among models in the Rashomon set for a target input sample. We derive properties of Rashomon Capacity and describe how it can be computed. Our numerical examples provide a methodology for reporting predictive multiplicity in probabilistic classifiers trained on a variety of datasets.

Remarkably, the computation of Rashomon Capacity also sheds light on a strategy for resolving predictive multiplicity. Different approaches have been proposed for dealing with multiplicity, including randomizing between competing classifiers [5] and bagging [1]. However, the size of the Rashomon set can be large (see Semenova et al. [3] for an example with 250k models), making strategies that require randomizing predictions across the entire Rashomon set potentially arbitrary and impractical. In contrast, we apply standard results from convex analysis to show that the predictive multiplicity for an input sample—as measured by Rashomon Capacity—can be entirely captured

[1] A probabilistic classifier is a model that maps an input sample onto a probability distribution, referred as a score, over a discrete set of classes. Examples of probabilistic classifiers include logistic regression, random forests, and a neural network with a softmax output layer.

by at most c models in the Rashomon set, where c is the number of predicted classes. This results holds *regardless of the size of the Rashomon set*. Thus, when c is small, the predictions produced by the competing classifiers can be communicated to a stakeholder, empowering them to decide how to resolve conflicting scores. Note that the c models that capture predictive multiplicity can be different across samples. Consequently, we also provide a greedy algorithm for identifying a subset of models in the Rashomon set that captures most of the score variations across a dataset. In summary, our main contributions are:

1. We introduce a new score-based metric for *quantifying* predictive multiplicty in probabilistic classifiers called Rashomon Capacity. We provide a principled derivation of this metric and demonstrate how it can be computed in practice.

2. We describe a methodology for *reporting* predictive multiplicity in probabilistic classifiers using Rashomon Capacity. We give examples of this methodology using different datasets and learning models. Rashomon Capacity must be reported to stakeholders in, for example, model cards [13].

3. We propose a procedure for *resolving* predictive multiplicity in probabilistic classifiers. Even though the Rashomon set may span a large (potentially uncountable) number of models, we show that the score variation for a sample is fully captured by a small, discrete subset of models in the Rashomon set. Communicating these predictions to stakeholders can empower them to decide how to resolve predictive multiplicity.

2 Background and related work

The Rashomon effect impacts model selection [14–16], explainability [17], and fairness [18]. Rudin et al. [14] suggested that, given the choice of competing models, machine learning (ML) practitioners should select interpretable models *a priori*, rather than selecting a black-box model with conjectural explanations afterwards. Hancox-Li [15] and D'Amour et al. [16] further argued that epistemic patterns, e.g., causality, should be specified in the ML pipeline, and the selected models from the Rashomon set should be able to reflect these patterns. Competing models in the Rashomon set may not only render conflicting explanations for predictions [17] and measures of variables' importance [2], but also have inconsistent performance across population sub-groups. Consequently, the arbitrary choice of a single model without regard for the Rashomon effect may result in unnecessary and discriminatory bias against vulnerable population groups [18]. See the Supplementary Materials (SM) for further discussion, including connections with individual fairness [19].

We outline the notation used in the paper next, then introduce existing metrics for predictive multiplicity and discuss their limitations. We consider a dataset $\mathcal{D} = \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^n$, e.g., a training or test set, for a classification task with c classes/labels, where each sample pair $(\mathbf{x}_i, \mathbf{y}_i)$ is drawn i.i.d. from $P_{X,Y}$ with support $\mathcal{X} \times \Delta_c$. Here, Δ_c denotes the c-dimensional probability simplex, i.e., each $\mathbf{y}_i$ is one-hot encoded. Let $\mathbf{e}_k$ be a length-c indicator vector with one in the kth position and zero elsewhere, i.e., $[\mathbf{e}_k]_k = 1$, and $[\mathbf{e}_k]_j = 0 \; \forall j \neq k$, where $[\cdot]_j$ denotes the jth entry of a vector. Thus, we have $\mathbf{y}_i \in \{\mathbf{e}_k\}_{k=1}^c$, and $\Delta_c = \mathsf{span}(\mathbf{e}_1, \mathbf{e}_2, \cdots, \mathbf{e}_c)$. $\mathbb{1}(\cdot)$ denotes the indicator function.

We denote by $\mathcal{H}$ a *hypothesis space*, i.e., a set of candidate probabilistic classifier is parameterized by $\theta \in \Theta \subseteq \mathbb{R}^d$ that approximate $P_{Y|X=\mathbf{x}_i}$, i.e., $\mathcal{H} \triangleq \{h_\theta : \mathcal{X} \to \Delta_c : \theta \in \Theta\}$. The loss function used to evaluate model performance is denoted by $\ell : \Delta_c \times \Delta_c \to \mathbb{R}^+$ (e.g., cross-entropy) and $L(\theta) \triangleq \mathbb{E}_{P_{X,Y}}[\ell(h_\theta(X), Y)]$ the population risk[2]. As usual, the population risk is approximated by the empirical risk $\hat{L}(\theta) \triangleq \frac{1}{n}\sum_{i=1}^n \ell(h_\theta(\mathbf{x}_i), \mathbf{y}_i)$. The definitions in the rest of this paper are cast in terms of empirical risk, and their population risk versions can be defined similarly.

Rashomon set, Rashomon ratio, and pattern Rashomon ratio. We start with the definition of the Rashomon set. An intuitive characterization is the set of all models in the hypothesis space that yield similar average loss. Given a Rashomon parameter $\epsilon \geq 0$, Semenova et al. [3] and Marx et

[2]We use the notations $L(h) \triangleq \mathbb{E}_{P_{X,Y}}[\ell(h(X), Y)]$ and $L(\theta) \triangleq \mathbb{E}_{P_{X,Y}}[\ell(h_\theta(X), Y)]$ interchangeably, depending on whether we intend to emphasize the parameter space Θ.

al. [4] respectively define a Rashomon set[3] and an ϵ-level set as $\mathcal{R}(\mathcal{H}, \epsilon) \triangleq \{h \in \mathcal{H}; \hat{L}(h) \leq \epsilon\}$. Note that the Rashomon set is determined by the hypothesis space and the Rashomon parameter, and also implicitly by the dataset due to the evaluation of $\hat{L}(h)$. The cardinality $|\mathcal{R}(\mathcal{H}, \epsilon)|$ or the volume[4] $\mathsf{vol}(\mathcal{R}(\mathcal{H}, \epsilon))$ of the Rashomon set (depending on whether $\mathcal{R}(\mathcal{H}, \epsilon)$ is finite) can then be used to quantify the size of the Rashomon set.

Estimating $\mathsf{vol}(\mathcal{R}(\mathcal{H}, \epsilon))$ is essentially a level set estimation problem [20], and is computationally infeasible when the hypothesis space $\mathcal{H}$ is large [21]. Semenova et al. [3, Section 5.1] provide an exact computation of $\mathsf{vol}(\mathcal{R}(\mathcal{H}, \epsilon))$ on ridge regression (where the Rashomon set forms an ellipsoid), and suggest rejection-sampling-with-replacement and importance sampling of the Rashomon set when an exact form is not feasible [3, Section 5.3]. However, there is no sample complexity guarantee for the sampling methods provided, and a significant amount of models might be needed to estimate the Rashomon volume. For example, in [3, Section 6], 250k decision trees are sampled in order to compute $\mathsf{vol}(\mathcal{R}(\mathcal{H}, \epsilon))$. Given $\mathcal{R}(\mathcal{H}, \epsilon)$, the *Rashomon ratio* is defined as $\hat{\mathcal{R}}(\mathcal{H}, \epsilon) \triangleq \frac{\mathsf{vol}(\mathcal{R}(\mathcal{H}, \epsilon))}{\mathsf{vol}(\mathcal{H})}$. $\hat{\mathcal{R}}(\mathcal{H}, \epsilon)$ represents the fraction of models in the hypothesis space that fit the data about equally well. Semenova et al. [3] suggested that when the Rashomon ratio is large, models with various desirable properties, such as better generalizability, can often exist inside the Rashomon set.

For predictive multiplicity in classification problems, Semenova et al. [3, Defn. 12] further proposed *pattern Rashomon ratio* for binary classification, which is the ratio of the count of all possible binary predicted classes given by the functions in the Rashomon set to that given by the functions in the hypothesis space. When the size of the hypothesis space goes to infinity, the pattern Rashomon ratio can be upper and lower bounded by the binary entropy function [3, Proposition 13]. Note that the computational complexity of the pattern Rashomon ratio grows exponentially with the number of samples, and could be an "expensive" metric for predictive multiplicity when applied on a large dataset. Moreover, even for a simplest logistic regression, the exact form of the pattern Rashomon ratio is not tractable due to the non-linearity of the maximum likelihood ratio [22].

Ambiguity and discrepancy of models in a Rashomon set. In an important paper (and the work closest to ours), Marx et al. [4] motivated predictive multiplicity and proposed two metrics, *ambiguity* and *discrepancy*, for quantifying predictive multiplicity in binary classification problems[5]. Marx et al. [4] measure multiplicity in terms of the thresholded outputs (i.e., predicted classes) of a classifier, defining the Rashomon set (or ϵ-level set in their terminology) in terms of classification accuracy. In contrast, the metrics proposed here are defined in terms of the raw outputs in Δ_c of a probabilistic classifier. As we shall see, this allows for a more nuanced characterization of multiplicity in probabilistic classification. We briefly overview the metrics by Marx et al. [4] next.

Ambiguity is the proportion of samples in a dataset that can be assigned conflicting predictions by competing classifiers in the Rashomon set. Discrepancy is the maximum number of predictions that could change in a dataset if we were to switch between models within the Rashomon set. More precisely, Given a model $\hat{h}$, the ambiguity $\alpha_\epsilon(\hat{h})$ and the discrepancy $\delta_\epsilon(\hat{h})$ are respectively defined as

$$\begin{aligned} \alpha_\epsilon(\hat{h}) &= \frac{1}{n} \sum_{i=1}^{n} \max_{h \in \mathcal{R}(\mathcal{H}, \epsilon)} \mathbb{1}\left[\arg\max h(\mathbf{x}_i) \neq \arg\max \hat{h}(\mathbf{x}_i)\right], \\ \delta_\epsilon(\hat{h}) &= \max_{h \in \mathcal{R}(\mathcal{H}, \epsilon)} \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}\left[\arg\max h(\mathbf{x}_i) \neq \arg\max \hat{h}(\mathbf{x}_i)\right]. \end{aligned} \tag{1}$$

Both quantities capture predictive multiplicity in terms of thresholded (i.e., 0 or 1) outputs of a classifier. For linear predictive models, the quantities in (1) can be estimated by mixed integer programming tools [4].

[3] When given an empirical risk minimizer $\hat{\theta} \in \arg\min_{\theta \in \Theta} \hat{L}(h_\theta)$, one can set $\epsilon = \hat{L}(h_{\hat{\theta}}) + \epsilon'$.

[4] We can define a monotonic function $\mathsf{vol} : \mathcal{H} \to \mathbb{R}^+$ for the volume of $\mathcal{R}(\mathcal{H}, \epsilon)$ in the hypothesis space. When $\mathcal{H}$ is parameterized by $\Theta \subseteq \mathbb{R}^d$, the volume of $\{\theta \in \Theta; h_\theta \in \mathcal{R}(\mathcal{H}, \epsilon)\}$ can be directly computed in $\mathbb{R}^d$; however, computing the volume of a set in high-dimensional space is hard even if its analytic form is known.

[5] The definitions in Marx et al. [4] can be directly extended to multi-class classification.

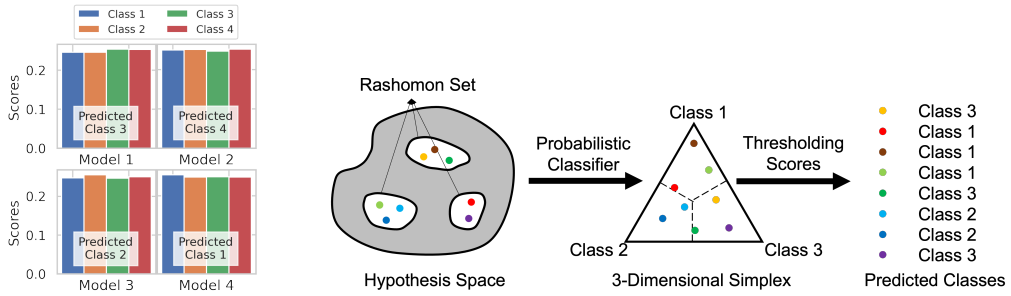


Figure 2: Left: 4 models output similar scores for 4 classes, but the predicted classes produced via thresholding are very different. This may lead to high ambiguity and/or discrepancy (cf. (1)). **Right**: Given a sample, the scores obtained from different models (colored dots) in the Rashomon set lead to different predicted classes by thresholding the scores (based on decision boundaries, i.e., dashed lines in the simplex). Prior work either measure multiplicity on the hypothesis set, e.g., the Rashomon ratio, or in terms of predicted classes, e.g., pattern Rashomon ratio and ambiguity/discrepancy. In this work, we measure predictive multiplicity directly in terms of scores on the probability simplex.

The above metrics for predictive multiplicity are based on predicted classes. Thus, they require finding the $\arg\max$ or thresholding the scores at the output of a classifier. In probabilistic classification, thresholding may mask similar predictions produced by competing models and artificially increase multiplicity: output scores can be almost equal across different classes, yet the (thresholded) predicted classes can be very different. For example, two scores $[0.49, 0.51]$ and $[0.51, 0.49]$ for a binary classification problem can lead to entirely different predicted classes—1 and 0, respectively—and ultimately overestimate predictive multiplicity (see Fig. 2 (Left) for another multi-class example). In fact, predictive multiplicity metrics based on predicted classes may yield multiplicity *even for a single fixed model* when, for example, the threshold criteria for output scores is changed[6]. This subtle, yet important difference motivates us to reconsider existing metrics and introduce a new predictive multiplicity metric for probabilistic classifiers that output scores (cf. Fig. 2 (Right) for an overview).

3 Measuring predictive multiplicity of probabilistic classifiers

We outline below desirable properties of predictive multiplicity metrics for probabilistic classifiers. These properties motivate our definition of Rashomon Capacity and provide guidelines for the creation of new multiplicity metrics in future research. Motivated by the potential individual-level harm incurred by an arbitrary choice of model in the Rashomon set, we focus on of per-sample multiplicity metrics. Next, we formally define Rashomon Capacity in terms of the KL-divergence between the output scores of classifiers in the Rashomon set. We then use Rashomon Capacity to define a predictive multiplicity metric for individual samples in a dataset (similar to ambiguity in (1)). Omitted proofs are in the SM, along with more information on Rashomon Capacity.

3.1 Properties of multiplicity metrics for probabilistic classifiers

Consider a dataset $\mathcal{D}$ and a corresponding Rashomon set $\mathcal{R}(\mathcal{H}, \epsilon)$ for a classification problem with c classes. We collect all possible output scores for a sample $\mathbf{x}_i \in \mathcal{D}$ and define the *ϵ-multiplicity set* as

$$\mathcal{M}_\epsilon(\mathbf{x}_i) \triangleq \{h(\mathbf{x}_i) \mid h \in \mathcal{R}(\mathcal{H}, \epsilon)\} \subseteq \Delta_c. \tag{2}$$

Let $m(\cdot)$ be a measure of predictive multiplicity, and $m(\mathcal{M}_\epsilon(\mathbf{x}_i))$ be the predictive multiplicity of sample $\mathbf{x}_i$. Which properties should $m(\cdot)$ have? Ideally, we expect $m(\mathcal{M}_\epsilon(\mathbf{x}_i))$ to be a bounded value in $[1, c]$, since at least one class is assigned to sample $\mathbf{x}_i$, and at most c different classes could be assigned to $\mathbf{x}_i$. Moreover, if $m(\mathcal{M}_\epsilon(\mathbf{x}_i)) = 1$ (i.e., a predictive multiplicity of 1), then one would

[6]That is not to say that multiplicity of predicted classes should not be reported. Ambiguity and discrepancy in (1) are important metrics for stakeholders.

expect that only one score is produced for $\mathbf{x}_i$ and, thus, all predictions in $\mathcal{M}_\epsilon(\mathbf{x}_i)$ are exactly the same. Similarly, if $m(\mathcal{M}_\epsilon(\mathbf{x}_i)) = c$ (i.e., predictive multiplicity equal to the number of classes), then there must exist c models $\{h_1, \cdots, h_c\} \subseteq \mathcal{R}(\mathcal{H}, \epsilon)$ such that $h_j(\mathbf{x}_i) = \mathbf{e}_j$. In other words, each of the c classes can be assigned to the sample, yielding a predictive multiplicity of c. Finally, $m(\mathcal{M}_\epsilon(\mathbf{x}_i))$ should be monotonic in $\mathcal{M}_\epsilon(\mathbf{x}_i)$, i.e., if $\mathcal{M}_\epsilon(\mathbf{x}_i) \subseteq \mathcal{M}'_\epsilon(\mathbf{x}_i)$, then $m(\mathcal{M}_\epsilon(\mathbf{x}_i)) \leq m(\mathcal{M}'_\epsilon(\mathbf{x}_i))$. We summarize these desirable properties of predictive multiplicity metrics in the following definition.

Definition 1. *Let $\mathcal{S}_c \triangleq \{\{\mathbf{y}\} \mid \mathbf{y} \in \Delta_c\}$ be the set of singleton sets in Δ_c, and $\sigma(\mathcal{S}_c)$ its corresponding σ-algebra*[7]*. We say that a function $m : \sigma(\mathcal{S}_c) \to \mathbb{R}$ is a* predictive multiplicity metric[8] *if for any $\mathcal{A}, \mathcal{B} \in \sigma(\mathcal{S}_c)$*

1. $1 \leq m(\mathcal{A}) \leq c$;
2. $m(\mathcal{A}) = 1$ *if and only if* $|\mathcal{A}| \leq 1$;
3. $m(\mathcal{A}) = c$ *if and only if* $\mathbf{e}_k \in \mathcal{A}$ *for* $k \in [c]$*, i.e., $\mathcal{A}$ contains the corner points of Δ_c;*
4. $m(\mathcal{A}) \leq m(\mathcal{B})$ *if* $\mathcal{A} \subseteq \mathcal{B}$.

We introduce next a predictive multiplicity metric called Rashomon Capacity that satisfies all properties above. In the rest of the paper, when the ϵ-multiplicity set $\mathcal{M}_\epsilon(\cdot)$ is clear from context, we use $m(\mathbf{x}_i)$ as shorthand for $m(\mathcal{M}_\epsilon(\mathbf{x}_i))$.

3.2 Rashomon Capacity

Our goal is to quantify predictive multiplicity by the score difference assigned to each point $\mathbf{x}_i$ in $\mathcal{D}$. Note that each point in $\mathcal{M}_\epsilon(\mathbf{x}_i)$ is a probability distribution over c classes. Thus, it is natural to adopt divergence measures for distributions to capture the "variation" of scores in $\mathcal{M}_\epsilon(\mathbf{x}_i)$. From a geometric viewpoint, a larger spread in scores indicates a greater amount of predictive multiplicity for a sample.

Assume a probability (or "weight") distribution P_M across models in $\mathcal{R}(\mathcal{H}, \epsilon)$ (and therefore each score in $\mathcal{M}_\epsilon(\mathbf{x}_i)$), where M denotes the random variable of selecting/sampling the models in the Rashomon set. Intuitively, if P_M assigns mass 1 to a single model and 0 to all other models in the Rashomon set, then the output of only one model is considered. Conversely, if P_M is the uniform distribution, then the outputs of every model in the set are equally weighed. Given a divergence measure between distributions $d(\cdot\|\cdot)$, we quantify the spread of the scores in $\mathcal{M}_\epsilon(\mathbf{x}_i)$ by

$$\rho(\mathcal{M}_\epsilon(\mathbf{x}_i), P_M) \triangleq \inf_{\mathbf{q} \in \Delta_c} \mathbb{E}_{h \sim P_M} d(h(\mathbf{x}_i)\|\mathbf{q}). \tag{3}$$

Here, the minimizing $\mathbf{q}$ acts as a "center of gravity" or "centroid" for the outputs of the classifiers in the Rashomon set for a chosen distribution P_M across models. Analogously, the quantity $\rho(\mathcal{M}_\epsilon(\mathbf{x}_i), P_M)$ can be understood as a measure of "spread" or "inertia" across model outputs. We select the distribution P_M that results in the largest spread in scores:

$$C_d(\mathcal{M}_\epsilon(\mathbf{x}_i)) \triangleq \sup_{P_M} \rho(\mathcal{M}_\epsilon(\mathbf{x}_i), P_M) = \sup_{P_M} \inf_{\mathbf{q} \in \Delta_c} \mathbb{E}_{h \sim P_M} d(h(\mathbf{x}_i)\|\mathbf{q}). \tag{4}$$

The missing element is the choice of divergence measure $d(\cdot\|\cdot)$. A natural candidate is cross-entropy (or log-loss) $d(h(\mathbf{x}_i)\|\mathbf{q}) = -h(\mathbf{x}_i)^\top \log \mathbf{q}$, since this is the standard loss used for training and evaluating probabilistic classifiers. Alas, the minimal cross-entropy $\min_{\mathbf{q}} -h(\mathbf{x}_i)^\top \log \mathbf{q} = h(\mathbf{x}_i)^\top \log h(\mathbf{x}_i)$ is not 0 and depends on $h(\mathbf{x}_i)$. Consequently, if one were to choose $d(\cdot\|\cdot)$ to be cross-entropy, the minimum value of (4) would not be consistent and would depend on $\mathbf{x}_i$ — even when the outputs across all models in the Rashomon set match! Thus, we shift cross-entropy by its minimum $-h(\mathbf{x}_i)^\top \log h(\mathbf{x}_i)$. This results in KL-divergence as our divergence measure of choice: $D_{KL}(h(\mathbf{x}_i)\|\mathbf{q}) = -h(\mathbf{x}_i)^\top \log \mathbf{q} + h(\mathbf{x}_i)^\top \log h(\mathbf{x}_i)$. We name the spread in scores measured using KL-divergence as *Rashomon Capacity*[9].

[7] In measure-theoretic notation, this corresponds to the set of all countable sets in Δ_c.

[8] We use the term "metric" loosely here, not in the usual sense of distance between two elements of a set.

[9] We consider logarithms in base 2, and the unit of Rashomon Capacity is bit.

Definition 2. *Given a sample $\mathbf{x}_i$ and the corresponding ϵ-multiplicity set $\mathcal{M}_\epsilon(\mathbf{x}_i)$, the Rashomon Capacity is defined as*

$$C(\mathcal{M}_\epsilon(\mathbf{x}_i)) \triangleq \sup_{P_M} \inf_{\mathbf{q}\in\Delta_c} \mathbb{E}_{h\sim P_M} D_{KL}(h(\mathbf{x}_i)\|\mathbf{q}). \tag{5}$$

Moreover, we define $m_C(\mathbf{x}_i) \triangleq 2^{C(\mathcal{M}_\epsilon(\mathbf{x}_i))}$.

The quantity $C(\mathcal{M}_\epsilon(\mathbf{x}_i))$ is ubiquitous in information theory; in fact, $C(\mathcal{M}_\epsilon(\mathbf{x}_i))$ is the *channel capacity* [23] of a channel $P_{Y|M}$ whose rows are the entries of $\mathcal{M}_\epsilon(\mathbf{x}_i)$. This connection motivates the name "Rashomon Capacity" and is useful for proving that $m_C(\mathbf{x}_i)$ is indeed a predictive multiplicity metric, as stated in the next proposition.

Proposition 1. *The function $m_C(\cdot) = 2^{C(\mathcal{M}_\epsilon(\cdot))} : \mathcal{X} \to [1, c]$ satisfies all properties of a predictive multiplicity metric in Definition 1.*

The Rashomon set is fundamentally different from Rashomon Capacity, in the sense that a larger Rashomon set does not necessarily lead to a larger Rashomon Capacity. Using a binary classification problem as an example, consider two Rashomon sets with scores

$$\begin{aligned}&\mathcal{R}_1(\mathcal{H},\epsilon) = \{h_1, h_2, h_3\},\ h_1(\mathbf{x}_i) = [0.45, 0.55], h_2(\mathbf{x}_i) = [0.50, 0.50], h_3(\mathbf{x}_i) = [0.60, 0.40],\\ &\mathcal{R}_2(\mathcal{H},\epsilon) = \{h_1, h_2\},\ h_1(\mathbf{x}_i) = [0.85, 0.15], h_2(\mathbf{x}_i) = [0.10, 0.90].\end{aligned} \tag{6}$$

$\mathcal{R}_2(\mathcal{H},\epsilon)$ has a larger Rashomon Capacity than $\mathcal{R}_1(\mathcal{H},\epsilon)$, albeit $|\mathcal{R}_2(\mathcal{H},\epsilon)| = 2 < |\mathcal{R}_1(\mathcal{H},\epsilon)| = 3$.

3.3 Computing Rashomon Capacity of discrete Rashomon sets

The definition of Rashomon Capacity does not assume a finite cardinality of the Rashomon set. However, in order to make Rashomon Capacity computationally tractable, we assume a discrete Rashomon set, and accordingly a discrete ϵ-multiplicity set. In this case, Rashomon Capacity can be computed by standard procedures such as the Blahut–Arimoto (BA) algorithm [24, 25]. We discuss the BA algorithm in the SM and provide an implementation in the accompanying code.

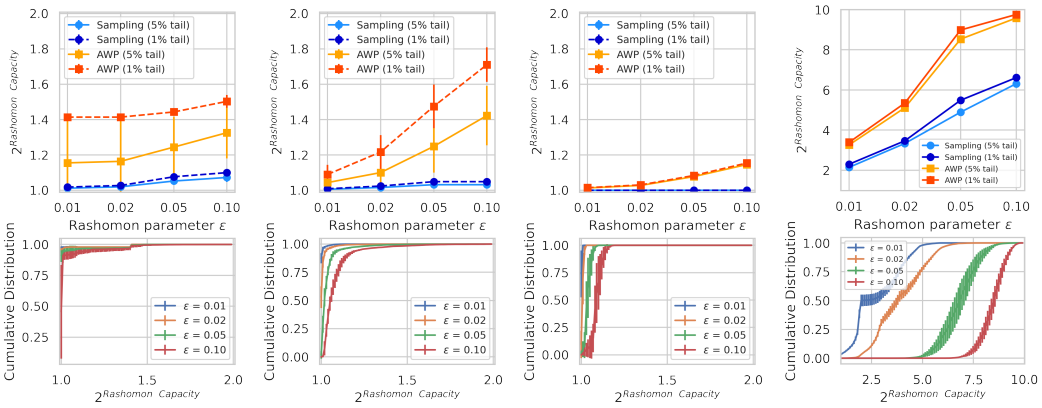
Remarkably, even when the Rashomon set has infinite cardinality, the value of Rashomon Capacity for a sample can be recovered by considering only a small number of models in the Rashomon set. In fact, for each sample $\mathbf{x}_i$, there exists a subset of at most c models that fully captures the variation in scores. This statement is formalized by the next proposition, which can be proven by applying Carathéodory's theorem [26].

Proposition 2. *For each sample $\mathbf{x}_i \in \mathcal{D}$, there exists a subset $\mathcal{A} \subseteq \mathcal{M}_\epsilon(\mathbf{x}_i)$ with $|\mathcal{A}| \leq c$ that fully captures the spread in scores for $\mathbf{x}_i$ across the Rashomon set, i.e., $m_C(\mathbf{x}_i) = 2^{C(\mathcal{A})}$. In particular, there are at most c models in $\mathcal{R}(\mathcal{H},\epsilon)$ whose output scores yield the same Rashomon Capacity for $\mathbf{x}_i$ as the entire Rashomon set.*

The above result point towards a natural strategy for empowering a stakeholder to resolve predictive multiplicity when they provide a sample $\mathbf{x}_i$: If c is small, report the (at most) c conflicting scores in $\mathcal{A} \subseteq \mathcal{M}_\epsilon(\mathbf{x}_i)$ that yield $m_C(\mathbf{x}_i) = 2^{C(\mathcal{A})}$. In terms of Rashomon Capacity, this small subset of scores captures the predictive multiplicity *across the entire Rashomon set*. The stakeholder can then choose to randomize between scores, accept the average score, or apply another appropriate strategy. The limitation of this approach is that it still requires evaluating outputs for all models in the Rashomon set. In the next section, we discuss a greedy algorithm to address this issue.

4 Empirical study

We illustrate how to measure, report, and resolve predictive multiplicity of probabilistic classifiers using Rashomon Capacity on UCI Adult [27], COMPAS [28], HSLS [29], and CIFAR-10 datasets [8]. The UCI Adult and the COMPAS are two binary classification datasets on income and recidivism prediction, respectively, and are widely used in fairness research [30]. The HSLS is an education



(a) UCI Adult (80.28%) **(b)** COMPAS (66.22%) **(c)** HSLS (70.39%) **(d)** CIFAR-10 (81.67%)

Figure 3: For each dataset (percentage is test accuracy), the top figure shows the mean and standard error of the largest 1% and 5% (1% tail and 5% tail in the legend) Rashomon Capacity among all the samples with difference Rashomon parameter ϵ. Two methods are used to obtain models from the Rashomon set, `AWP` (7) and random `sampling`. The bottom figure shows the cumulative distribution of the Rashomon Capacity of all the samples obtained by `AWP`. Each point is generated with 5 repeated splits of the dataset.

dataset, collected from high school students in the USA, whose features include student and parent information (see SM for details). We created a binary label Y from students' 9th-grade math test scores (i.e., top 50% vs. bottom 50%). We select the first three datasets to illustrate the effect of predictive multiplicity on individuals. Finally, we include the CIFAR-10 dataset to demonstrate how to report Rashomon Capacity in multi-class classification.

For the classifiers, we adopt feed-forward neural networks for the first three datasets, and a convolutional neural network VGG16 [9] for CIFAR-10. For more information on the datasets, neural network architectures, and training details, see SM Section 3. To the best of the authors' knowledge, this is the first time that predictive multiplicity metrics are measured and reported in neural networks[10]. All numbers reported are evaluated on the test set.

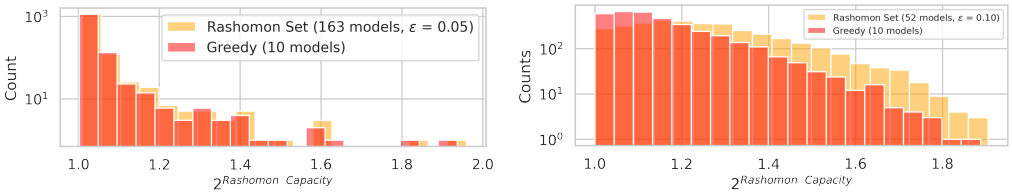
Measuring and reporting predictive multiplicity via Rashomon Capacity. The first challenge in computing Rashomon Capacity is obtaining the Rashomon set. We describe one simple method for navigating the Rashomon set next and, later in the section, we also consider sampling models in the Rashomon set via random initialization of parameters prior to training (see Section 2 for alternative strategies). Assume the models in $\mathcal{H}$ are parameterized. Given a sample $\mathbf{x}_i$, we obtain models with output predictions $\mathbf{p}_k$ by approximately solving the following optimization problem which maximizes the output score for class k:

$$\mathbf{p}_k = h_{\hat{\theta}}(\mathbf{x}_i), \text{ where } \hat{\theta} = \underset{\theta\in\Theta,\ h_\theta\in\mathcal{R}(\mathcal{H},\epsilon)}{\arg\max} [h_\theta(\mathbf{x}_i)]_k, \ \forall k = 1, 2, \cdots, c. \tag{7}$$

To solve (7), for each k, we set the objective to be $\min_{\theta\in\Theta} -[h_\theta(\mathbf{x}_i)]_k$, compute the gradients, and update the parameter θ until $h_\theta \notin \mathcal{R}(\mathcal{H}, \epsilon)$, i.e., $\hat{L}(h_\theta) > \epsilon$. Given a pre-trained model in the Rashomon set, (7) can be viewed as an adversarial weight perturbations (`AWP`) technique to explore the Rashomon set [31, 32] (see SM Section 2 for exact weight perturbation on logistic regression). With the discrete set of scores collected by solving (7), the Rashomon Capacity can be computed by the Blahut–Arimoto (BA) algorithm [24, 25] (implementation provided in accompanying code).

We preform two methods, random `sampling` with different weight initialization seeds and `AWP` (7), to obtain 100 models from the Rashomon set, and report the Rashomon Capacity in Fig. 3. In particular, we show the mean of the largest 1% and 5% Rashomon Capacity, and the cumulative distribution of the Rashomon Capacity across the samples. As the Rashomon parameter increases, both `sampling` and `AWP` lead to higher Rashomon Capacity since the Rashomon set gets larger. The `AWP` (7) achieves

[10] Breiman [1, Section 8] reported the Rashomon effect in neural networks on regression problems.



(a) COMPAS (67.35%) **(b)** HSLS (69.91%)

Figure 4: The distributions of the Rashomon Capacity for COMPAS and HSLS datasets (ϵ is percentage of mean test accuracy) obtained by sampling models from the Rashomon set and applying greedy model selection procedure (Greedy in the legend) on the sampled models.

higher Rashomon Capacity than random `sampling` as `AWP` intentionally explores the Rashomon set that maximizes the scores variations. It is important to keep in perspective that *each sample* in the high-Rashomon Capacity tail displayed in Fig. 3. corresponds to an *individual* who receive conflicting predictions. In applications such as criminal justice and education, conflicting predictions for even one individual should be reported in, e.g., model cards [13]. We report other metrics of multiplicity based such as ambiguity/discrepancy and Rashomon ratio in the SM.

Resolving predictive multiplicity by greedy model selection. We propose a greedy model selection procedure to reduce the number of competing classifiers for resolving predictive multiplicity. Given R competing classifiers, the goal is to select r models ($r < R$) that result in distributions of the Rashomon Capacity similar to that of the original R models. Starting from a dataset $\mathcal{D}$ and a Rashomon set $\mathcal{R}(\mathcal{H}, \epsilon)$, this can be implemented by (i) initializing a set $\mathcal{A}$ of models by randomly selecting a model in $\mathcal{R}(\mathcal{H}, \epsilon)$, (ii) growing $\mathcal{A}$ by adding one model from $\mathcal{R}(\mathcal{H}, \epsilon)$ that maximizes the average Rashomon Capacity across $\mathcal{D}$, and (iii) stopping until there are r models in $\mathcal{A}$. This greedy model selection is inspired by Property 4 (monotonicity) in Definition 1, since including the models to the set $\mathcal{A}$ does not reduce capacity.

In Fig. 4, we sampled 163 and 52 models from the Rashomon sets for COMPAS and HSLS datasets respectively. Here, the hypothesis space are feed-forward neural networks (see details in SM). Observe that only a small subset of the sampled models, selected by the greedy model selection procedure, is required to recover the distribution of the Rashomon Capacity. On COMPAS dataset, the 10 models obtained by the greedy model selection procedure capture the Rashomon Capacity computed with the original 163 models., i.e., these 10 models "explain" most of the score variations.

Final remarks and Limitations. We conclude by discussing the limitations and future directions. For the computation of the Rashomon Capacity, the adversarial weight perturbation (cf. (7)), despite capturing more score variations than random sampling (cf. Fig. 3), still requires training/perturbing a significant amount of models. This approach is computationally burdensome when scaling up to large datasets with millions of samples. One important limitation of Rashomon Capacity is that, due to the convexity of KL-divergence, in certain cases $m_C(\cdot)$ may seem small for already significant score variations across the classes. Indeed, this issue will occur for any strictly convex measure of divergence in (3). We provide a further discussion of how to interpret the numerical values of Rashomon Capacity in the SM.

We highlight two future directions. First, one of the core problems in studying the Rashomon effect is how to efficiently explore the Rashomon set. This is not only the computational bottleneck for estimating Rashomon Capacity, but also for other metrics including the Rashomon ratio and ambiguity/discrepancy. Consequently, designing a more efficient algorithm, e.g., an algorithm similar to the Fast Gradient Sign Method [33] in adversarial training, to explore the Rashomon set is an impactful direction. Second, the Rashomon Capacity could be generalized to other probability divergences, e.g., f-divergences [34], Rényi divergence [35], or Wasserstein distance [36]. The generalization could potentially provide further operational significance and tunability for measuring multiplicity.

References

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

[2] 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." *J. Mach. Learn. Res.*, vol. 20, no. 177, pp. 1–81, 2019.

[3] L. Semenova, C. Rudin, and R. Parr, "A study in rashomon curves and volumes: A new perspective on generalization and model simplicity in machine learning," *arXiv preprint arXiv:1908.01755*, 2019.

[4] C. Marx, F. Calmon, and B. Ustun, "Predictive multiplicity in classification," in *International Conference on Machine Learning*. PMLR, 2020, pp. 6765–6774.

[5] K. Creel and D. Hellman, "The algorithmic leviathan: Arbitrariness, fairness, and opportunity in algorithmic decision making systems," *Virginia Public Law and Legal Theory Research Paper*, 2021.

[6] H. Smith, "Algorithmic bias: should students pay the price?" *AI & society*, vol. 35, no. 4, pp. 1077–1078, 2020.

[7] INEP, "Instituto nacional de estudos e pesquisas educaionais anísio teixeira, microdados do ENEM," https://www.gov.br/inep/pt-br/assuntos/noticias/enem/ministerio-da-educacao-divulga-o-novo-enem, 2020.

[8] A. Krizhevsky, G. Hinton *et al.*, "Learning multiple layers of features from tiny images (technical report)," *University of Toronto*, 2009.

[9] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[10] X. Zhang, J. Zhao, and Y. LeCun, "Character-level convolutional networks for text classification," *Advances in neural information processing systems*, vol. 28, 2015.

[11] J. Salamon, C. Jacoby, and J. P. Bello, "A dataset and taxonomy for urban sound research," in *Proceedings of the 22nd ACM international conference on Multimedia*, 2014, pp. 1041–1044.

[12] K. Greff, R. K. Srivastava, J. Koutník, B. R. Steunebrink, and J. Schmidhuber, "Lstm: A search space odyssey," *IEEE transactions on neural networks and learning systems*, vol. 28, no. 10, pp. 2222–2232, 2016.

[13] M. Mitchell, S. Wu, A. Zaldivar, P. Barnes, L. Vasserman, B. Hutchinson, E. Spitzer, I. D. Raji, and T. Gebru, "Model cards for model reporting," in *Proceedings of the conference on fairness, accountability, and transparency*, 2019, pp. 220–229.

[14] C. Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," *Nature Machine Intelligence*, vol. 1, no. 5, pp. 206–215, 2019.

[15] L. Hancox-Li, "Robustness in machine learning explanations: does it matter?" in *Proceedings of the 2020 conference on fairness, accountability, and transparency*, 2020, pp. 640–647.

[16] 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," *arXiv preprint arXiv:2011.03395*, 2020.

[17] M. Pawelczyk, K. Broelemann, and G. Kasneci, "On counterfactual explanations under predictive multiplicity," in *Conference on Uncertainty in Artificial Intelligence*. PMLR, 2020, pp. 809–818.

[18] 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.

[19] C. Dwork and C. Ilvento, "Individual fairness under composition," *Proceedings of Fairness, Accountability, Transparency in Machine Learning*, 2018.

[20] B. Mason, R. Camilleri, S. Mukherjee, K. Jamieson, R. Nowak, and L. Jain, "Nearly optimal algorithms for level set estimation," *arXiv preprint arXiv:2111.01768*, 2021.

[21] F. Bachoc, T. Cesari, and S. Gerchinovitz, "The sample complexity of level set approximation," in *International Conference on Artificial Intelligence and Statistics*. PMLR, 2021, pp. 424–432.

[22] T. Hastie, R. Tibshirani, J. H. Friedman, and J. H. Friedman, *The elements of statistical learning: data mining, inference, and prediction*. Springer, 2009, vol. 2.

[23] T. M. Cover, *Elements of information theory*. John Wiley & Sons, 1999.

[24] R. Blahut, "Computation of channel capacity and rate-distortion functions," *IEEE transactions on Information Theory*, vol. 18, no. 4, pp. 460–473, 1972.

[25] S. Arimoto, "An algorithm for computing the capacity of arbitrary discrete memoryless channels," *IEEE Transactions on Information Theory*, vol. 18, no. 1, pp. 14–20, 1972.

[26] C. Carathéodory, "Über den variabilitätsbereich der fourier'schen konstanten von positiven harmonischen funktionen," *Rendiconti Del Circolo Matematico di Palermo (1884-1940)*, vol. 32, no. 1, pp. 193–217, 1911.

[27] M. Lichman, "UCI machine learning repository," 2013. [Online]. Available: http://archive.ics.uci.edu/ml

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

[29] S. J. Ingels, D. J. Pratt, D. R. Herget, L. J. Burns, J. A. Dever, R. Ottem, J. E. Rogers, Y. Jin, and S. Leinwand, "High school longitudinal study of 2009 (hsls: 09): Base-year data file documentation. nces 2011-328." *National Center for Education Statistics*, 2011.

[30] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan, "A survey on bias and fairness in machine learning," *ACM Computing Surveys (CSUR)*, vol. 54, no. 6, pp. 1–35, 2021.

[31] D. Wu, S.-T. Xia, and Y. Wang, "Adversarial weight perturbation helps robust generalization," *Advances in Neural Information Processing Systems*, vol. 33, pp. 2958–2969, 2020.

[32] Y.-L. Tsai, C.-Y. Hsu, C.-M. Yu, and P.-Y. Chen, "Formalizing generalization and adversarial robustness of neural networks to weight perturbations," *Advances in Neural Information Processing Systems*, vol. 34, 2021.

[33] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," *arXiv preprint arXiv:1412.6572*, 2014.

[34] I. Csiszár, "Generalized projections for non-negative functions," in *Proceedings of 1995 IEEE International Symposium on Information Theory*. IEEE, 1995, p. 6.

[35] A. Rényi, "On measures of entropy and information," in *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*, vol. 4. University of California Press, 1961, pp. 547–562.

[36] L. N. Vaserstein, "Markov processes over denumerable products of spaces, describing large systems of automata," *Problemy Peredachi Informatsii*, vol. 5, no. 3, pp. 64–72, 1969.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] See Section 4.
   (c) Did you discuss any potential negative societal impacts of your work? [Yes] See Section 1 and Section 2.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes]
   (b) Did you include complete proofs of all theoretical results? [Yes] Due to space limit, we provide the complete proofs and extra explanations in the SM, and provide sketch of the proofs in the main text.
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] We include all the code, data, and experimental details in the SM.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] All provided in the SM.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] Experiments are repeated or 5 times to compute the error bars.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] Provided in the SM.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes] We cite as thoroughly as we can.
   (b) Did you mention the license of the assets? [Yes] Datasets, code, and models are publicly available.
   (c) Did you include any new assets either in the supplemental material or as a URL? [Yes] We do not use more datasets, but include more metrics on predictive multiplicity in the SM.
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [Yes] Datasets are publicly available.
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [No] The datasets do not contain any personally identifiable information or offensive content.
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [No] Not applicable.
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [No] Not applicable.
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [No] Not applicable.