A Study in Rashomon Curves and Volumes: A New Perspective on Generalization and Model Simplicity in Machine Learning

Lesia Semenova LESIA@CS.DUKE.EDU
Department of Computer Science
Duke University
Durham, NC 27708, USA

Cynthia Rudin CYNTHIA@CS.DUKE.EDU
Departments of Computer Science, Electrical and Computer Engineering, and Statistical Science
Duke University
Durham, NC 27708, USA

Ronald Parr PARR@CS.DUKE.EDU
Department of Computer Science
Duke University
Durham, NC 27708, USA

Abstract

The *Rashomon effect* occurs when many different explanations exist for the same phenomenon. In machine learning, Leo Breiman used this term to characterize problems where many accurate-but-different models exist to describe the same data. In this work, we study how the Rashomon effect can be useful for understanding the relationship between training and test performance, and the possibility that simple-yet-accurate models exist for many problems. We consider the *Rashomon set*—the set of almost-equally-accurate models for a given problem—and study its properties and the types of models it could contain. We present the *Rashomon ratio* as a new measure related to simplicity of model classes, which is the ratio of the volume of the set of accurate models to the volume of the hypothesis space; the Rashomon ratio is different from standard complexity measures from statistical learning theory. For a hierarchy of hypothesis spaces, the Rashomon ratio can help modelers to navigate the trade-off between simplicity and accuracy. In particular, we find empirically that a plot of empirical risk vs. Rashomon ratio forms a characteristic Γ-shaped *Rashomon curve*, whose elbow seems to be a reliable model selection criterion. When the Rashomon set is large, models that are accurate—but that also have various other useful properties—can often be obtained. These models might obey various constraints such as interpretability, fairness, or monotonicity.

Keywords: Rashomon Set, Model Multiplicity, Simplicity, Generalization, Interpretable Machine Learning, Model Selection

1. Introduction

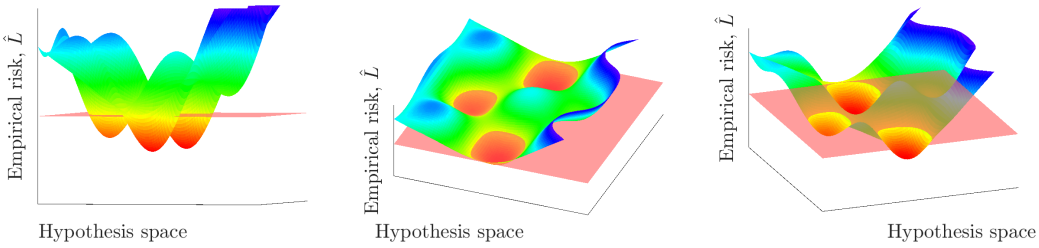
The 1950 Kurosawa film "Rashomon"(Kurosawa, 1950) revolves around four characters describing entirely different perspectives on the same crimes. Based on this idea that there could be many seemingly accurate descriptions of the same data, Leo Breiman (Breiman et al., 2001) coined the term "Rashomon effect" to describe cases when there exist many

different approximately-equally accurate models. He noticed that this effect happens very often; there is no "best" model from most finite data sets, only many good descriptions. Of course, in machine learning, our goal is not necessarily to find out the truth; it is to predict well out-of-sample.

Decades of study about generalization in machine learning have provided many different mathematical theories. Many of them measure the complexity of classes of functions without considering the data (e.g., VC theory, Vapnik, 1995), or measure properties of specific algorithms (e.g., algorithmic stability, see Bousquet and Elisseeff, 2002). However, none of these theories seems to directly capture a phenomenon that occurs throughout practical machine learning. In particular, there are a vast number of data sets for which many standard machine learning algorithms perform similarly. In these cases, the machine learning models tend to generalize well. Furthermore, in these same cases, there is often a simpler model that performs similarly and also generalizes well. Perhaps the Rashomon effect can help us to explain that phenomenon.

In this work, we aim to quantify the Rashomon effect, show that it has implications for generalization, and show it has implications for the existence of simpler models. If the Rashomon effect is large, it means that there exists a large number of models (the *empirical Rashomon set*) that perform approximately-equally-well on the training data. When the empirical Rashomon set is large, it is reasonable to assume that models with various desirable properties can exist inside it. For example, sparser, more transparent models, models that obey domain-specific constraints such as monotonicity along a given set of features, models that obey fairness constraints, and models that rely more on features that we can trust—these models may all live within the same large Rashomon set. Proving whether an interpretable (or fair, or monotonic) model exists in the Rashomon set is a challenging practical problem in general, but solving for such a model directly can be even harder. For instance, finding optimal sparse, accurate models of various forms (linear models with integer coefficients, decision sets, rule lists, decision trees) can be NP hard, sometimes with no polynomial time approximation. Let us thus return to the possibility of aiming to prove that desirable ("simpler") models exist within the Rashomon set prior to finding them. As we will show in Section 4.2, there are assumptions that allow us to prove existence of simpler models within the Rashomon set. If the assumptions are satisfied, a model from a simpler class is approximately as accurate as the most accurate model within the hypothesis space, which consequently leads to better generalization guarantees. The assumptions are based in approximation theory, which models how one class of functions can approximate another.

Proving the existence of interpretable (or fair, or otherwise constrained) models before aiming to find them differs from the current approach to machine learning in practice. Increasingly, machine learning practitioners have access to a plethora of machine learning techniques, such as deep networks, that, with some tweaking, can produce reasonable test and training performance. Such exceedingly complex models are often among the first models practitioners try to fit, rather than the last. The practical question that can arise in such cases is whether such complex models are really necessary, or if similarly accurate models that satisfy other criteria, such as interpretability or fairness, might be proven to exist with little computational effort. As we discuss later, the knowledge of a large



(a) From the side (b) From above (c) From underneath

Figure 1: An illustration of a possible Rashomon set in two dimensional hypothesis space $\mathcal{F}$. Models below the red plane belong to the Rashomon set $\hat{R}_{set}(\mathcal{F}, \theta)$, where the height of the red plane is adjusted by the Rashomon parameter θ defined in Section 3

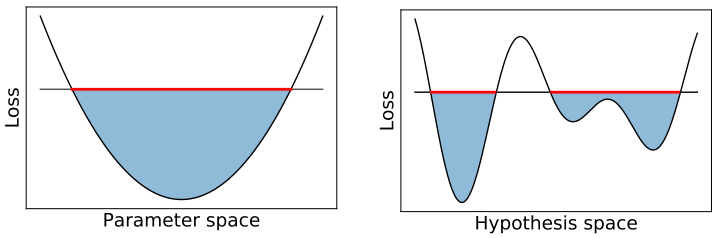
.

Rashomon set provides evidence in advance that such a search may be fruitful, and that there are computationally inexpensive ways to gauge whether the Rashomon set is large.

We quantify the magnitude of the Rashomon effect through the *Rashomon volume*, which is the size of the set of models that performs almost equally well on the training data. An illustration of the Rashomon set is shown in Figure 1. The *Rashomon ratio* is a ratio of the Rashomon volume to the volume of a hypothesis space. In this manuscript, we explore the connections between the Rashomon volume, Rashomon ratio, hierarchies of hypothesis spaces, training performance and generalization.

The Rashomon ratio can serve as a gauge of simplicity for a learning problem. As a property of both a data set and a hypothesis space, it differs from the VC dimension (Vapnik and Chervonenkis, 1971) (as Rashomon ratio is specific to a data set), it differs from algorithmic stability (see Rogers and Wagner, 1978; Kearns and Ron, 1999) (as the Rashomon ratio does not rely on robustness of an algorithm with respect to changes in the data), it differs from local Rademacher complexity (Bartlett et al., 2005) (as the Rashomon ratio does not measure the ability of the hypothesis space to handle random changes in targets and actually benefits from multiple similar models), and it differs from geometric margins (Vapnik, 1995) (as one can have a small margin with a large Rashomon ratio, and margins are measured with respect to one model, whereas the Rashomon ratio considers the existence of many). We provide theorems that show simple cases when size of the Rashomon set does not correlate with these complexity measures in Section 6. The Rashomon set is not simply a flat minimum; it could consist of many non-flat local minima as illustrated in Figure 2b, and it works for discrete hypothesis spaces where gradients, and thus "sharpness" (Dinh et al., 2017) do not exist. We provide basic generalization bounds showing how the Rashomon ratio gauges the existence of simpler-yet-accurate solutions that generalize well.

We empirically observe a trend across different data sets between the Rashomon ratio and empirical risk when considering a hierarchy of hypothesis spaces (in particular, decision trees with increasing depth, or linear models with increasing model size). This trend, which we discuss in Section 8 is in the form of a Γ-shaped curve, which is formed as we move



(a) The ϵ-flatness (b) The Rashomon set

Figure 2: Difference between ϵ-flatness as defined in Dinh et al. (2017) and the Rashomon volume. Red lines represent (a) ϵ-flatness, (b) the Rashomon volume. The height of the shaded area represents (a) the parameter ϵ or the 2σ-sharpness, (b) the Rashomon parameter θ. The ϵ-flatness is defined by a connected component in a parameter space for a given local minimum, while the Rashomon set is defined with respect to an empirical risk minimizer over the full hypothesis space $\mathcal{F}$ and may contain models from multiple local minima. Rashomon sets are also defined for discrete spaces.

up the hierarchy to increase the size of the hypothesis space. Specifically, as the size of the hypothesis space increases, first the empirical risk of the best model in the class decreases, and the Rashomon ratio grows or remains approximately the same; then the empirical risk of the best model in the class is relatively constant, but the Rashomon ratio decreases. This trend, which we call the *Rashomon curve*, occurs for all 52 out of the 52 data sets (38 classification and 14 regression) that we downloaded from the UCI Machine Learning Repository (Dua and Graff, 2019) for decision tree classifiers of various depths, and polynomial regressors of varying degrees.

In our experiments there is a connection between model selection and the Rashomon ratio. In particular, we find that the turning point in the Γ-shaped Rashomon curve, which can be formulated as a simple minimax optimization problem, is often a good choice for model selection. This *Rashomon elbow* balances between a low complexity (or a smaller size) hypothesis space (corresponding to a small Rashomon volume) and a low empirical risk (corresponding to a high accuracy). We show how the Rashomon elbow can be useful to choose the complexity of a hypothesis space to balance training accuracy with generalization.

Our results have implications beyond those where the size of the Rashomon volume can be estimated in practice. *In particular, our results indicate that when many machine learning methods perform similarly on the same data set (without overfitting), it could be because the Rashomon set of the functions these algorithms consider is large.* In that case, it may be worthwhile to find models within the Rashomon set that have desirable properties, such as interpretability. Since it is harder to optimize for constrained models to achieve interpretibility rather than simply to run several different standard machine learning methods, it is beneficial to run the standard machine learning methods first. This would allow the practitioner to gauge whether the constrained optimization is likely to yield a model that is approximately as accurate as the standard methods.

We also discuss different methods of measuring the Rashomon volume and the Rashomon ratio. For linear regression, we derive a closed form solution for the Rashomon volume in parameter space. When the Rashomon set is bounded and convex in parameter space, we design a separating oracle and use known randomized algorithm guarantees as discussed in Appendix D.2. In the more general case, we propose to use rejection sampling and importance sampling methods in the hypothesis space for estimation of the Rashomon ratio, and use these for our experiments.

We summarize the contributions of this work as follows: (i) We define the Rashomon volume and the corresponding *Rashomon ratio* as important characteristics of the Rashomon set. (ii) We provide generalization bounds for models from the Rashomon set, and show that the size of the Rashomon set serves as a barometer for the existence of accurate-yet-simpler models that generalize well. These are different from standard learning theory bounds that consider the distance between the true and empirical risks for the same function. (iii) We illustrate the Γ-shaped Rashomon curve on many data sets, and show that the Rashomon elbow can be a useful choice for model selection. (iv) We show empirically that in cases when a large Rashomon set occurs, most machine learning methods tend to perform similarly, and also in these cases, interpretable or sparse (yet accurate) models exist. (v) We provide several approaches for estimating the size of the Rashomon set. (vi) We demonstrate that the Rashomon ratio, as a gauge of simplicity of a machine learning problem, is different from other known complexity measures such as VC-dimension, algorithmic stability, geometric margin, and Rademacher complexity.

2. Related Work

There are several bodies of relevant literature as discussed below.

Rashomon sets: Rashomon sets have been used for various purposes (Breiman et al., 2001; Srebro et al., 2010; Fisher et al., 2019; Coker et al., 2018; Tulabandhula and Rudin, 2014b; Meinshausen and Bühlmann, 2010; Letham et al., 2016; Nevo and Ritov, 2017). For instance, Srebro et al. (2010) consider a loss restricted class of close-to-optimal models, and with an assumption of H-smoothness of a loss function, they obtain a tighter excess risk bound through local Rademacher complexity (Bartlett et al., 2005). Our bounds do not work the same way and aim to prove a different type of result. Other works aim to search through the Rashomon set to find the most extreme models within it, rather than looking at the size of the Rashomon set, as we do in this work. Fisher et al. (2019) leverages the Rashomon set in order to understand the spectrum of variable importance and other statistics across the set of good models. Our work considers the existence of models from simpler classes rather than exploring the Rashomon set to find a range of variable importance or other statistics. The work of Tulabandhula and Rudin (2013, 2014a,b) uses the Rashomon set to assist with decision making, by finding the range of downstream operational costs associated with the Rashomon set. Rashomon sets are related to p-hacking and robustness of estimation, because the Rashomon set is a set over which one might conduct a sensitivity analysis to choices made by an analyst (Coker et al., 2018).

Flat minima or wide valleys: The concept of flat minima (wide valleys) has been explored in the deep learning literature as a possible way to understand convergence properties of the complicated, non-convex loss functions that deep networks traverse during training

(Hochreiter and Schmidhuber, 1997; Dinh et al., 2017; Keskar et al., 2016; Chaudhari et al., 2016; Keskar et al., 2016). Based on a minimum-message-length argument (Wallace and Boulton, 1968), several works claim that flat loss functions lead to better generalization due to a robustness to noise around the minimum (Hochreiter and Schmidhuber, 1997; Keskar et al., 2016; Chaudhari et al., 2016). Following Hochreiter and Schmidhuber (1997), Dinh et al. (2017) define ϵ-flatness, which constitutes a special case of our Rashomon sets, as shown in Figure 2. In particular, our Rashomon set is defined over the hypothesis (functional) space, while ϵ-flatness is defined in a parameter space (though sometimes we use parameter space for ease of computation), and the Rashomon set is not necessarily a single connected component (although it might be in the case of a convex loss over a continuous domain), while ϵ-flatness pertains only to a connected set. This means that the Rashomon set can contain models from different local minima, or can be defined on discrete spaces, while ϵ-flatness is relevant only for continuous loss functions. Another way of quantifying flatness is σ-sharpness (Keskar et al., 2016; Dinh et al., 2017), which measures the change of the loss function inside a σ-ball in a parameter space. In the case of a connected Rashomon set, this loss difference corresponds to the Rashomon parameter θ.

Statistical learning theory: Numerous works provide generalization bounds based on different complexity measures, and under different assumptions. Some of them include Rademacher (Srebro et al., 2010; Kakade et al., 2009) and Gaussian complexities (Kakade et al., 2009), PAC-Bayes theorems (Langford and Shawe-Taylor, 2003), covering numbers bounds (Zhou, 2002), and margin bounds (Vapnik and Chervonenkis, 1971; Schapire et al., 1998; Koltchinskii et al., 2002), etc. In contrast, under assumptions elaborated in Section 4, the Rashomon ratio provides a certificate of the existence of a simpler model that generalizes, rather than acting itself as a simplicity measure. The use of approximating sets, as used extensively in this paper, is used throughout the literature on learning theory (Lecué, 2011; Lugosi and Wegkamp, 2004; Schapire et al., 1998; Mendelson, 2003). An excellent example of this is the classical generalization bound for boosting and margins (Schapire et al., 1998), which uses combinations of several random draws of base classifiers to represent combinations of base classifiers. This is an instance of the so-called "Maurey's lemma," which often provides this approximating set for linear model classes.

Model selection: The closest model selection literature to our work (Lugosi and Nobel, 1999; Shawe-Taylor et al., 1998) focuses on separately estimating the complexity of each hypothesis space within a hierarchy. The complexity of each member of the hierarchy is measured through VC-dimension (Shawe-Taylor et al., 1998) or empirical covers (Lugosi and Nobel, 1999); once a member of the hierarchy is chosen (a specific hypothesis space), they would choose a model that minimizes risk within the hypothesis space. Our work also chooses models that minimize the risk in each hypothesis space, and provides a specific guideline on how to choose the hypothesis space using the Rashomon elbow.

We know of no other works that illustrate the Rashomon curve, which we have observed universally among data sets we have considered.

3. Rashomon Set Definitions and Notation

Consider a training set of n data points $S = \{z_1, z_2, ..., z_n\}$, $z_i = (x_i, y_i)$ drawn i.i.d. from an unknown distribution $\mathcal{D}$ on a bounded set $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$, where $\mathcal{X} \subset \mathbb{R}^p$ and $\mathcal{Y} \subset \mathbb{R}$ are an

input and an output space respectively. Our hypothesis space is $\mathcal{F} = \{f : \mathcal{X} \to \mathcal{Y}\}$. We limit the hypothesis space $\mathcal{F}$ to contain only models that vary within the bounded domain $\mathcal{Z}$ where the data reside. We will assume that the hypothesis space is bounded and that there is a prior distribution ρ over functions in $\mathcal{F}$. Often we will use a term functional space to refer to a non-parametric hypothesis space. To measure the quality of a prediction made by a hypothesis, we use a loss function $\phi : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}^+$. Specifically, for each given point $z = (x, y)$ and a hypothesis f, the loss function is $\phi(f(x), y)$. For a given f we will also overload notation by writing $l : \mathcal{F} \times \mathcal{Z} \to \mathbb{R}^+$ that takes f explicitly as an argument: $l(f, z) = \phi(f(x), y)$. We are interested in learning a model f that minimizes the *true risk* $L(f) = \mathbb{E}_{z\sim\mathcal{D}}[\phi(f(x), y)]$, which depends on an unknown distribution $\mathcal{D}$ and therefore is estimated with an *empirical risk*: $\hat{L}(f) = \frac{1}{n}\sum_{i=1}^{n} \phi(f(x_i), y_i)$. For the rest of this paper, data are drawn from the unknown distribution $\mathcal{D}$ on $\mathcal{X} \times \mathcal{Y}$, unless otherwise specified.

3.1 Basic Rashomon Set and Ratio Definition

We define the *empirical Rashomon set* (or simply *Rashomon set*) as a subset of models of the hypothesis space $\mathcal{F}$ that have training performance close to the best model in the class, according to a loss function (Breiman et al., 2001; Srebro et al., 2010; Fisher et al., 2019; Coker et al., 2018; Tulabandhula and Rudin, 2014b). More precisely:

Definition 1 (Rashomon set) *Given $\theta \geq 0$, a data set S, a hypothesis space $\mathcal{F}$, and a loss function ϕ, the Rashomon set $\hat{R}_{set}(\mathcal{F}, \theta)$ is the subspace of the hypothesis space defined as follows:*

$$\hat{R}_{set}(\mathcal{F}, \theta) := \{f \in \mathcal{F} : \hat{L}(f) \leq \hat{L}(\hat{f}) + \theta\},$$

where $\hat{f}$ is an empirical risk minimizer for the training data S with respect to a loss function ϕ: $\hat{f} \in \operatorname{argmin}_{f\in\mathcal{F}} \hat{L}(f)$.

If we want to specify the data set S that is used to compute the Rashomon set, we indicate the data set in the subscript, as $\hat{R}_{set_S}(\mathcal{F}, \theta)$. Fisher et al. (2019)'s definition of Rashomon set is distinct from ours in that we typically use an empirical risk minimizer to define the Rashomon set instead of a prespecified reference model which is independent of the sample.

The hypothesis space $\mathcal{F}$ in the definition of the Rashomon set can be a specific well-defined hypothesis space, such as the space of decision trees of depth D or neural nets with D hidden layers, or it can be a more general space (a meta-hypothesis space) that contains models from different hypothesis spaces (e.g., linear functions, polynomials up to degree D, and piece-wise constant functions) with the training error as a loss function.

We call θ the *Rashomon parameter*. To compute the size of the Rashomon set we will use the *Rashomon volume* $\mathcal{V}(\hat{R}_{set}(\mathcal{F}, \theta))$, where $\mathcal{V}(\cdot) : \mathcal{F} \to \mathbb{R}_{\geq 0}$ measures the volume of a set in the hypothesis space, potentially weighted by the prior ρ over functions. The practitioner can define the Rashomon volume in a way that would be specific to a learning problem and hypothesis space. In general, the Rashomon volume is $\mathcal{V}(\hat{R}_{set}(\mathcal{F}, \theta)) = \int_{f\in\mathcal{F}} 1_{f\in\hat{R}_{set}(\mathcal{F},\theta)}\rho(f)d\rho$, where ρ is a prior on the hypothesis space. We assume that the Rashomon set is bounded and therefore $\mathcal{V}(\hat{R}_{set}(\mathcal{F}, \theta)) < \infty$. If the hypothesis space is discrete with uniform prior then the Rashomon volume can be calculated

by counting how many models are inside $\hat{R}_{set}(\mathcal{F}, \theta)$ or, in other words, by computing the cardinality of the Rashomon set directly: $\mathcal{V}(\hat{R}_{set}(\mathcal{F}, \theta)) = \sum_{f \in \mathcal{F}} 1_{f \in \hat{R}_{set}(\mathcal{F}, \theta)}$.

When we compare the size of the Rashomon set for different hypothesis spaces, the Rashomon volume might not be an informative measure, especially if the hypothesis spaces we consider have very different complexity. To normalize the Rashomon volume, we introduce the *Rashomon ratio*, which uses the hypothesis space to form the denominator of the ratio. Assumptions that $\mathcal{F}$ is bounded and contains models within the bounded domain $\mathcal{Z}$ allow us to keep the ratio denominator from containing functions that are irrelevant (and thus increase the denominator without a chance of increasing the numerator). Consequently, $\mathcal{V}(\mathcal{F}) < \infty$. We define the Rashomon ratio as follows:

Definition 2 (Rashomon ratio) *Let $\mathcal{F}$ be a hypothesis space given a data set S. The Rashomon ratio is a ratio of the volume of models inside the Rashomon set $\hat{R}_{set}(\mathcal{F}, \theta)$ to the volume of models in the hypothesis space $\mathcal{F}$:*

$$\hat{R}_{ratio}(\mathcal{F}, \theta) = \frac{\mathcal{V}(\hat{R}_{set}(\mathcal{F}, \theta))}{\mathcal{V}(\mathcal{F})}. \tag{1}$$

By our definitions, this ratio is always between 0 and 1. It represents the fraction of models that are good (the fraction of models that fit the data about equally well). If θ is small, a larger Rashomon ratio implies that more models perform about equally well. When θ is large enough, the Rashomon set contains all models in $\mathcal{F}$ and the ratio is 1.

As before, we indicate the data set S that is used to compute the Rashomon ratio in the subscript, as $\hat{R}_{ratio_S}(\mathcal{F}, \theta)$.

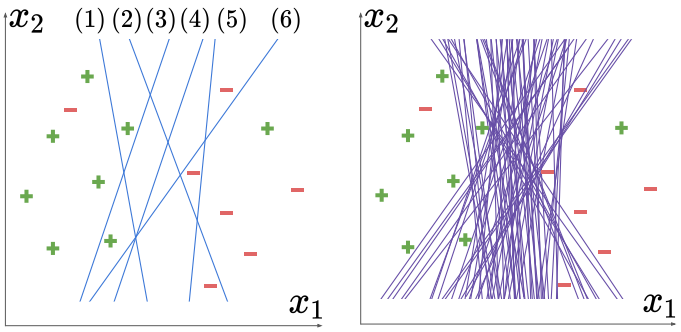
In the definitions above, the Rashomon set considered multiplicity of models. In the definition in the next subsection, we consider multiplicity of predictions instead. Whereas in Definition 1, two models that are different but make the same predictions would be considered different, these two models would join the same equivalence class in the definition of the pattern Rashomon ratio.

3.2 Pattern Rashomon Ratio for Binary Classification

For a binary classification on a given data set S we introduce the pattern Rashomon ratio as follows.

Definition 3 (Pattern Rashomon ratio) *Given a hypothesis space $\mathcal{F}$, data set S, and a binary-valued function $\zeta : \mathcal{F} \times \mathcal{Z}^n \to \{0, 1\}$ (which is usually either $sign(f(x_i))$ or the loss $1_{[sign(f(x_i)) \neq y_i]}$), the pattern Rashomon ratio is the ratio of possible realizations of ζ vectors from functions within the Rashomon set. Denote $\boldsymbol{\zeta}(f, S) = [\zeta(f, z_1), \zeta(f, z_2), ..., \zeta(f, z_n)]$. Also denote $binary(i)$ as the vectorized binary representation of i of size n (e.g., $binary(1) = [0, 0, 0, 0, 1]$ when n=5). The pattern Rashomon ratio is:*

$$\hat{R}^{pat}_{ratio}(\mathcal{F}, \theta) = \frac{\sum_{j=0}^{2^n-1} \mathbb{1}[\exists f \in \hat{R}_{set_S} : \boldsymbol{\zeta}(f, S) = binary(j)]}{\sum_{j=0}^{2^n-1} \mathbb{1}[\exists f \in \mathcal{F} : \boldsymbol{\zeta}(f, S) = binary(j)]}.$$



(a) (b)

Figure 3: (a) Classifiers in the figure each define a different loss pattern. The number of distinct patterns (in Figure (a) there are six patterns) created by functions in the Rashomon set comprises the numerator of the pattern Rashomon ratio. (b) Each classifier in the figure is a different model in the Rashomon set. The fraction of models in the Rashomon set is the Rashomon ratio.

The ratio based on patterns is different from the Rashomon ratio defined in (1) as a multiplicity of models, as shown in Figure 3. The pattern Rashomon ratio measures the diversity of predictions made by functions in the Rashomon set compared to the diversity of prediction of functions within the hypothesis space. If the pattern Rashomon ratio is high, it means that the Rashomon set contains not only multiple models, but also multiple models with different properties, depending on the definition of ζ.

The pattern Rashomon ratio has useful approximation guarantees. In particular as the size of the model space D grows to be infinitely large (e.g., the depth of the decision tree grows infinitely, or number of parameters grows to infinity), the pattern Rashomon ratio approaches a fixed value that depends on the Rashomon parameter and number of points in the data set only. This intuition is summarized in the next proposition.

Proposition 4 (Approximation guarantees for the pattern Rashomon ratio) *Let D represent the size of the hypothesis space $\mathcal{F}$. For binary classification and sign performance function $\zeta(f, z) = sign(f(x))$, as $D \to \infty$, the pattern Rashomon ratio $\hat{R}^{pat}_{ratio}(\mathcal{F}, \theta) \to \bar{R}^{pat} = \frac{\sum_{i \leq \lfloor \theta n \rfloor} \binom{n}{i}}{2^n}$, and for $\theta \leq 1/2$: $\frac{2^{n(H(\theta)-1)}}{\sqrt{8n\theta(1-\theta)}} \leq \bar{R}^{pat} \leq 2^{n(H(\theta)-1)}$, where n is the size of the training data set, and $H(\theta) = -\theta \log_2 \theta - (1-\theta)\log_2(1-\theta)$ is the binary entropy.*

In contrast, there is no obvious limit value for the Rashomon ratio. There exist data distributions such that for a fixed value θ, as the size of the hypothesis space grows, the Rashomon ratio will converge to 0. There also exist data distributions such that the Rashomon ratio may not converge to either zero or one. For example, separable data with a large margin may lead to a limiting Rashomon ratio that is greater than zero.

The Rashomon ratio and the pattern Rashomon ratio, as properties of a data set and a hypothesis space, serve as gauge of simplicity of the learning problem. A large Rashomon

ratio means that the Rashomon set contains multiple models within the hypothesis space that have approximately constant empirical risk. These accurate models could either come from multiple local minima or from one wide local minimum. In this case, potentially every reasonable optimization procedure could lead to a hypothesis from the Rashomon set. Therefore, for large Rashomon sets, accurate models tend to be easier to find (as optimization procedures can find them). *In other words, if the Rashomon ratio is large, the Rashomon set could contain many accurate and simple models, and the learning problem becomes simpler.* On the other hand, smaller Rashomon ratios might imply a harder learning problem, especially in the case of few deep and narrow local minima.

We introduce one more variation of the Rashomon set, where the threshold that restricts the risk does not depend on the empirical risk minimizer. Given a parameter $\gamma \geq 0$, we call the Rashomon set with restricted empirical risk an *anchored Rashomon set*:

$$\hat{R}^{anc}_{set}(\mathcal{F}, \gamma) := \{f \in \mathcal{F} : \hat{L}(f) \leq \gamma\}.$$

We define also the *true anchored Rashomon set* based on the true risk as follows:

$$R^{anc}_{set}(\mathcal{F}, \gamma) := \{f \in \mathcal{F} : L(f) \leq \gamma\}.$$

We will denote $\hat{R}^{anc}_{ratio}(\mathcal{F}, \gamma)$ and $R^{anc}_{ratio}(\mathcal{F}, \gamma)$ as the Rashomon ratios computed on the anchored Rashomon set and the true anchored Rashomon set. Potentially, we could choose γ so that the true anchored Rashomon set definition mirrors Definition 1 and the true risk is restricted by quantity $\gamma = L(f^*) + \theta$, where $f^* \in \mathcal{F}$ is a model that minimizes the true risk.

The true anchored Rashomon set, as it turns out, can be a (practically unmeasurable) certificate of the existence of a simpler model. Since we can never actually explore the anchored Rashomon set, we would never know whether it will be (or has been) useful for a particular problem. We explain this in the next section, and then spend most of our effort considering *empirical* Rashomon sets, which are easier to work with in practice. In the next section, we discuss the simplicity and generalization properties of models that are in the Rashomon set.

4. Rashomon Set Models: Simplicity and Generalization

Building on the discussions from the previous section, let us consider two hypothesis (functional) spaces with different levels of complexity, where the lower-complexity space serves as a good *approximating set* for the higher-complexity space. The functional spaces are called $\mathcal{F}_1$, for the simpler space, and $\mathcal{F}_2$, for the more complex space, where $\mathcal{F}_1 \subset \mathcal{F}_2$. Here, to determine the complexity of a functional space, we use traditional notions of complexity (conversely, simplicity) such as covering numbers or VC dimension. For a useful example of a simple and a more complex space, consider $\mathcal{F}_2$ to be the space of linear models with real-valued coefficients in a space of d dimensions, and consider $\mathcal{F}_1$ to be the space of scoring systems (Ustun and Rudin, 2016), which are sparse linear models, with at most d' nonzero integer coefficients, $d' \ll d$.

Generalization bounds would be tighter if we could use the lower complexity space $\mathcal{F}_1$, but as we are considering functions from $\mathcal{F}_2$, learning theory often has us include the complexity of $\mathcal{F}_2$ in the bound. Given this, we have several questions to answer:

1. What if the higher-complexity hypothesis space we chose was more complex than necessary for modeling the data? In that case, can we still have guarantees on test performance of the best classifier in the complex space $\mathcal{F}_2$ that leverage the complexity of the lower-complexity space $\mathcal{F}_1$ instead of that of space $\mathcal{F}_2$? In particular, perhaps special properties of the more complex space can help our analysis; if these properties hold, then we can get all the guarantees we need about the best possible attainable test performance on the more complex space $\mathcal{F}_2$ by looking only at optimal training performance on the less complex space $\mathcal{F}_1$. (As a preview to Section 4.1 where we answer this question, the property on the complex space that will help us is that the true anchored Rashomon set of $\mathcal{F}_2$ is large. However, we cannot know in practice when this property holds, which is a disadvantage of this analysis. When the property holds, *even if we do not know it holds*, then this property still becomes helpful in practice as it guarantees when it is beneficial to choose a simper hypothesis space.)

2. Before looking at the data, let us assume we chose to examine the more complex space $\mathcal{F}_2$, not knowing that the lower-complexity space $\mathcal{F}_1$ would suffice. We may have done this for computational reasons, since perhaps it is much easier to optimize over the higher-complexity space than the lower-complexity space. For instance, as before, perhaps the lower-complexity space consists of scoring systems, which are combinatorially hard to optimize over, whereas the larger space considers standard linear models with real-valued coefficients, which are much easier to optimize. Our question is, after having done some analysis on the higher complexity space, can we still have generalization guarantees that use only the complexity of the lower-complexity space? Specifically, can we guarantee the existence of a simple-yet-accurate model (a model from the lower-complexity space with low training error) that will generalize well? Can we guarantee that many such simple-yet-accurate models exist? (As a preview to Section 4.2, we will use smoothness and the size of the Rashomon set as key tools to prove the existence of simple-yet-accurate models.)

3. Again let us assume we chose to examine the more complex hypothesis space, $\mathcal{F}_2$. If $\mathcal{F}_1$ serves as a good approximating set for $\mathcal{F}_2$, then we can easily get a generalization bound for all $f_2 \in \hat{R}_{set}(\mathcal{F}_2, \theta)$ that uses the complexity of $\mathcal{F}_1$. This bound would indicate that the learning problem is not actually as complicated as it might seem if we had only looked at the more complex space of functions $\mathcal{F}_2$. In Section 4.3 we will again use smoothness to create this bound.

In creating these bounds, we hoped to distill the problems to their essence, so that the bounds are as close as possible to basic learning theory bounds. These bounds, including those for discrete hypothesis spaces can be generalized to more complex statistical learning theory analyses if desired. Note that the bounds in Section 4.1 do not serve the same purpose as standard statistical learning theoretic bounds, as they do not aim to bound generalization error for a single function (that is, the difference between training and test loss for a function). Rather, we are interested in differences between training loss of one function and test loss of another. Standard learning theory analysis handles the single function case nicely; we are concerned with other questions here.

4.1 The True Anchored Rashomon Set Can Be Very Helpful... But You Might Not Know When

As in classic Occham's razor bounds, we start with finite hypothesis spaces. Let us consider finite hypothesis spaces $\mathcal{F}_1$ and $\mathcal{F}_2$, where $\mathcal{F}_1 \subset \mathcal{F}_2$. Consider the first question discussed above: Given $\mathcal{F}_1$ and $\mathcal{F}_2$, can we have guarantees on the best possible test performance of models in $\mathcal{F}_2$, which depend only on $\mathcal{F}_1$'s complexity and empirical performance? In the following theorem, we will make a key assumption that allows us to do this: we assume a sufficiently large anchored true Rashomon set for $\mathcal{F}_2$. Specifically, we assume that there are a large number of functions from $\mathcal{F}_2$ that have true risk below γ. This large number of functions is assumed to be large enough to contain at least one function from $\mathcal{F}_1$ (later, in Section 4.2 we show conditions under which simple models from $\mathcal{F}_1$ exist in the Rashomon set of $\mathcal{F}_2$). Since this special function from $\mathcal{F}_1$ is likely to generalize between training and test error (due to learning theory results on $\mathcal{F}_1$), we will be able to analyze the best possible true risk from $\mathcal{F}_2$ in terms of what we can observe from $\mathcal{F}_1$ on the data.

Here, $|\mathcal{F}|$ denotes the cardinality of the finite space $\mathcal{F}$. These bounds can be generalized to infinite hypothesis spaces, but they are designed for intuition, which works nicely with finite hypothesis spaces.

Theorem 5 (The advantage of a large true anchored Rashomon set I) *Consider finite hypothesis spaces $\mathcal{F}_1$ and $\mathcal{F}_2$, such that $\mathcal{F}_1 \subset \mathcal{F}_2$. Let the loss l be bounded by b, $l(f_2, z) \in [0, b] \;\; \forall f_2 \in \mathcal{F}_2, \forall z \in \mathcal{Z}$. Define an optimal function $f_2^* \in \operatorname{argmin}_{f_2 \in \mathcal{F}_2} L(f_2)$. Let us assume that the true anchored Rashomon set is large enough to include a function from $\mathcal{F}_1$, so there exists a model $\tilde{f}_1 \in \mathcal{F}_1$ such that $\tilde{f}_1 \in R^{anc}_{set}(\mathcal{F}_2, \gamma)$. In that case, for any $\epsilon > 0$ with probability at least $1 - \epsilon$ with respect to the random draw of data:*

$$|L(f_2^*) - \hat{L}(\hat{f}_1)| \leq \gamma + 2b\sqrt{\frac{\log |\mathcal{F}_1| + \log 2/\epsilon}{2n}},$$

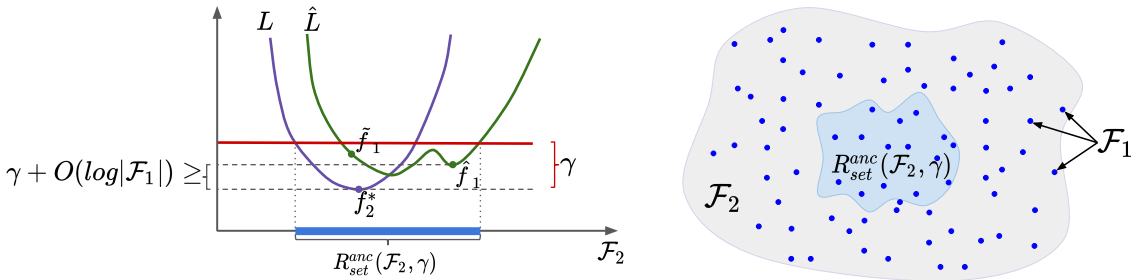
where $\hat{f}_1 \in \operatorname{argmin}_{f_1 \in \mathcal{F}_1} \hat{L}(f_1)$.

That is, when the assumption on the geometry of the true risk landscape is correct, we can approximate an optimal model from $\mathcal{F}_2$ with the best empirical model within $\mathcal{F}_1$, and the bound depends only on the complexity of $\mathcal{F}_1$ and not $\mathcal{F}_2$ as shown in Figure 4(a). In this case, if we work only with $\mathcal{F}_1$ empirically, we would achieve test performance on par with the best test set performance achievable in $\mathcal{F}_2$. We can further improve the bound in Theorem 5 by eliminating the complexity term, at the expense of doubling the γ term.

Theorem 6 (The advantage of a good approximating set) *Consider hypothesis spaces $\mathcal{F}_1$ and $\mathcal{F}_2$, such that $\mathcal{F}_1 \subset \mathcal{F}_2$. Let the loss l be bounded by b, $l(f_2, z) \in [0, b] \;\; \forall f_2 \in \mathcal{F}_2, \forall z \in \mathcal{Z}$. Define an optimal function $f_2^* \in \operatorname{argmin}_{f_2 \in \mathcal{F}_2} L(f_2)$. Let us assume that the true anchored Rashomon set is large enough to include a function from $\mathcal{F}_1$, so there exists a model $\tilde{f}_1 \in \mathcal{F}_1$ such that $\tilde{f}_1 \in R^{anc}_{set}(\mathcal{F}_2, \gamma)$. In that case, for any $\epsilon > 0$ with probability at least $1 - \epsilon$ with respect to the random draw of data:*

$$|L(f_2^*) - \hat{L}(\hat{f}_1)| \leq 2\gamma + b\sqrt{\frac{\log 1/\epsilon}{2n}},$$

where as before, $\hat{f}_1 \in \operatorname{argmin}_{f_1 \in \mathcal{F}_1} \hat{L}(f_1)$.



(a) (b)

Figure 4: (a) For $\mathcal{F}_1 \subset \mathcal{F}_2$, the true risk of $\mathcal{F}_2$ and the empirical risk of $\mathcal{F}_1$ are close if there exists a model $\tilde{f}_1$ in the intersection of $\mathcal{F}_1$ and the anchored Rashomon set of $\mathcal{F}_2$ as shown in Theorem 5. (b) $\mathcal{F}_1$ is formed by random sampling of $\mathcal{F}_2$. If we sample sufficiently many models from $\mathcal{F}_2$ to be included in $\mathcal{F}_1$, with high probability there will be a model from $\mathcal{F}_1$ that will be within the Rashomon set of $\mathcal{F}_2$.

Theorem 6 can be tighter than Theorem 5 if the approximating set $\mathcal{F}_1$ contains a function that is very close to the minimizer of the true loss, relative to the log of the size of $\mathcal{F}_1$. Here, smaller values of γ will result in a tighter bound, though on the other hand, larger values may be needed to ensure the existence of a model $\tilde{f}_1$ in the Rashomon set. In this way, the bound indicates that finding a good approximating set can be important: a better approximating set could allow Theorem 6 to yield a tighter bound than Theorem 5.

The main assumption (of a sufficiently large anchored true Rashomon set) in Theorems 5 and 6 is an assumption about the population, and does not rely on the sample. It relies only on the existence of one special function in the true anchored Rashomon set. There are no smoothness assumptions on the loss function. If the main assumption of these theorems holds, then we gain the benefit of guarantees on $\mathcal{F}_2$ from looking only at $\mathcal{F}_1$ empirically. We cannot check whether the assumption holds since it involves the true risk, but practitioners can reap the benefits of it anyway: when minimizing over $\mathcal{F}_1$, they may unknowingly be achieving test error close to the best of $\mathcal{F}_2$.

To make the connection of this result to Rashomon sets more explicit, we will choose a specific relationship between $\mathcal{F}_1$ and $\mathcal{F}_2$, specifically, $\mathcal{F}_1$ will be a random sample of $\mathcal{F}_2$ that is chosen prior to, and separately from, learning. This is an artificial example in that $\mathcal{F}_1$ would never actually be chosen as a random sample from $\mathcal{F}_2$ in reality. However, the random sampling assumption permits $\mathcal{F}_1$ to be distributed fairly evenly around $\mathcal{F}_2$, which, arguably, could approximate the way some simpler spaces are embedded in more complex functional spaces. For instance, approximation theory results (discussed later) ensure that functions from some spaces can approximate all functions from other spaces.

If $\mid \mathcal{F}_2 \mid = 100000$ and $\hat{R}_{ratio}(\mathcal{F}_2, \gamma) \geq 0.1\%$ then if $\mid \mathcal{F}_1 \mid \geq 5156$ then with probability at least 99% the bound 2 holds.
If $\mid \mathcal{F}_2 \mid = 100000$ and $\hat{R}_{ratio}(\mathcal{F}_2, \gamma) \geq 0.5\%$ then if $\mid \mathcal{F}_1 \mid \geq 1051$ then with probability at least 99% the bound 2 holds.
If $\mid \mathcal{F}_2 \mid = 100000$ and $\hat{R}_{ratio}(\mathcal{F}_2, \gamma) \geq 1\%$ then if $\mid \mathcal{F}_1 \mid \geq 526$ then with probability at least 99% the bound 2 holds.
If $\mid \mathcal{F}_2 \mid = 100000$ and $\hat{R}_{ratio}(\mathcal{F}_2, \gamma) \geq 2\%$ then if $\mid \mathcal{F}_1 \mid \geq 262$ then with probability at least 99% the bound 2 holds.
If $\mid \mathcal{F}_2 \mid = 100000$ and $\hat{R}_{ratio}(\mathcal{F}_2, \gamma) \geq 5\%$ then if $\mid \mathcal{F}_1 \mid \geq 104$ then with probability at least 99% the bound 2 holds.

Table 1: Examples of the possible usage of Theorem 7.

4.1.1 An Example of $\mathcal{F}_1$ and $\mathcal{F}_2$ Where We Can Estimate the Probability with which a Simpler[1] Model is Guaranteed to Be in the Rashomon Set.

If $\mathcal{F}_1$ is a random sample of functions from $\mathcal{F}_2$ (as illustrated in Figure 4(b)), and if $\mathcal{F}_2$ has a large true anchored Rashomon set, then $\mathcal{F}_1$ is likely to include at least one model from the true anchored Rashomon set. In that case, Theorem 5 applies. Conversely, if $\mathcal{F}_2$ has a small true anchored Rashomon set, $\mathcal{F}_1$ is unlikely to contain a model from the true anchored Rashomon set, in which case, Theorem 5 does not apply, and there is no guarantee.

Theorem 7 (The advantage of a large true anchored Rashomon set II) *Consider finite hypothesis spaces $\mathcal{F}_1$ and $\mathcal{F}_2$, such that $\mathcal{F}_1 \subset \mathcal{F}_2$ and $\mathcal{F}_1$ is uniformly drawn from $\mathcal{F}_2$ without replacement. Define an optimal function $f_2^* \in \operatorname{argmin}_{f_2 \in \mathcal{F}_2} L(f_2)$. For a loss l bounded by b and any $\epsilon > 0$, with probability at least $(1-\epsilon)\, p$ with respect to the random draw of functions from $\mathcal{F}_2$ to form $\mathcal{F}_1$ and with respect to the random draw of data:*

$$|L(f_2^*) - \hat{L}(\hat{f}_1)| \leq \gamma + 2b\sqrt{\frac{\log \mid \mathcal{F}_1 \mid + \log 2/\epsilon}{2n}}, \tag{2}$$

where $p = 1 - \frac{\binom{(1-R^{anc}_{ratio}(\mathcal{F}_2,\gamma))\mid \mathcal{F}_2 \mid}{\mid \mathcal{F}_1 \mid}}{\binom{\mid \mathcal{F}_2 \mid}{\mid \mathcal{F}_1 \mid}} = 1 - \prod_{i=1}^{|R^{anc}_{set}(\mathcal{F}_2,\gamma)|} \left(1 - \frac{\mid \mathcal{F}_1 \mid}{\mid \mathcal{F}_2 \mid - |R^{anc}_{set}(\mathcal{F}_2,\gamma)| + i}\right)$, *and* $\hat{f}_1 \in \operatorname{argmin}_{f_1 \in \mathcal{F}_1} \hat{L}(f_1)$.

Please refer to Table 1 for lower bounds of $\mid \mathcal{F}_1 \mid$ from Theorem 7, given values of $\mid \mathcal{F}_2 \mid$ and $\hat{R}_{ratio}(\mathcal{F}_2, \gamma)$.

Theorem 7 holds with a probability that depends on the likelihood of drawing $\mathcal{F}_1$ from $\mathcal{F}_2$. If $\mid \mathcal{F}_1 \mid$ is small compared with $\mid \mathcal{F}_2 \mid$, the probability p in the theorem statement is small as well. In the following lemma, we provide a lower bound on the Rashomon ratio so that the bound in Theorem 9 holds with probability $(1-\epsilon)^2$, instead of $(1-\epsilon)p$. As a key step of the proof of the lemma, we will lower bound the probability of sampling without

1. "Simpler" in this section means that the model comes from a discrete hypothesis space with smaller cardinality.

If $\|F_1\| = 100000$ then to get the bound 2 to hold with probability at least 99% the Rashomon ratio should be $\hat{R}_{ratio}(\mathcal{F}_2, \gamma) \geq (5.026 \times 10^{-8})\%$.
If $\|F_1\| = 10000$ then to get the bound 2 to hold with probability at least 99% the Rashomon ratio should be $\hat{R}_{ratio}(\mathcal{F}_2, \gamma) \geq (5.026 \times 10^{-7})\%$.
If $\|F_1\| = 1000$ then to get the bound 2 to hold with probability at least 99% the Rashomon ratio should be $\hat{R}_{ratio}(\mathcal{F}_2, \gamma) \geq (5.026 \times 10^{-6})\%$.

Table 2: Examples of the possible usage of Theorem 9.

replacement with the probability of sampling with replacement. Please see the details of the proof in Appendix C.4.

Lemma 8 *For a finite hypothesis space $\mathcal{F}_2$ of size $|\mathcal{F}_2|$, we will draw $|\mathcal{F}_1|$ functions uniformly without replacement from $\mathcal{F}_2$ to form $\mathcal{F}_1$. If the true anchored Rashomon ratio of the hypothesis space $\mathcal{F}_2$ is at least*

$$R^{anc}_{ratio}(\mathcal{F}_2, \gamma) \geq 1 - \epsilon^{\frac{1}{|\mathcal{F}_1|}}$$

then with probability at least $1 - \epsilon$ with respect to the random draw of functions from $\mathcal{F}_2$ to form $\mathcal{F}_1$, the Rashomon set contains at least one model $\tilde{f}_1$ from $\mathcal{F}_1$.

Combining Lemma 8 and Theorem 5 we get the following theorem:

Theorem 9 (The advantage of a large true anchored Rashomon set III) *Consider finite hypothesis spaces $\mathcal{F}_1$ and $\mathcal{F}_2$, such that $\mathcal{F}_1 \subset \mathcal{F}_2$ and $\mathcal{F}_1$ is uniformly drawn from $\mathcal{F}_2$ without replacement. For a loss l bounded by b if the Rashomon ratio is at least*

$$R^{anc}_{ratio}(\mathcal{F}_2, \gamma) \geq 1 - \epsilon^{\frac{1}{|\mathcal{F}_1|}}$$

then for any $\epsilon > 0$, with probability at least $(1-\epsilon)^2$ with respect to the random draw of functions from $\mathcal{F}_2$ to form $\mathcal{F}_1$ and with respect to the random draw of data:

$$|L(f_2^*) - \hat{L}(\hat{f}_1)| \leq \gamma + 2b\sqrt{\frac{\log|\mathcal{F}_1| + \log 2/\epsilon}{2n}},$$

where $f_2^* \in \operatorname{argmin}_{f_2 \in \mathcal{F}_2} L(f_2)$, and $\hat{f}_1 \in \operatorname{argmin}_{f_1 \in \mathcal{F}_1} \hat{L}(f_1)$.

Please refer to Table 2 for possible values of the lower bound on the Rashomon ratio, given $|\mathcal{F}_1|$ and ϵ.

Note that if each model from the Rashomon set has different performance according to the performance function ζ, then Theorem 7 and 9 apply to the pattern Rashomon ratio as well.

Theorems 7 and 9 guarantee that if the true anchored Rashomon set is sufficiently large and if $\mathcal{F}_1$ is selected at random from $\mathcal{F}_2$, then with high probability, the best empirical risk over the simpler space $\mathcal{F}_1$ is close to the best possible true risk over the larger space $\mathcal{F}_2$. The generalization guarantee comes from the size of the simpler space $\mathcal{F}_1$.

The bound shows directly how the size of the Rashomon set could potentially impact generalization guarantees. The intuition for Theorems 7 and 9 holds beyond the case when $\mathcal{F}_1$ is randomly sampled from $\mathcal{F}_2$, it holds, for example, when $\mathcal{F}_1$ covers $\mathcal{F}_2$ sufficiently well. In particular, as the true anchored Rashomon ratio increases, it is more likely that the empirical risk minimum of $\mathcal{F}_1$ will be close to the minimum of the true risk of $\mathcal{F}_2$.

4.1.2 Membership in Anchored Rashomon Sets

Theorems 5, 7, and 9 show the advantage of a large Rashomon set through the possibility of choosing a model from a simpler hypothesis space. However, they have caveats as discussed earlier. As discussed, the theorems' assumption that the true anchored Rashomon set contains a model from a simpler class is unverifiable. Even if it holds, we cannot check it, as we cannot actually compute the true anchored Rashomon set in practice. However, there might be cases when the empirical and the true Rashomon sets are close (e.g., largely overlapping, or one is a cover for the other), and therefore it is beneficial to know the properties of one to understand of the properties of the other. In particular, with high probability, if a fixed model is contained within the anchored Rashomon set, it also belongs to a slightly larger true anchored Rashomon set. The reverse statement holds as well.

Proposition 10 (Empirical anchored Rashomon set is close to true) *For a loss l bounded by b and for any $\epsilon > 0$ with probability at least $1 - e^{-2n(\epsilon/b)^2}$ with respect to the random draw of training data, if $f \in \hat{R}^{anc}_{set}(\mathcal{F}, \gamma)$ then $f \in R^{anc}_{set}(\mathcal{F}, \gamma + \epsilon)$.*

An analogous statement holds for a model from a true anchored Rashomon set:

Proposition 11 (True anchored Rashomon set is close to empirical) *For a loss l bounded by b and for any $\epsilon > 0$, if $f \in R^{anc}_{set}(\mathcal{F}, \gamma)$ then with probability at least $1 - e^{-2n(\epsilon/b)^2}$ with respect to the random draw of training data,*

$$f \in \hat{R}^{anc}_{set}(\mathcal{F}, \gamma + \epsilon).$$

Proposition 11 is based on the same intuition as Lemma 23 in the work of Fisher et al. (2019), which is used to bound the probability with which a given model is not in the empirical Rashomon set; this is used in a proof of a bound for model class reliance. We use the proposition to indicate the probability with which the empirical anchored Rashomon set is as close as possible to the true anchored Rashomon set for a given model.

Propositions 10 and 11 show that, for a fixed model, with high probability, its membership in the true anchored and the anchored Rashomon sets are closely related. Therefore, when we consider a given data set S to compute a model in the (empirical) Rashomon set, we can infer that this model is likely to belong to a related true Rashomon set.

Theorems 5, 7, 9 in this section do not take advantage of the fact that we can investigate $\mathcal{F}_2$ empirically, and more easily than we can investigate $\mathcal{F}_1$; these theorems instead only discuss the exploration of $\mathcal{F}_1$. In what follows, we will study empirical Rashomon sets. Because we are studying empirical Rashomon sets, and because we will work with $\mathcal{F}_2$ instead of $\mathcal{F}_1$, we will need some mechanism to approximate $\mathcal{F}_2$ in terms of $\mathcal{F}_1$ and to ensure that functions from $\mathcal{F}_2$ generalize; for these purposes, we will use smoothness of the loss over functional space.

4.2 Existence of Simple-yet-Accurate Models with Good Generalization

As discussed above, we will now consider empirical Rashomon sets, rather than true Rashomon sets as the assumptions in the earlier theorems are unverifiable, and rely on optimization of the less complex functional space $\mathcal{F}_1$ rather than the more complex functional space $\mathcal{F}_2$. If optimization over $\mathcal{F}_2$ is easier (as it is less constrained), we may want to optimize over $\mathcal{F}_2$ first, and be guaranteed the existence of at least one function in $\mathcal{F}_1$ based on what we observe with $\mathcal{F}_2$. Thus, what we aim to prove in this section is the existence of functions in $\mathcal{F}_1$ that are in the Rashomon set of $\mathcal{F}_2$. In order to do this using an approximating set argument, we use more assumptions than in the previous section, specifically smoothness. The following theorem shows that, under certain conditions, *if* there is a function close to $\mathcal{F}_2$'s minimizer in hypothesis space that is also in $\mathcal{F}_1$, then it is a function we would be looking for: it is also in the Rashomon set of $\mathcal{F}_2$ and it probably generalizes.

For a hypothesis space $\mathcal{F}$ and some $f' \in \mathcal{F}$ let us define the δ-ball of functions centered at f' as $B_\delta(f') = \{f \in \mathcal{F} : \|f' - f\|_p \leq \delta\}$.

A loss $l : \mathcal{F} \times \mathcal{X} \to \mathcal{Y}$ is said to be K-*Lipschitz*, $K \geq 0$, if for all $f_1, f_2 \in \mathcal{F}$ and for all $z \in \mathcal{Z}$: $|l(f_1, z) - l(f_2, z)| \leq K\|f_1 - f_2\|_p$. The p-norm can be defined for example as $\|f\|_p = \left(\int_{\mathcal{X}} |f|^p d\mu\right)^{1/p}$, where μ is a measure on $\mathcal{X}$.

Theorem 12 (Existence of a simpler-but-accurate model I) *For K-Lipschitz loss l bounded by b consider hypothesis spaces $\mathcal{F}_1$ and $\mathcal{F}_2$ such that $\mathcal{F}_1 \subset \mathcal{F}_2$. With probability greater than $1 - \epsilon$ w.r.t. the random draw of training data, if there exists $\bar{f}_1 \in \mathcal{F}_1$ such that $\|\hat{f}_2 - \bar{f}_1\|_p \leq \frac{\theta}{K}$, where $\hat{f}_2$ is the empirical risk minimizer within $\mathcal{F}_2$, then for a fixed parameter $\epsilon \in (0, 1)$:*

1. *$\bar{f}_1$ is in the Rashomon set $\hat{R}_{set}(\mathcal{F}_2, \theta)$.*
2. *$\left|L(\bar{f}_1) - \hat{L}(\bar{f}_1)\right| \leq 2KR_n(\mathcal{F}_1) + b\sqrt{\frac{\log(2/\epsilon)}{2n}}$, where $R_n(\mathcal{F})$ is the standard Rademacher complexity of a functional space $\mathcal{F}$. (This bound arises from standard learning theory.)*

In the case of one local minimum, if that minimum is wide, then it contains many models that perform similarly and thus yields a large Rashomon set. A wide minimum can occur when, for example, the loss is somewhat flat with a bounded derivative. These are often cases where the loss is locally Lipschitz continuous on the Rashomon set with a small Lipschitz constant K, which would create a tighter bound.

Theorem 12 also illustrates how we can use a known Lipschitz constant to choose a Rashomon parameter θ. In particular, if we would like to consider a δ-ball in the hypothesis space $\|\hat{f} - f\|_p < \delta$, then we would choose $\theta = K\delta$, which would keep the bound as small as possible but still permit the result to hold.

Theorem 13 below is a variation on Theorem 12. It still uses the approximating set argument, but also requires the Rashomon set to be large enough to include a ball of functions. As long as the set of simpler functions is distributed well among the full functional space, the ball contains at least one function from the simpler class.

Theorem 13 (Existence of a simpler-but-accurate model II) *For a K-Lipschitz loss l bounded by b, and hypothesis spaces $\mathcal{F}_1$ and $\mathcal{F}_2$ such that $\mathcal{F}_1 \subset \mathcal{F}_2$. With probability greater*

than $1-\epsilon$ *w.r.t. the random draw of training data, if for every model* $f_2 \in \hat{R}_{set}(\mathcal{F}_2, \theta)$ *there exists a model* $f_1 \in \mathcal{F}_1$ *such that* $\|f_2 - f_1\|_p \leq \delta$ *and if the Rashomon set is large, e.g. it contains a ball of size at least* δ*, that is,* $\hat{R}_{set}(\mathcal{F}_2, \theta) \supset B_\delta(\centerdot)$*, then there exists a model* $\bar{f}_1 \in \hat{R}_{set}(\mathcal{F}_2, \theta)$*, such that for a fixed parameter* $\epsilon \in (0, 1)$*:*

1. $\bar{f}_1$ *is from the simpler space* $\mathcal{F}_1$*.*
2. $\left|L(\bar{f}_1) - \hat{L}(\bar{f}_1)\right| \leq 2KR_n(\mathcal{F}_1) + b\sqrt{\frac{\log(2/\epsilon)}{2n}}$*, where* $R_n(\mathcal{F})$ *is the standard Rademacher complexity of a functional space* $\mathcal{F}$*. (This bound arises from standard learning theory.)*

As we have made approximating set arguments several times, with the most recent theorem (Theorem 13) making an assumption that all models from $\mathcal{F}_2$'s Rashomon set are close to a model in $\mathcal{F}_1$, let us discuss this assumption. The field of Approximation Theory provides general conditions under which classes of functions can approximate each other. Given a target function from one space, we want to know whether a sequence of functions from another space can converge to the target. Table 3 shows classes of functions $\mathcal{F}_2$ that can be approximated with functions from classes $\mathcal{F}_1$ within δ using a specified norm. For instance, piecewise constant functions, such as decision trees, can be approximated by smooth functions.

The previous theorems showed the existence of a single function from $\mathcal{F}_1$ with desirable properties. Ideally, we would want multiple functions in $\mathcal{F}_1$ with these properties, since in practice, when we search $\mathcal{F}_1$, we may not find the single function guaranteed by the previous theorem. Theorem 14 below guarantees the existence of more such functions.

For a hypothesis space $\mathcal{F}$, define a ϵ-packing as a finite set $\Xi = \{\xi_1, ..., \xi_k | \xi_i \in \mathcal{F}\}$ such that $\|\xi_i - \xi_j\|_p > \delta$ and $B_{\delta/2}(\xi_i) \cap B_{\delta/2}(\xi_j) = \varnothing$ for all $i \neq j$. The *packing number* $\mathcal{B}(\mathcal{F}, \delta)$ is the largest δ-packing. Then we have the following:

Theorem 14 (Existence of multiple simpler models) *For* K*-Lipschitz loss* l *bounded by* b*, consider hypothesis spaces* $\mathcal{F}_1$ *and* $\mathcal{F}_2$ *such that* $\mathcal{F}_1 \subset \mathcal{F}_2$*. With probability greater than* $1-\epsilon$ *w.r.t. the random draw of training data, if for every model* $f_2 \in \hat{R}_{set}(\mathcal{F}_2, \theta)$ *there exists a model* $f_1 \in \mathcal{F}_1$ *such that* $\|f_2 - f_1\|_p \leq \delta$ *then there exists at least* $B = \mathcal{B}(\hat{R}_{set}(\mathcal{F}_2, \theta), 2\delta)$ *functions* $\bar{f}_1^1, \bar{f}_1^2 ..., \bar{f}_1^B \in \hat{R}_{set}(\mathcal{F}, \theta)$ *such that:*

1. *They are from a simpler space:* $\bar{f}_1^1, \bar{f}_1^2 ..., \bar{f}_1^B \in \mathcal{F}_1$*.*
2. $\left|L(\bar{f}_1^i) - \hat{L}(\bar{f}_1^i)\right| \leq 2KR_n(\mathcal{F}_1) + b\sqrt{\frac{\log(2/\epsilon)}{2n}}$*, for all* $i \in [1, .., B]$*, where* $R_n(\mathcal{F})$ *is the Rademacher complexity of a functional space* $\mathcal{F}$*. (This is from standard learning theory.)*

From Theorem 14, we see that since larger Rashomon sets have larger packing numbers, they therefore contain more simpler models with good generalization guarantees.

Note that in Theorems 12, 13, and 14, other complexity measures from learning theory could be used, such as VC dimension or fat-shattering dimension, to bound the generalization of the hypothesis space $\mathcal{F}_1$. We chose the Rademacher complexity as it provides the tightest bound among standard complexity measures.

$\mathcal{F}_2$	$\mathcal{F}_1$	δ	Source
$f \in L_\infty(\Omega)$, $\|f\|_\infty \in [m, M]$	$s_N \in S(\Omega)$, s_N—piecewise constant, N—number of constants	$\|f - s_N\|_\infty \leq \frac{M-m}{2N}$	DeVore (1998); Davydov (2011)
$f \in W_p^1(\Omega)$, $1 \leq p \leq \infty$, where W_p^1 is a Sobolev space	$s_\Delta(f) \in S(\Omega)$, s_Δ—piecewise constant, Δ—fixed partition, $\Omega = (0,1)^d$, N—number of constants	$\|f - s_\Delta(f)\|_p \leq CN^{-1/d}\|f\|_{W_p^1\Omega}$	Davydov (2011)
$f \in \{x^k, k \in N\}$	$P(n)$—polynomials of degree at most $n \in N$	$\|f - P(n)\|_\infty \leq \frac{1}{2^{k-1}} \sum_{j>(n+k)/2} \binom{k}{j}$	Newman and Rivlin (1976)
$f \in C[0,1]$ is a non-constant symmetric boolean function on $x_1,..,x_n$	$P(d)$—algebraic polynomials of degree d	$\|f - P(d)\|_\infty \leq \mathcal{O}(\sqrt{n(n - \Gamma(f))})$	Paturi (1992)
$f \in Lip_M(\alpha)$, f is Lipschitz continuous with constant M	$N_n : [a,b] \to \mathbb{R}$ is a feedforward neural network with one layer and bounded, monotone and odd defined activation function, $n \in \mathbb{N}$	$\sup_{x\in[a,b]} \|f(x) - N_n(x)\| \leq \frac{5M}{2}\left(\frac{b-a}{n}\right)^\alpha$	Cao et al. (2008)
$f \in L_p(I)$, where $I \subset \mathbb{R}^d$ is a cube in $\mathbb{R}^d$, $\|\cdot\|_{W^r(L_p(I))}$—Sobolev semi norm	P_r—space of polynomials of order r in d, constant C depends on r	$\inf_{p\in P_r} \|f - p\|_{L_p(I)} \leq C\|I\|^{r/d}\|f\|_{W^r(L_p(I))}$	DeVore (1998)

Table 3: Examples of function approximation in different functional spaces: a function from spce $\mathcal{F}_1$ approximates a function in space $\mathcal{F}_2$ with given guarantee δ.

As mentioned briefly earlier, Theorem 14 could have implications in practice, because if our data and algorithm admit a large Rashomon set on a complex space, Theorem 14 suggests that it could be beneficial to locate models from simpler classes within the Rashomon set. These simpler models could be, for instance, models that are constrained to be interpretable. Finding interpretable models can often be computationally demanding, since this generally involves minimizing training loss subject to interpretability constraints, which are often discrete or challenging in other ways. The existence of a large Rashomon set on a more complex space of functions implies that there exist possible many solutions to the constrained optimization problem over the simpler space, and thus it is worthwhile to actually solve this optimization problem over the simpler space. In other words, if the Rashomon set is large, and the other conditions are obeyed, Theorem 14 shows that many interpretable-yet-accurate models would exist, prior to actually finding them.

4.3 Generalization Bound for All Models from the Rashomon Set

By using $\mathcal{F}_1$ as an approximating set for $\mathcal{F}_2$ under smoothness assumptions, we can show that not only do $\mathcal{F}_1$'s functions generalize, but also the entire set of functions within $\mathcal{F}_2$'s Rashomon set generalize. The generalization guarantee uses the complexity measure of the simpler space $\mathcal{F}_1$, rather than that of the more complex space $\mathcal{F}_2$.

The importance of this result, stated formally in Theorem 15, is the implication that good approximating sets mean that the learning problem is inherently lower complexity than a naïve learning theory analysis (which uses $\mathcal{F}_2$'s complexity measure) might suggest. This bound is basic and is a distilled version of standard analyses that use approximating sets.

Theorem 15 (Generalization and reduced complexity of the Rashomon set) *For a K-Lipschitz loss l bounded by b consider two hypothesis spaces $\mathcal{F}_1 \subset \mathcal{F}_2$ such that for any model $f_2 \in \hat{R}_{set}(\mathcal{F}_2, \theta)$ there exists a model $f_1 \in \mathcal{F}_1$ such that $\|f_2 - f_1\|_p \leq \delta$, then for all $f_2 \in \hat{R}_{set}(\mathcal{F}_2, \theta)$ and for any $\epsilon > 0$ with probability at least $1 - \epsilon$:*

$$\left| L(f_2) - \hat{L}(f_2) \right| \leq 2K \left(\delta + R_n(\mathcal{F}_1) \right) + b\sqrt{\frac{\log(2/\epsilon)}{2n}},$$

where $R_n(\mathcal{F})$ is the standard Rademacher complexity of a functional space $\mathcal{F}$.

Theorem 15 shows that if $\mathcal{F}_1$ is a good approximation set for the Rashomon set of $\mathcal{F}_2$, then we can obtain a generalization guarantee for any function from the Rashomon set of the complex space $\mathcal{F}_2$ using only the complexity of the simpler space $\mathcal{F}_1$. As a reminder, Table 3 shows examples of function classes where good approximating sets occur. The approximating set is better when δ is small, and this is when the generalization bound is tighter.

A large Rashomon set is thus a certificate of better generalization, because it allows us to find models from a simpler space, and/or use complexity measures of a simpler space. Despite the connection of the Rashomon set to generalization, the size of the Rashomon set is not the same as the typical complexity measures of functional spaces. Section 6 relates the size of the Rashomon set to several established complexity measures. Before that, in Section 5, we discuss computation.

5. Computation of the Rashomon Ratio and Volume

Often in practice, we do not need to compute the Rashomon ratio, as there is a practical way to check if the Rashomon ratio could be large as we discuss in Section 7. However there are cases when we actually can derive a closed-form solution for the Rashomon volume or estimate the Rashomon ratio using sampling techniques. We discuss these methods of the Rashomon ratio computation in this section.

The Rashomon ratio can be viewed as a probability that a model is contained within the Rashomon set, taking a prior over the hypothesis space. In Equation 1 we compute this probability by comparing the volume of the Rashomon set to that of the full hypothesis space. We can compute this probability directly by sampling models from the hypothesis

space, and taking the fraction of times the sample lies inside the Rashomon set. By Hoeffding's inequality, this rejection sampling procedure has probabilistic guarantees and can be used in cases when the hypothesis space is discrete, e.g., tree structures. We discuss this in Section 5.1.

When the hypothesis space has a parameterized representation, we can compute the Rashomon volume in parameter space. We assume that we can parameterize each model $f \in \mathcal{F}$ in a hypothesis space with a unique, finite number of parameters and denote $f(z) = f_\omega(z)$, where ω is a parameter vector. For a parameter space Ω, the parametric hypothesis is denoted: $\mathcal{F}_\Omega = \{f_\omega(z) : \omega \in \Omega \subseteq \mathbb{R}^p\}$. To be consistent with how the Rashomon set is defined on hypothesis spaces (using similarity directly between functions, rather than similarity between parameters), we will assume that there exists a parametrization such that distances according to the provided metric are similar in both the hypothesis and the parameter space. Otherwise, due to overparametrization or reparametrization, the Rashomon volume can be artificially changed (Dinh et al., 2017). This problem can be avoided by a choice of parameterization that encourages distances in parameter space to be similar to distances in hypothesis space. In Section 5.2, we show how to compute the Rashomon volume in parameter space directly in closed form for ridge and least squares regression.

5.1 Sampling Methods

As before, assume that there exists a prior distribution ρ over the hypothesis space $\mathcal{F}$. From the definition, the Rashomon ratio can be computed as a probability of a model being in the Rashomon set:

$$\hat{R}_{ratio}(\mathcal{F}, \theta) = P[f \in \hat{R}_{set}(\mathcal{F}, \theta)] = \mathbb{E}_{f\sim\rho}\, \mathbb{1}_{[f\in\hat{R}_{set}(\mathcal{F},\theta)]}\,.$$

To approximate the Rashomon ratio, we perform rejection sampling with replacement. In particular, after k draws from distribution ρ,

$$\hat{P}[f \in \hat{R}_{set}(\mathcal{F}, \theta)] = \frac{1}{k}\sum_{i=1}^{k} \mathbb{1}_{[f\in\hat{R}_{set}(\mathcal{F},\theta)]}\,.$$

By Hoeffding's inequality: $P\left(|\hat{P}[f \in \hat{R}_{set}(\mathcal{F}, \theta)] - P[f \in \hat{R}_{set}(\mathcal{F}, \theta)]| \geq t\right) \leq 2e^{-2kt^2}$, or alternatively $1-\alpha = P\left(|\hat{P}[f \in \hat{R}_{set}(\mathcal{F}, \theta)] - P[f \in \hat{R}_{set}(\mathcal{F}, \theta)]| < t\right) \leq 1-2e^{-2kt^2}$. Then, in order to estimate the Rashomon ratio $P[f \in \hat{R}_{set}(\mathcal{F}, \theta)]$ to within t, with a $(1-\alpha)$ confidence interval, we need to sample at least $k \geq \frac{\log 2/\alpha}{2t^2}$ hypotheses from $\mathcal{F}$. The guarantees from this rejection sampling approach are tight enough to be used in practice, and can be used in most hypothesis spaces where hypotheses can be randomly generated.

There are cases when the Rashomon set is very small and therefore rejection sampling contributes very little to the Rashomon ratio approximation. It makes sense to draw models from the region around the Rashomon set instead of from the set of all reasonable models. Importance sampling allows us to sample from an alternative distribution, namely the *proposal distribution*, that is concentrated on the importance region. However, after sampling is done, we need to adjust the weight of the sample to match the probability of sampling it

from the original distribution. Let ρ be a target distribution over the hypothesis space $\mathcal{F}$ and q be a proposal distribution that is focused around the Rashomon set. We can estimate the Rashomon ratio through importance sampling as follows:

$$\hat{R}_{ratio}(\mathcal{F},\theta) = \mathbb{E}_{f\sim q}\frac{\rho(f)}{q(f)} \times \mathbb{1}_{[f\in\hat{R}_{set}(\mathcal{F},\theta)]}.$$

In Section 7 and Appendix H we discuss how to design a target distribution for the hypothesis space of decision trees.

5.2 Analytical Calculation of Rashomon Volume for Ridge Regression

A special case of when the Rashomon volume can be computed in closed form in a parameter space is ridge regression. For a space of linear models $\mathcal{F}_\Omega = \{\omega^T x, \omega \in \mathbb{R}^p\}$, ridge regression chooses a parameter vector by minimizing the penalized sum of squared errors for a training data set $S = [X, Y]$:

$$\min_\omega \hat{L}(\omega) = \min_\omega (X\omega - Y)^T(X\omega - Y) + C\omega^T\omega, \tag{3}$$

where the optimal solution of the ridge regression estimator is $\hat{\omega} = (X^TX + CI_p)^{-1}X^TY$.

Geometrically, the optimal solution to ridge regression will be a parameter vector that corresponds to the intersection of ellipsoidal isosurfaces of the sum of squares term and a hypersphere centered at the origin, with the regularization parameter C determining the trade off between the loss and the radius of the sphere. More generally, isosurfaces of the ridge regression loss function are ellipsoids, and the volume of such an ellipsoid corresponds to the Rashomon volume. Using this geometric intuition, we compute the Rashomon volume in parameter space by the following theorem:

Theorem 16 (Rashomon volume for ridge regression) *For a parametric hypothesis space of linear models $\mathcal{F}_\Omega = \{f_\omega(x) = \omega^T x, \omega \in \mathbb{R}^p\}$ and a data set $S = X \times Y$, the Rashomon set $\hat{R}_{set}(\mathcal{F}_\Omega, \theta)$ of ridge regression is an ellipsoid, containing vectors ω such that:*

$$(\omega - \hat{\omega})^T\frac{X^TX + CI_p}{\theta}(\omega - \hat{\omega}) \leq 1,$$

and the Rashomon volume can be computed as:

$$\mathcal{V}(\hat{R}_{set}(\mathcal{F}_\Omega,\theta)) = J(\theta,p)\prod_{i=1}^{p}\frac{1}{\sqrt{\sigma_i^2 + C}}, \tag{4}$$

where σ_i are singular values of matrix X, $J(\theta,p) = \frac{\pi^{p/2}\theta^{p/2}}{\Gamma(p/2+1)}$ and $\Gamma(\cdot)$ is the gamma function.

Note that for least squares regression, we can use results of Theorem 16 with penalization constant $C = 0$. When features in matrix X are linearly dependent, some singular values will be 0. In this case, the volume of the Rashomon set based on Equation 4 goes to infinity. We avoid this problem by making sure that the Gram matrix of the feature matrix X is always positive definite, meaning that all singular values are non-zero. One way to ensure

this is to perform principal component analysis and use the most significant components as replacement features in order to reduce the hypothesis space prior to learning. We follow this technique in our experiments in Section 8.3.

Interestingly, from Theorem 16, it follows that for ridge regression, *the Rashomon volume depends on the feature space only and does not depend on the regression targets* Y. Indeed, assume that every parameter vector ω such that $f_\omega \in \hat{R}_{set}(\mathcal{F}_\Omega, \theta)$ can be represented as $\omega = \hat{\omega} + \delta$. By a simple transformation we have that $\hat{L}(f_\omega) - \hat{L}(f_{\hat{\omega}}) = \delta^T X^T X \delta$, meaning that if we take a step in parameter space, the empirical risk difference will depend only on the feature space and the step itself, and not on the targets of the problem. This observation can help us choose the parameter θ as $\theta = \delta^T X^T X \delta$ if we want to ensure some dependence between the optimal model $\hat{\omega}$ and a model of interest ω. Then, by choosing the direction as $\delta = \omega - \hat{\omega}$ we can compute the Rashomon parameter θ.

For other algorithms, the Rashomon volume generally depends on the targets; in that sense, ridge regression is unusual.

In Appendix D.1.3 we discuss lower bounds on the Rashomon volume that follow from Theorem 16. These bounds do not depend on the singular values of the feature matrix X and might be easier to compute in practice in order to estimate the Rashomon volume faster.

The cases we considered in this section restrict the structure or properties of the learning problem, but these restrictions allow us to compute the Rashomon volume directly, or estimate the Rashomon ratio with high probability. In Appendix D we further discuss how to approximate the Rashomon volume to any pre-specified precision when the Rashomon set is convex. We also provide an optimization problem to under-approximate the Rashomon volume in the parameter space for support vector machine classification.

We will use both sampling techniques and the ridge regression closed-form solution for the experiments in later sections.

Over the decades, the statistical learning theory community developed beautiful measures that show the expressive power and richness of hypothesis spaces, and how they relate to data and algorithms. The most popular are VC dimension, algorithmic stability, geometric margins, and Rademacher complexity. The Rashomon ratio is different from all of these well-known complexity measures: we can find cases where there is no correspondence between them. In Section 6, we illustrate that there exist data sets and distributions that illuminate differences between the Rashomon ratio and the standard complexity measures.

6. Rashomon Ratio as Compared to Simplicity Measures from Learning Theory

The Rashomon ratio can give insight into the simplicity of a learning problem, but it is different from well-known complexity measures from learning theory. The Rashomon ratio depends on a loss function, the hypothesis space, and a data set, while the majority of other measures are either data set agnostic or focus on properties of a specific model in the space. We will compare the Rashomon ratio to different quantities that are used for generalization in statistical learning theory, including VC dimension (Vapnik and Chervonenkis, 1971), the stability of a learning algorithm (Bousquet and Elisseeff, 2002), geometric margins (Schapire et al., 1998; Burges, 1998), Rademacher complexity and local Rademacher

complexity (Bartlett et al., 2005), because the Rashomon ratio is both similar and different to each of these measures. We will use demonstrations to show the differences.

VC dimension. Vapnik-Chervonenkis (VC) dimension (Vapnik and Chervonenkis, 1971) and the Rashomon set are completely different concepts in terms of data-dependence. The VC dimension is the cardinality of the largest set of points that the learning algorithm can shatter. The hypothesis space shatters a set of points if it can achieve any possible target labeling on this set. In other words, the VC dimension shows the expressive power of a hypothesis space for *any* data set including *an extreme* arrangement of data points and labels. On the contrary, the Rashomon set depends on an empirical risk minimizer that we compute directly for a specific data set, which may not be extreme.

Algorithmic stability. The main motivation for algorithmic stability theory is to ensure robustness of a learning algorithm. Following Bousquet and Elisseeff (2002), we define the hypothesis stability of a learning algorithm as follows.

Definition 17 (Hypothesis stability) *A learning algorithm $\mathcal{A}$ has β hypothesis stability with respect to the loss l if for all $i \in \{1, ..., n\}$,*

$$\mathbb{E}_{S,z}[|l(f_S, z) - l(f_{S^{\setminus i}}, z)|] \leq \beta,$$

where $\beta \in R_+$, hypothesis f_S is learned by an algorithm $\mathcal{A}$ on a data set S, loss $l(f_S, z) = \phi(f_S(x), y)$ for $z = (x, y)$, data set $S = \{z_1, ...z_n\}$, and $S^{\setminus i}$ is modified from the training data by removing the i^{th} element of the data set: $S^{\setminus i} = \{z_1, ..., z_{i-1}, z_{i+1}, ...z_n\}$.

In Section 5.2 we showed that in the case of linear least squares regression, the Rashomon volume depends on features of X only, and does not depend on regression targets Y. In contrast, hypothesis stability depends heavily on Y. In fact, if we can control how we change the set of targets, hypothesis stability can be made to change by an arbitrarily large amount—the Rashomon volume is fundamentally different from hypothesis stability. This is formalized in Theorem 18.

Theorem 18 (Rashomon ratio and algorithmic stability) *Consider a distribution P_X over a discrete domain $\mathcal{X} = \{x_1, ...x_N\}$ and a learning algorithm A that minimizes ridge regression's empirical risk $\hat{L}$ for a linear hypothesis space $\mathcal{F}_\Omega$, as in Equation (3). For any $\lambda > 0$ there exist joint distributions P_{X,Y_1} and P_{X,Y_2} where for $\mathbf{X}$ drawn i.i.d. from P_X, $\mathbf{Y}_1$ is drawn from $P_{Y_1|\mathbf{X}}$ over $\mathcal{Y} \,|\, \mathcal{X}$ and $\mathbf{Y}_2$ is drawn from $P_{Y_2|\mathbf{X}}$ over $\mathcal{Y} \,|\, \mathcal{X}$, such that the expected Rashomon ratios are the same:*

$$\mathbb{E}_{P_{X,Y_1}}[R_{ratio_{\mathbf{S}_1}}(\mathcal{F}_\Omega, \theta)] = \mathbb{E}_{P_{X,Y_2}}[R_{ratio_{\mathbf{S}_2}}(\mathcal{F}_\Omega, \theta)],$$

yet hypothesis stability constants are different by an arbitrarily chosen value of λ:

$$\tilde{\beta}_2 - \tilde{\beta}_1 \geq \lambda,$$

where $\mathbf{S}_1$ and $\mathbf{S}_2$ denote data sets $\mathbf{S}_1 = [\mathbf{X}, \mathbf{Y_1}]$ and $\mathbf{S}_2 = [\mathbf{X}, \mathbf{Y_2}]$, $\tilde{\beta}_1$ is the hypothesis stability coefficient of algorithm $\mathcal{A}$ for distribution P_{X,Y_1} and $\tilde{\beta}_2$ is the hypothesis stability coefficient for distribution P_{X,Y_2}.

Geometric margin. Intuitively both the Rashomon ratio and the width of the geometric margin are data-dependent and show how expressive the hypothesis space is with respect to a given data set. However, the margin depends on the closest data points to the decision boundary (e.g., support vectors), while the Rashomon set does not necessarily rely on the support vectors and may depend on the full data set. Theorem 19 summarizes this idea.

Before stating the theorem, we provide a definition of the margin. For the parametric hypothesis space of linear models $\mathcal{F}_\Omega = f(x) = \omega^T x, \omega \in \mathbb{R}^p$ and binary classification, denote d_+ and d_- as the shortest distances from a decision boundary to the closest points with targets $y = 1$ and $y = -1$ respectively. Then the margin d is a sum of these distances $d = d_+ + d_-$ (Burges, 1998). Moreover, for the model $f_{\hat{\omega}}$ that maximizes the margin, the margin width is $\frac{2}{\|\hat{\omega}\|_2}$.

Theorem 19 (Rashomon ratio and geometric margin) *For any fixed $0 < \lambda < 1$, there exists a fixed hypothesis space $\mathcal{F}_\Omega$, a Rashomon parameter θ, and there exist two data sets S_1 and S_2 with the same empirical risk minimizer $\hat{f} \in \mathcal{F}_\Omega$ such that the width of the geometric margin d is the same for both data sets, yet the Rashomon ratios are different:*

$$|R_{ratio_{S_1}}(\mathcal{F}_\Omega, \theta) - R_{ratio_{S_2}}(\mathcal{F}_\Omega, \theta)| > \lambda.$$

Empirical local Rademacher complexity. The empirical Rademacher complexity is another complexity measure of the hypothesis space. Following Bartlett et al. (2005), for binary classification we define it as follows.

Definition 20 (Empirical Rademacher complexity) *Given a data set S, and a hypothesis space $\mathcal{F}$ of real-valued functions, the empirical Rademacher complexity of $\mathcal{F}$ is defined as:*

$$\hat{R}_n^S(\mathcal{F}) = \frac{1}{n}\mathbb{E}_\sigma \left[\sup_{f \in F} \sum_{i=1}^{n} \sigma_i f(z_i) \right],$$

where $\sigma_1, \sigma_2, \ldots, \sigma_n$ are independent random variables drawn from the Rademacher distribution i.e. $P(\sigma_i = +1) = P(\sigma_i = -1) = 1/2$ for $i = 1, 2, \ldots, n$.

Since we are interested only in models that are inside the Rashomon set, in this section we will consider local empirical Rademacher complexity (Bartlett et al., 2005), which is defined using the Rashomon set $\hat{R}_{set}(\mathcal{F}, \theta)$. Empirical Rademacher complexity measures how well the hypothesis space can fit to random assignments of the labels. In contrast, the Rashomon volume is different, as it measures the number of models that are close to optimal. In other words, the Rashomon set benefits from having multiple similar models, while Rademacher complexity treats them as equivalent. In the following theorem, we provide a simple example to show this discrepancy between the two measures.

Theorem 21 (Rashomon ratio and local Rademacher complexity) *For $0 < \lambda < 1$, there exist two data sets S_1 and S_2, a hypothesis space $\mathcal{F}_\Omega$, and a Rashomon parameter θ such that the local Rademacher complexities defined on the Rashomon sets for S_1 and S_2 are the same:*

$$\hat{R}_n^{S_1}\left(\hat{R}_{set}(\mathcal{F}_\Omega, \theta)\right) = \hat{R}_n^{S_2}\left(\hat{R}_{set}(\mathcal{F}_\Omega, \theta)\right),$$

yet the Rashomon ratios are different:

$$\left| R_{ratio_{S_1}}(\mathcal{F}_\Omega, \theta) - R_{ratio_{S_2}}(\mathcal{F}_\Omega, \theta) \right| > \lambda.$$

Pattern Rashomon ratio. The pattern Rashomon ratio defined in (3) is different from both the Rademacher complexity and geometric margins. Intuitively the pattern Rashomon ratio is closer to the Rademacher complexity, as it tries to find the number of models that fit the best under different label permutations; in contrast, the standard multiplicity-based Rashomon ratio is closer to geometric margins (the multiplicity-based Rashomon ratio tends to be larger when the classification margins are larger).

Additionally, there is a straightforward connection between the growth function and the pattern Rashomon ratio. Recall that the *growth function*, or shattering coefficient, is the maximum number of ways any n data points can be classified using functions from the hypothesis space. The connection is that the volume of the hypothesis space measured using pattern distance is exactly the growth function defined on the current data set. More specifically, the pattern Rashomon ratio and the growth function are equivalent under very specific conditions: (i) the Rashomon set is the full hypothesis space (this is unlikely in practice), (ii) we consider classification with 0-1 loss as the performance measure ζ, and (iii) we consider only one data set and do not take supremum over all data sets (as is usual for the growth function).

Now that we have established what we expect to see theoretically from the Rashomon ratio, we move to experiments.

7. Larger Rashomon Ratios Correlate with Similar Performance of Machine Learning Algorithms, and Good Generalization

In this section, we present several observed properties of larger Rashomon ratios.

We would like to measure the Rashomon ratio of a hypothesis space that includes a broad variety of functions, including some that are capable of fitting the data approximately as well as boosted decision trees, support vector machines with gaussian kernels, or other complex machine learning algorithms. Since it is not clear how to perform direct sampling of this class in functional space, we chose to approximate this space by a function class that (1) we could sample from, and (2) would potentially be a good approximating set for the desired hypothesis space. The class we chose was decision trees of depth 7, which is flexible and large, yet easy to sample from. Since decision trees can refine an input space arbitrarily finely as their depth increases, we can view sufficiently deep decision trees as a rich hypothesis space that approximates many other types of hypothesis spaces, including those used by other machine learning methods. Thus, it is conceivable that large Rashomon sets for decision trees translate into large Rashomon sets for the hypothesis spaces we would like to consider in functional space.

To estimate the Rashomon ratio of depth 7 decision trees, we used importance sampling, as discussed in Section 5.1. The proposal distribution assigns the correct labels to the leaves of the tree based on the training data. Since the data are populated on a bounded domain, to grow a tree up to a depth D fully, we make 2^{D-1} splits. For each data set and each

depth we average our results over ten folds for datasets with less than 200 points and over five folds for datasets with more than 200 points, and we sample 250,000 decision trees per fold. We choose the Rashomon parameter θ to be 5%, and, therefore, all the models in the Rashomon set have empirical risk not more than $\hat{L}(\hat{f}) + 0.05$, where $\hat{L}(\hat{f})$ is the lowest achievable empirical risk across all algorithms we considered. (We vary the θ value later; the results did not seem to be sensitive to that choice.)

Separately, we assess whether many different machine algorithms perform similarly on the data set. If many different algorithms with different functional forms and different levels of smoothness perform similarly on a data set, the Rashomon set contains all of these diverse functions. In that case, we conjecture that the Rashomon set could be large. In this first experiment, we test this conjecture, by investigating whether large Rashomon sets (as measured with decision trees of depth 7) correlate with many machine learning methods performing similarly on the same data set. Here, large Rashomon ratios are on the order of 10^{-37}% or 10^{-38}%, whereas small Rashomon ratios are 10^{-40}% or less. We further discuss experimental setup in Appendix H.

Our experiments considered five popular machine learning algorithms: logistic regression, CART, random forests, gradient boosted trees, and support vector machines with RBF kernels. CART, random forests and gradient boosted trees were regularized by varying the tree depth, the minimum number of samples required to split a node, the minimum number of samples required to create a leaf node, and the number of estimators. Support vector machines were tuned by varying the regularization parameter and the kernel coefficient. We used 38 machine learning classification data sets from the UCI Machine Learning Repository (Dua and Graff, 2019), among which 16 have categorical features and 22 have real-valued features. The majority of the data sets are binary classification data sets and we adapted the rest of the data sets to binary classification as well. The number of features varies from 3 to 784, with the majority of the data sets being in the 15—25 feature range. Appendix F contains a description of the data sets we considered.

To recap, we used decision trees of depth 7 to estimate the Rashomon ratio directly, and separately use a variety of machine learning methods on the same data to assess whether large Rashomon ratios correlate with similar training performance across algorithms, as well as good generalization between training and test sets.

7.1 Similar Performance Across Algorithms

Figure 5(a) illustrates the performance of the five machine learning algorithms with the largest Rashomon ratios in the space of decision trees of depth 7. Across all of these cases, larger Rashomon ratios led to approximately similar training results (within $\sim 5\%$ difference between algorithms). Moreover, all of the models chosen by the algorithms generalized well and produce very similar test accuracy (within $\sim 5\%$ difference between training and generalization errors).

Interestingly, the converse statement, that similar performance across different algorithms should lead to large Rashomon sets, does not always hold; sometimes, generalization occurs with small Rashomon ratios. This observation could be explained in several different ways. First, the Rashomon ratio is not the only driver of good generalization performance. Second, even when the Rashomon ratio is a good driver of generalization performance,

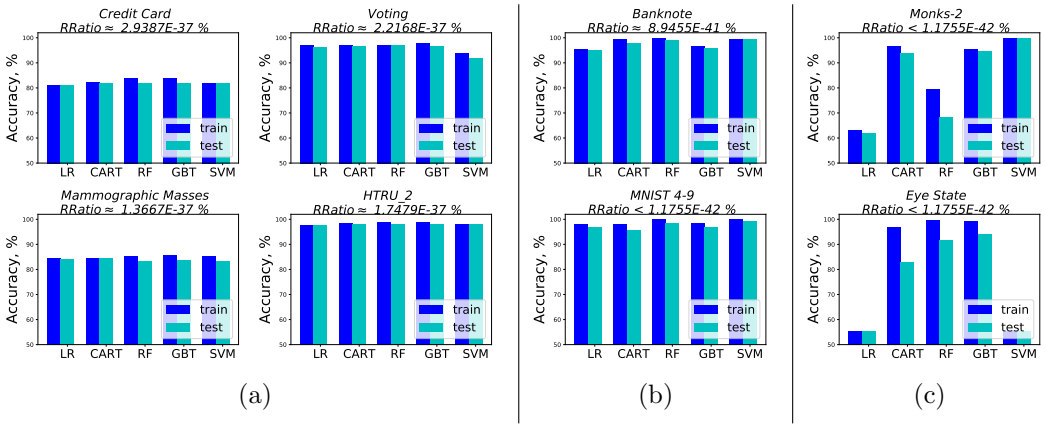


Figure 5: (a) Examples of experiments on four data sets showing that larger Rashomon ratios lead to similar performance of five machine learning algorithms with regularization. All the algorithms generalize well and have similar test accuracy. (b)-(c): Examples showing that smaller Rashomon ratios do not necessarily imply a performance difference between machine learning algorithms. Even with low Rashomon ratios, algorithms can be highly accurate and generalize well, as shown in Figure (b). On the other hand, when the Rashomon ratio is small, sometimes algorithms can perform differently or fail to generalize, as shown in Figure (c). In the figure, test accuracies, training accuracies and the Rashomon ratio are averaged over ten folds.

it may appear artificially small because of a poor representation of data or poor choice of the ratio's denominator. For instance, if the features are highly correlated, this artificially deflates the size of the Rashomon ratio, as discussed in Appendix G. Moreover, if the denominator of the Rashomon ratio (which is the size of the hypothesis space) is poorly designed to include an overly large number of models, then the Rashomon ratio may appear artificially small.

The issues with measuring the Rashomon ratio may be a possible explanation for some of the results in Figure 5(b), which includes some data sets with high-performing algorithms, yet (by the way we measured it) a small Rashomon ratio.

7.2 Good Generalization

In *all* cases we observed, if training performance was consistent across algorithms, test performance was also similar. One thing we notably did not observe were cases where algorithms did not generalize, performance differed across algorithms, and the Rashomon set was large. Actually, we did not observe cases where the Rashomon set was large and performance differed among algorithms. All of these observations are consistent with (but do not definitively prove) the theory that consistent performance across algorithms occur because of large Rashomon sets, which in turn leads to generalization.

Figure 5(c) shows small Rashomon sets and wildly different performance across algorithms, and in that case, sometimes the models generalize and sometimes they do not. We

show one example of each of these cases in Figure 5(c). Our theory does not apply to the case of small Rashomon sets, but again the appearance of small Rashomon sets could be due to poor ways of measuring the size of the Rashomon sets in our experiments.

7.3 Regularization Changes the Hypothesis Space and the Rashomon Set

Regularization limits the hypothesis space and thus changes the nature of the Rashomon set's measurements. Each value of the regularization parameter corresponds to a soft constraint on the hypothesis space, which in turn can be realized as a hard constraint on this space. The Rashomon ratio in the regularized case will typically be larger or equal to the Rashomon ratio in the unregularized case. There are two reasons for this, explained below.

First, regularization reduces the hypothesis space. Hypotheses that were available when learning without regularization may be excluded when learning with regularization. As a result, the volume of the hypothesis space reduces, which decreases the denominator of the Rashomon ratio.

Second, the empirical risk minimizer changes between the regularized and unregularized hypothesis sets, which means the criterion for falling into the Rashomon set changes as well. Recall that the Rashomon set is defined based on the best performing model on the training set. The regularized hypothesis space is less likely to contain overfitted models than the unregularized space. This means the regularized hypothesis space's empirical risk minimizer typically has higher empirical risk than that of the unregularized hypothesis space. Then, if the Rashomon parameter θ is fixed when comparing the two hypothesis spaces, there may be more models in the Rashomon set for the regularized case. Thus, in the regularized case, the Rashomon volume would be larger, and, therefore, the Rashomon ratio would be larger too.

In Appendix H, we show a performance comparison of different machine learning algorithms with and without generalization for all datasets. In both cases, the experiments within Sections 7.1 and 7.2 lead to the same conclusions.

7.4 Diversity Across Algorithms

In Section 7.1, our experimental results suggested that similar performance across machine learning algorithms correlates with larger Rashomon sets. This result implicitly relies on the ability of these algorithms to produce diverse models: it could happen that all of the algorithms produce exactly the same model, in which case our Rashomon set could be small, even if all algorithms perform the same. So how do we know whether our algorithms actually produce diverse results?[2] That is, what assumptions and measurements would we make in order to determine that the Rashomon set is indeed large?

We consider two different sets of assumptions on the structure of the hypothesis space that allow us to show the connection between diversity of models within the Rashomon set and the size of the Rashomon set. In the first case, we estimate the lower bound on the number of diverse models in the Rashomon set for a hypothesis space that itself contains a hierarchy of hypothesis spaces. To do so, we introduce a growth assumption that allows us to navigate through the hierarchy. In the second case, we assume that the hypothesis

2. Thank you to Vipin Kumar for asking this question.

space is diverse enough to contain subspaces of models of different complexity. In that case, the machine learning algorithms that search these subspaces should produce hypotheses of different complexity as well. If these hypotheses realize different patterns in the Rashomon set, then the Rashomon set is diverse. We discuss each of these cases in more detail below.

Hierarchy-based diversity. One general idea for guaranteeing a large Rashomon set requires an assumption and a particular observation. Specifically, we would *observe* that at least two of the algorithms produce models of very different levels of complexity (one simple, one complicated), and that these two models are both in the empirical Rashomon set. We would *assume* that for each simpler class there exist at least C times as many distinct models in a slightly more complicated model class. This creates a Rashomon set that grows at least exponentially in the number of complexity levels between the two observed models. Thus, since the two models we observe are both in the Rashomon set, then there must exist a large number of models in the Rashomon set with complexity in between them. We will formalize the simple conditions described above that allow us to conclude from these models that a large Rashomon set could exist.

Consider a hierarchy of discrete hypothesis spaces $\mathcal{F}_1 \subset \mathcal{F}_2 \subset \cdots \subset \mathcal{F}_T$. Let us make an assumption on models in the Rashomon set of a simple hypothesis space $\mathcal{F}_t$: we assume there exists at least C times more distinct models from a more complicated hypothesis space $\mathcal{F}_{t+1}$, where these models are also in the Rashomon set. We can propagate the growth condition across the hierarchy and compute a lower bound on the Rashomon volume. This intuition is summarized in Proposition 23.

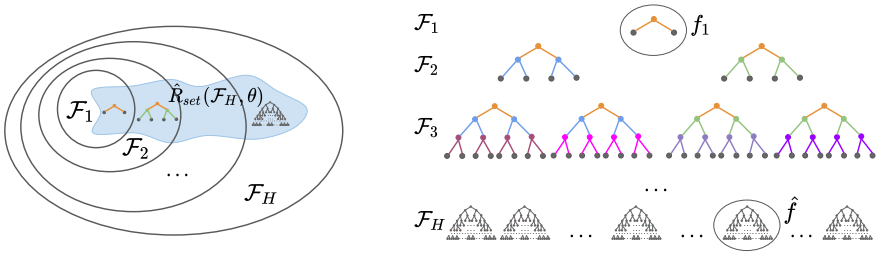
Assumption 22 (Growth assumption) *Consider a hierarchy of discrete hypothesis spaces $\mathcal{F}_1 \subset \mathcal{F}_2 \subset \cdots \subset \mathcal{F}_T$ and a given $B_1 = \{f_1\} \subset \mathcal{F}_1$. Assume that for $t \in [2, \dots T]$, there exist subsets $B_t \subset \mathcal{F}_t$, $B_{t-1} \cap B_t = \varnothing$ such that $\frac{|B_t|}{|B_{t-1}|} > C > 1$ and $\hat{L}(f_{t-1}) \geq \hat{L}(f_t)$, where $|\cdot|$ denotes set cardinality.*

This assumption leads directly to the following proposition.

Proposition 23 *For a hierarchy of discrete hypothesis spaces $\mathcal{F}_1 \subset \mathcal{F}_2 \subset \cdots \subset \mathcal{F}_T$ such that there exists $f_1 \in \hat{R}_{set}(\mathcal{F}_T, \theta) \cap \mathcal{F}_1 \neq \varnothing$, if Assumption 22 holds for $B_1 = \{f_1\}$ then the Rashomon set $\hat{R}_{set}(\mathcal{F}_T, \theta)$ contains at least $\frac{C^T - 1}{C - 1}$ models.*

For Proposition 23 to ensure the existence of a large Rashomon set, we would be required to observe the existence of only two models from the hierarchy. One model is an empirical risk minimizer from the most complicated class $f^* \in \mathcal{F}_T$ that defines the Rashomon set. Another model is from the simplest class $f_1 \in \mathcal{F}_1$ and serves as a base model for the growth assumption.

When considering how realistic Assumption 22 is, we might consider the possibility that it approximately holds: perhaps for each class there are actually sometimes more or sometimes less than C functions from the next complexity class. The complexity classes can be defined any reasonable way, including decision trees, where t is the number of leaves, or one could use linear models with t nonzero terms. Similarly, one could use boosted stumps with t being the total number of stumps. Adding a stump could reasonably provide a new, distinct, model that lowers the loss, as in the assumption. In that way, one could reasonably



(a) (b)

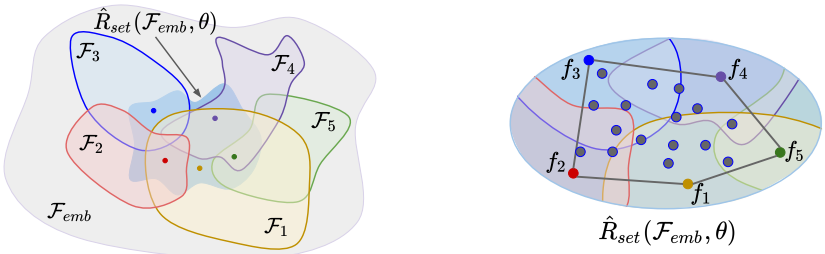
Figure 6: Figure 6(a) illustrates a case where the Rashomon set contains functions across the hierarchy of models. Figure 6(b) shows how a function in f_1, which is a sparse decision tree, can be grown to (gradually) reduce impurities, so that its children stay within the Rashomon set.

expect that if, within the Rashomon set, there exist both simple and complex models, that the Rashomon set is large.

Assumption 22 is complicated to observe, because we would need to lower-bound the count of models in each complexity class. The pattern Rashomon ratio could help with this. While it is not easy to determine whether two functions are the same, it is easy to determine whether they predict the same way on all of the data points. That leads us back to the pattern Rashomon ratio. If we observe a new pattern when we arrive at a complex hypothesis space, the function creating that pattern does not belong to the simpler hypothesis space (we would have seen it when examining that simpler space). Thus, if for every $B_t \subset \mathcal{F}_t$ there exist at least C unique patterns in $\mathcal{F}_{t+1}$ such that $\hat{L}(f_t) \geq \hat{L}(f_{t+1})$, then Assumption 22 holds and the Rashomon set will contain at least $\frac{C^T-1}{C-1}$ models with unique patterns.

Let us consider a more concrete specification of Assumption 22, where we are given the hierarchy of hypothesis spaces as fully-grown decision trees, with respect to depth. In this case, to propagate through the hierarchy, we assume that after each split there will be some correctable impurity remaining in each leaf. In other words, we assume each tree in the Rashomon set of the simpler class will be grown into at least C more distinct trees that are within the next more complicated class. The total number of trees in the Rashomon set for each depth t is at least C^t, which leads to exponential growth. Such a growth assumption for decision trees is a more specific form of Assumption 22, where for each model in a simpler class there exists at least C distinct models in a slightly more complicated model class. An illustration of this is in Figure 6.

Complexity-based diversity. For more general cases (no hypothesis space hierarchy and exponential growth assumption) we can use simple empirical observations. Consider an embedded space $\mathcal{F}_{emb} = \{\mathcal{F}_1, \mathcal{F}_2, \ldots \mathcal{F}_H\}$ that consists of all the hypothesis spaces on which we run different algorithms. Let $H_{alg} \subset \mathcal{F}_{emb}$ be a set of models that different machine learning algorithms output. If we choose hypothesis spaces $\mathcal{F}_1, \mathcal{F}_2, \ldots \mathcal{F}_H$ so that they have different complexity (e.g, VC dimension or Rademacher complexity) then similar performance of models from H_{alg} might indicate a larger Rashomon set that includes



(a) (b)

Figure 7: (a) An illustration of a function class that contains functions of different complexity, arising from the search spaces of different algorithms. (b) An illustration of a convex Rashomon set. Because the Rashomon set is convex, all functions in the convex hull of H_{alg} must be contained within the Rashomon set.

functions from all of these complexity classes, as illustrated in Figure 7(a). Indeed, the random forest models are different from linear models and different from models produced by support vector machines with RBF kernels.

Again, one problem with the above argument is that multiple algorithms might compute the same model. In this case, the Rashomon set might be small. We can use a simple empirical measure based on the average Hamming distance to check how diverse the models are within H_{alg}. The Hamming distance between two vectors of targets is the number of predictions that are different for the two models. Therefore, computing the Hamming distance between pairs of H_{alg} models' predictions and averaging them is an indicator of the diversity of models in H_{alg}.

Let us consider a different special-case assumption than Assumption 22. In particular, if there is some notion of smoothness for functions in the hypothesis space, and if we have found several diverse functions within the Rashomon set, then there are probably more functions in the Rashomon set near them that we did not find. Beyond smoothness, let us additionally assume that the diverse models lie in a single connected component of the Rashomon set. In that case, the existence of diverse models in one connected component of the Rashomon set suggests that the Rashomon set is likely to contain even more models.

For example, let us assume $\hat{R}_{set}(\mathcal{F}_{emb}, \theta)$ is convex. Then, it contains at least the convex hull of models from H_{alg}. (That is, the convex hull of all models found by the different algorithms, all which are within the Rashomon set.) An illustration of this is shown in Figure 7(b). The larger the average Hamming distance is, the more scattered the models from H_{alg} are within the connected component of $\hat{R}_{set}(\mathcal{F}_{emb}, \theta)$, and therefore the larger the Rashomon set might be.

To define smoothness between functions of different complexity, let us assume that hypothesis space $\mathcal{F}_H \in \mathcal{F}_{emb}$ can well-approximate models from other hypothesis spaces. For example, support vector machines with small RBF kernels can approximate other models we consider such as decision forests or gradient boosting trees. Now that all models have a representative in the hypothesis space $\mathcal{F}_H$, we can work only in $\mathcal{F}_H$, defining smoothness in the parameter space of $\mathcal{F}_H$.

To summarize, in this subsection we discussed two cases when models computed by different algorithms are diverse enough to infer the large Rashomon set. The first case is for a hierarchy of hypothesis spaces, that, given models from the simplest and most complicated spaces, requires the growth assumption, leading to exponential growth of the Rashomon set. The second case suggests considering machine learning algorithms that each search hypothesis spaces of different complexity. Empirically, we can compute the average Hamming distance of the computed models to check if they are diverse. If so, we may have a large Rashomon set.

There are several conclusions we can make from our experiments. The most important conclusion is that when the Rashomon ratio is observed to be large, all algorithms perform similarly, and their models tend to generalize. Given these conclusions, one may ask whether it is desirable to aim for the largest possible Rashomon ratio in all cases. In the next section, we introduce *Rashomon curves* and address this question empirically.

8. Rashomon Curves

In this section we introduce a trend, namely the *Rashomon curve*, that we observe experimentally across *all* classification data sets we downloaded from the UCI Machine Learning Repository (Dua and Graff, 2019).

Consider a hierarchy of hypothesis spaces $H_0 \subset H_1 \subset \cdots \subset H_T$, where each $H_t = \{h : \mathcal{X} \to \mathcal{Y}\}$, and $\mathcal{X} = [0,1]^p$ is a unit hypercube. We consider the empirical risk over the loss function as follows $\hat{L}(H) = \min_{h \in H} \frac{1}{n} \sum_{i=1}^{n} \phi(h(x_i), y_i)$.

The *Rashomon curve* is a function from the empirical risk to the log of the Rashomon ratio for a hierarchy of hypothesis spaces. More formally, for a hierarchy of hypothesis spaces $H_0, \dots H_T$, the Rashomon curve is obtained by connecting the following points: $\left(\hat{L}(H_0), \hat{R}_{ratio}(H_0, \theta_0)\right), \dots, \left(\hat{L}(H_t), \hat{R}_{ratio}(H_t, \theta_t)\right), \dots, \left(\hat{L}(H_T), \hat{R}_{ratio}(H_T, \theta_T)\right)$, where $\theta_t = \theta$ is a fixed Rashomon parameter, $t \in [0, T]$.

8.1 Rashomon Curves Tend to Exist Often

For a hierarchy of hypothesis spaces, the Rashomon curve shows that as the size of the hypothesis space grows, the empirical risk of classifiers within the space first decreases and then the Rashomon ratio decreases. As a result, the Rashomon curve has a Γ-shaped trend as illustrated in Figure 8(a).

The horizontal part of the Γ-shape corresponds to a decrease in the empirical risk as we move through the hierarchy of hypothesis spaces. If the hypothesis space with the largest size we consider does not achieve a low value of the empirical risk (e.g., in a case of a complex learning problem) we will observe only the horizontal part of the Rashomon curve. This pattern indicates that none of the hypothesis spaces considered are complex enough to learn the training data well.

The vertical part of the Rashomon curve corresponds to changes in the Rashomon ratio. When a learning problem is easy (e.g., separable data sets, data sets with large margin, data sets with only one or two relevant features), a high accuracy on the training data is easily achievable with a smaller hypothesis space. In that case, we will observe only the

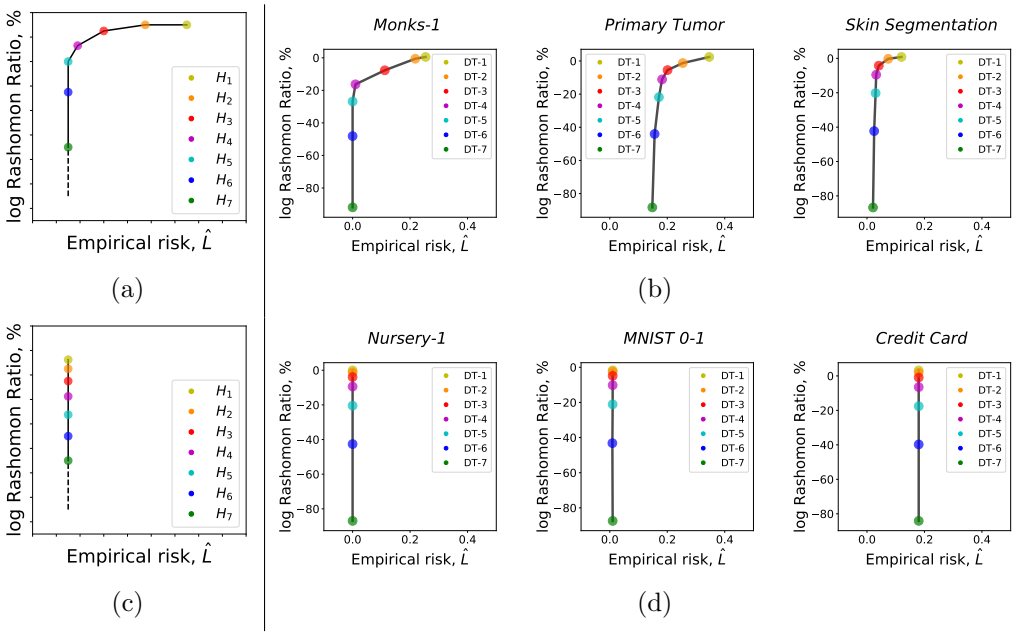


Figure 8: (a) and (c): Illustrations of the Rashomon curve's general shape. For a hierarchy of hypothesis spaces, each space is represented with a colored dot, where different colors corresponds to different hypothesis spaces. As the size of the hypothesis spaces grow, first the empirical risk decreases (horizontal trend) and then the Rashomon ratio decreases (vertical trend). We empirically observed either the full or partial Rashomon curve in every data set we considered. (b) and (d): We show the Rashomon curves for some of the UCI classification data sets. The hierarchy of hypothesis spaces is the set of decision trees from depth one to seven. The Rashomon parameter θ was set to be 0.05 for all experiments. For each hypothesis space H_t in the hierarchy, we say that the model is in the Rashomon set if its empirical risk is less than $\hat{L}(H_t) + \theta$. We average empirical risks and Rashomon ratios over ten folds to plot the Rashomon curve.

vertical part of the Rashomon curve as illustrated in Figure 8(c). The vertical part of the curve corresponds to more complex hypothesis spaces, which is where overfitting can occur. However, the steep drop in Rashomon ratio tends to overwhelm the increases in training accuracy that correspond to overfitting, which is why the curve appears vertical, rather than diagonally slanted.

As before, for our experiments, we used 38 UCI data sets from the UCI Machine Learning Repository. For the hierarchy of hypothesis spaces, we chose fully-grown decision trees up to depth D, where $D \in [1, 7]$. we denote each space by DT-D. We computed the Rashomon ratios by importance sampling of the decision trees for each depth D. We choose the Rashomon parameter θ to be 5%, and all the models in the Rashomon sets have empirical risk not more than $\hat{L}(\hat{f}) + 0.05$, where $\hat{L}(\hat{f})$ is the lowest achievable empirical risk of the hypothesis space at the given level in our hierarchy.

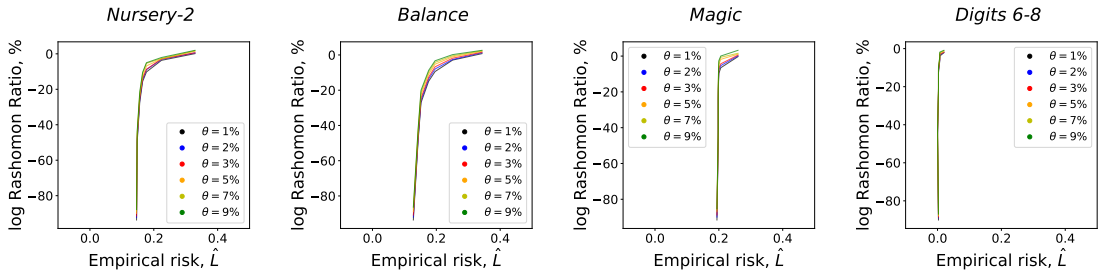


Figure 9: Different values of the Rashomon parameter θ all produce the Γ-shape trend. Empirical risks and Rashomon ratios are averaged over ten folds.

Figure 8 (b),(d) show the Rashomon curves for some categorical and real-valued data sets. *All* of the Rashomon curves, for all 38 data sets as described in Appendix J in Figures 24–28, follow the same trends illustrated in Figure 8 (a),(c). Most of the curves we have observed have a full Γ-shape pattern, while some (e.g., MNIST-0-1, Credit Card) follow only the vertical trend of the Rashomon curve, as in Figure 8 (c). As discussed earlier, this trend indicates a form of simplicity of the data set, and, indeed, Nursery-1 is separable, and others can even be separated with a single decision stump.

As we change the value of the Rashomon parameter θ, the general shape of the Rashomon curves is preserved across all data sets. Figure 9 shows the Rashomon curves for some of the data sets with various values of the Rashomon parameter θ. As we decrease the value of θ, the Rashomon ratio decreases.

Along a Rashomon curve, there is a point that balances between simplicity and empirical error, as illustrated in Figure 10(a). We call this point the *Rashomon elbow* and define it as follows:

Definition 24 (Rashomon elbow) *For a hierarchy of hypothesis spaces $H_0 \subset H_1 \subset \cdots \subset H_T$ the Rashomon elbow is a hypothesis space H_e that both minimizes the empirical risk and maximizes the Rashomon ratio:*

$$H_e \in \underset{\{H_t : \hat{L}(h|_{[h\in H_t]}) \approx \hat{L}(h|_{[h\in H_T]})\}}{\operatorname{argmax}} \hat{R}_{ratio}(H_t, \theta). \tag{5}$$

Models on the elbow should theoretically be both accurate and simple, and therefore generalize well. As we will discuss later, the elbow model is a good choice for model selection. The Rashomon elbow model can differ from a model selected using typical bias-variance trade-offs. To find the Rashomon elbow model, we use only training data, whereas model selection using a bias-variance trade-off requires test (or validation) data.

Note that when comparing Rashomon sets across a hierarchy, the Rashomon ratios decrease, but the sizes of the Rashomon sets increase, because these Rashomon sets are contained within each other. The Rashomon set for decision trees of depth five is contained within the Rashomon set of depth six, even though the Rashomon ratio at size five is smaller than that of size six; the denominator of the ratio increases faster than the numerator across the hierarchy.

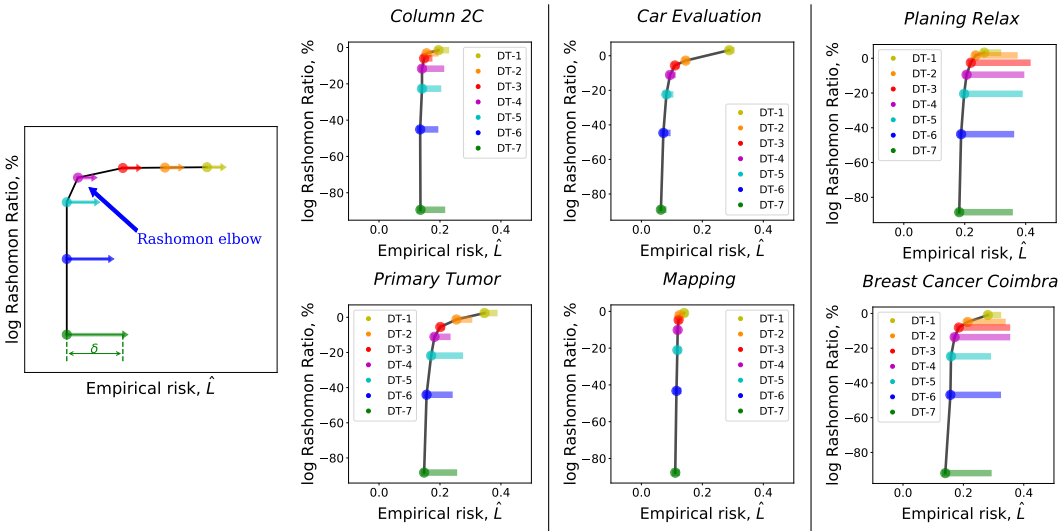


Figure 10: (a): An illustration of the Rashomon elbow and its generalization. For a hierarchy of hypothesis spaces, each hypothesis space is represented with a colored dot. The Rashomon elbow is shown with a blue arrow. Arrows point from the training empirical risk for each hypothesis space to the test risk, where the length of the arrow δ is the generalization error. (b)–(d): The generalization ability of the Rashomon elbow for selected UCI classification data sets. The hierarchy of hypothesis spaces are fully grown decision trees from depth one to seven. The Rashomon parameter θ was set to be 0.05 for all experiments. For each hypothesis space H_t in the hierarchy, we say that the model is in the Rashomon set if its empirical risk is less that $\hat{L}(H_t) + \theta$. In (b), the illustrated curve agrees with theory, whereas in (c), all models generalize, and in (d), no models generalize. We average train and test empirical risks as well as Rashomon ratios over ten folds to plot the generalization curve for data sets with more than 200 points and over five folds with data sets with less than 200 points.

To summarize, the Rashomon curve seems to be a fairly universal trend across data sets, which is that as the size of the hypothesis space increases, empirical risk is decreased until the elbow is reached, after which point, the risk stays approximately constant and the Rashomon ratio rapidly decreases. We will study the curve, its elbow, and their implication to generalization in the next section.

8.2 Rashomon Elbow and Empirical Generalization Properties

The Rashomon elbow, as illustrated in Figure 10(a), is a balancing point between low-error empirical performance and large Rashomon ratio. Let us now consider the generalization error for each of the hypothesis spaces that are represented as arrows on Figure 10(a). Notice

that the generalization error at the elbow is the lowest among all larger-sized hypothesis spaces achieving high accuracy.

Figure 10(a) is an idealized curve that abstractly represents a trend that we observed largely, but not universally, in the data: see the Rashomon elbow generalization error for the UCI-data sets we considered in Figure 10(b-d). From the figures we can divide the data sets into three categories. For the first category (Figure 10(b)) the Rashomon elbow generalizes similarly to Figure 10(a). We can see, for example, on the Column 2C data set, that all models starting from DT-4 overfit. The second category (Figure 10(c)) shows approximately the same generalization error across all complexities of the hypothesis spaces, meaning that the elbow model is still a good choice because it yields simpler models and generalizes as well as the most complex models. The third category (Figure 10(d)) shows large generalization errors across all of the hypothesis spaces; again the elbow model is not worse than all other models considered. Thus, we seem to find that the elbow model selection criterion either helps, or has no effect, but never achieves worse performance than other possible choices.

The Rashomon elbow determines a useful trade-off between generalization and estimation error. Figure 10 illustrates why the Rashomon elbow model might be a good choice for model selection—it often achieves the lowest test error, as discussed in Section 4. The elbow model's class is the smallest (simplest) among hypothesis spaces that achieves low training empirical risk. As we showed empirically in Figure 9, the location of the Rashomon elbow is not particularly sensitive to the choice of θ. For a hierarchy of fully-grown decision trees DT-D, $D \in [1, 7]$, Figure 8 shows the Rashomon curves. The Rashomon elbow model tends to have both high test accuracy and high Rashomon ratio, e.g., DT-4 for Monks-1, DT-3 for Skin Segmentation, DT-1 for Nursery-2 data.

Possible ways to formulate the optimization problem to find the elbow are in Appendix K.

In cases when the Rashomon ratio is complicated to compute, we can approximately find the elbow model based on changes in the empirical risk as we vary the complexity of the hypothesis space. In particular, consider a hierarchy of hypothesis spaces $H_0 \subset H_1 \subset \cdots \subset H_T$ and corresponding empirical risks $\hat{L}(H_0), \hat{L}(H_1), \ldots, \hat{L}(H_T)$ for the best models in these classes. Starting with the most complicated hypothesis space H_T and decreasing the size of the hypothesis space H_t, we stop when there is a jump in the empirical risk $\hat{H}_t$. The hypothesis space before this significant increase in the empirical risk, which we denote as $H_{\bar{e}}$, has the smallest size among $H_T, \ldots H_{\bar{e}}$ and thus is near the top of the vertical trend of the Rashomon curve. Moreover, $H_{\bar{e}}$ has low empirical risk and, therefore, is approximately at the Rashomon elbow.

8.3 Rashomon Curves for Ridge Regression

The Rashomon curve trend holds beyond classification problems, and here we show that it holds for ridge regression as well. In particular, Figure 11 shows the Rashomon curves for a hierarchy of polynomial hypothesis spaces, where the Rashomon volume was plotted against the empirical risk. For the Rashomon volume computation, we used results of Theorem 16, which allows an analytical computation of the volume. The formula depends on the feature matrix, the Rashomon parameter and the number of features in the data

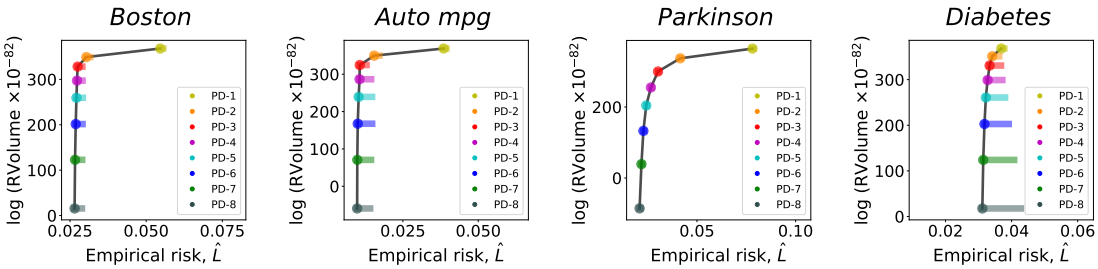


Figure 11: The generalization ability of the Rashomon elbow for the UCI regression data sets with real features. We consider 3 features for every data set with the largest corresponding singular values. The hierarchy of hypothesis spaces consists of polynomials of degree (PD) from one to eight. The Rashomon parameter θ was set to be $0.1\hat{L}(\hat{f}_t)$, $t \in [1, 8]$ for all experiments. The regularization parameter for ridge regression was set to 0.01. We average train and test accuracies as well as Rashomon volumes over ten folds.

set. For our experiments, we choose fourteen real-valued regression data sets from the UCI repository and we choose the three first principal components to form new features that are not redundant with other features. Then, to create a hierarchy of hypothesis spaces, we applied a polynomial transformation to these three features for polynomials of degree 1 through degree 8. As in the case of classification, we see that the Rashomon curve pattern holds, the elbow model exists, and it produces competitive generalization error compared with higher-dimensional spaces.

Although it was possible to compute Rashomon ratios analytically for ridge regression and to estimate them through sampling for decision trees, exhaustive computation of an entire Rashomon curve may not be a practical model selection technique in most cases. We discuss more practical implications in the next section.

9. Practical Implications of the Rashomon Sets and Ratios

We begin by recalling the main conclusions from this paper that would be most impactful for practitioners:

- Large Rashomon sets can embed models from simpler hypothesis spaces (Section 4).
- Similar performance across different machine learning algorithms may correlate with large Rashomon sets (Section 7).
- Large Rashomon sets correlate with existence of models that have good generalization performance (Section 7, 8).
- The size of the Rashomon set is a measure of model class complexity that trades off with training loss to form Rashomon curves (Section 8).

How can a machine learning practitioner benefit from these insights? Let us consider a researcher conducting a standard set of machine learning experiments in which the performance of several different algorithms are compared, and generalization is assessed.

Although it may may not be desirable to compute an entire Rashomon curve explicitly, some commonly occurring scenarios can give an insight into where we are on the Rashomon curve. Consider the possible scenario where all algorithms perform similarly, and when their models tend to generalize well. Since we can assess generalization ability on validation data, we can determine directly whether models generalize. As we discussed in Section 7.4, we can also determine whether all the models form a diverse set by considering the model forms that each algorithm produces. For instance, a random forest model has a different form than a single decision tree: they are both piecewise-constant, but forests have many more pieces. Forests and trees have a different form than a support vector machine model with Gaussian kernels, which is smooth. In the case we are discussing, where the models from the various algorithms are different, yet perform equally well, what we have found is that *there are a large number of different well-performing models. These functions can thus constitute different members of a large Rashomon set in an embedding space* $\mathcal{F}_2 \supset \mathcal{F}_{svm}, \mathcal{F}_{boosting}, \mathcal{F}_{forest}$*, etc. of reasonable models.* Here, $\mathcal{F}_2$ has limited complexity, which permits generalization of the various members of the Rashomon set, as well as other models within $\mathcal{F}_2$.

If, indeed, $\mathcal{F}_2$ exists and has the properties we claim (limited complexity class, large Rashomon set, models achieving highest test performance achievable on that data set), then several doors open. At that point, the researcher could:

Delve in: find specialized models with specific properties, such as interpretability. If the researcher is interested in interpretable models, they can search the large Rashomon set of $\mathcal{F}_2$ to locate simpler models within that set. Based on the result in Section 4, such simpler models are likely to exist in a large Rashomon set of not-too-complex models.

Look up: improve generalization without losing training accuracy. Since all algorithms perform similarly, the algorithms could be producing models along the vertical part of a Rashomon curve of hypothesis spaces. In that case, it is worth looking higher up the Rashomon curve for simpler models that maintain the same training accuracy. This moves the researcher upwards along the curve towards the Rashomon elbow.

In the converse case to the one considered above, the researcher's algorithms perform differently from each other. Based on our experiments in Section 7, our theory bestows none of the advantages listed above in this setting. There are two possible reasons for this. First, the complexity of the hypothesis spaces considered by the researcher may not be adequate for the task. The researcher thus could:

Broaden the horizon: use a more complex model class. If all the algorithms perform differently, the researcher could be choosing models along the horizontal trend of the Rashomon curve. In that case, the simpler models are losing accuracy over the more complex models. This suggests that there still is room to move towards the elbow, by selecting a more com-

plex model class that achieves better performance yet does not overfit.

A second reason for non-uniformity in performance could be that the task might benefit from specialized hypothesis spaces provided by some machine learning algorithms. For example, convolutional neural networks are particularly well-suited to certain types of vision tasks, where they outperform many general-purpose machine learning algorithms. We would not expect uniformity in performance across different algorithms for such tasks.

In the cases considered above, we have shown how an understanding of the Rashomon curve can influence decisions in most cases where a researcher is exploring a data set and iteratively selecting algorithms or hypothesis spaces. As with other fundamental concepts in machine learning, such as the bias-variance tradeoff, an understanding of Rashomon ratios and curves can inform practice even if such quantities are not computed explicitly but are inferred indirectly, such as through experimentation with a variety of algorithms, as we have suggested.

A perspective on modern machine learning applications and Rashomon curves: Recall that the Rashomon set is defined by the interaction between the hypothesis space and the training data. Since algorithms and their performance relative to benchmark data sets change over time as researchers compete on these benchmarks, the placement of problems on the Rashomon curve should not be viewed as something that is static, but that gives insight into the state of the art at a particular point in time. Perhaps the differing perspectives that researchers have on simplicity and generalization are based on *what portion of the Rashomon curve* their algorithms are exploring.

At one time, the MNIST data set was considered a challenging benchmark problem, though accuracies from many modern, general purpose machine learning algorithms are all close to 100%. This suggests that the field is on the vertical part of the Rashomon curve, with no advantage left to be gained from more complex or specialized algorithms. In this case, searching within the Rashomon set for models with other desirable properties, such as simplicity or computational efficiency may be desirable. On the other hand, the newer ImageNet challenge data set has seen increasing performance with increasingly complex model classes in recent years. Perhaps this suggests movement along the horizontal portion of the Rashomon curve towards the elbow, where further gains may yet be obtained.

The above vision examples with relatively high accuracy are in contrast with criminal recidivism prediction, where many different machine learning algorithms have essentially identical performance on many different recidivism prediction problems (Zeng et al., 2017; Tollenaar and van der Heijden, 2013; Angelino et al., 2018), where performance is measured across the full ROC curve. Recidivism prediction exemplifies the case where our theories become relevant: many different algorithms perform similarly, and very sparse interpretable models are as accurate as more complicated models. Interestingly, complicated black box models are still used in the justice system (Angwin et al., 2016), despite the fact that there is little evidence to support the need for this level of complexity (see, e.g., Rudin et al., 2019).

There is evidence that some credit risk assessment problems, and some medical problems, such as stroke risk in atrial fibrillation patients (see, e.g., Letham et al., 2015), diabetes prediction (Razavian et al., 2015), and pneumonia risk and readmission prediction (Caruana

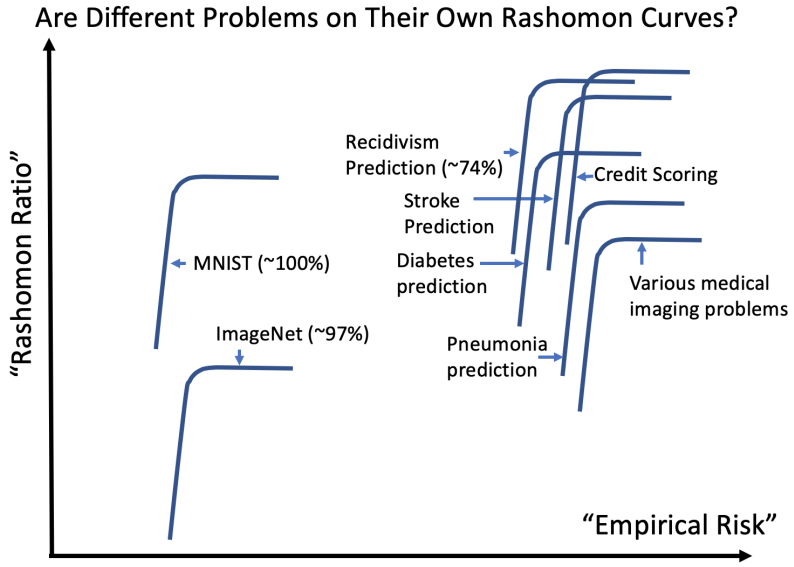


Figure 12: A possible perspective on modern machine learning and Rashomon curves: at a given point in time, the state-of-the-art for different data sets, data types, and algorithmic performance may be viewed as a location on that problem's Rashomon curve. The maximal accuracy for each problem is different, as well as the location and shape of the Rashomon elbow. This figure is just an illustration—the values on the axes are not precise.

et al., 2015) are in a similar state. For these problems, fully interpretable models have been derived that are approximately as accurate as the most accurate (potentially most complicated) machine learning models; perhaps these problems are on the vertical part of the Rashomon curve. Credit risk scoring leads to particularly interesting questions of accuracy-complexity trade-offs: there are financial incentives to keep credit risk models proprietary, which incentivizes the use of potentially overly complex models. Yet, a recent credit risk data set released by the Fair Isaac Corporation (FICO) for the purpose of a data mining competition yielded fully-interpretable models whose accuracy was on par with neural networks and other more complex model classes (Chen et al., 2018).

Figure 12 illustrates the perspective illustrated in this section: if we view current performance as a possible point on a Rashomon curve, it can be useful in determining how to proceed forward with analysis, including whether to delve in, look up, or broaden the horizon.

Conclusion and Future Work

This work studies the Rashomon set, which is the set of almost-equally-accurate models. It also studies the Rashomon ratio, which is the fraction of models that are in the Rashomon set. A large Rashomon set serves as certificate of existence for simpler-yet-accurate models, for a given data set and hypothesis space. Although similar to complexity or simplicity

concepts in machine learning, these Rashomon sets and ratios have key differences that lead to new insights. In cases where many different algorithms have similar and good performance, we hypothesize that a large Rashomon ratio may be the cause.

We also introduce the Rashomon curve, which is a function of the empirical risk versus the Rashomon ratio. The Rashomon curve follows a characteristic Γ-shape, and has occurred across all 52 data sets that we considered. The Rashomon curve reveals a behavior of a hierarchy of hypothesis spaces: The accuracy first increases, then stabilizes, after which the Rashomon ratio decreases. The Rashomon curve's elbow model serves as a useful trade-off between performance and simplicity, and empirically tends to generalize well.

In some cases, it may be practical to compute Rashomon ratios analytically or through sampling methods, but it is not essential to compute these quantities to benefit from the insight they provide. Clues, such as the similar performance of many different algorithms, can give insight into where a practitioner is on the Rashomon curve, and these insights can inform subsequent actions such as enriching the hypothesis space, or searching for more convenient models within the space(s) already considered.

There is also room for further work on techniques for estimating the size of the Rashomon set or ratio, either by sampling or in closed form. We provided a closed form solution for ridge regression only, but closed form solutions may exist for other hypothesis spaces. Better methods of computing or estimating these quantities may facilitate through empirical studies for additional machine learning models, which could bolster the empirical observations made in this paper. There are some challenges in calculating the Rashomon ratio discussed in Section 7, specifically that the Rashomon ratio may appear artificially smaller than it actually is, because of poor parameterization or overparameterization. As we attempt to calculate Rashomon ratios for larger functional spaces, we should remember that the Rashomon ratio should be measured in functional space or pattern space—Rashomon ratios computed in parameter space may not always serve as a good substitute. If we choose not to estimate the Rashomon ratio or size of the Rashomon set directly, but instead choose to look at differing performance among algorithms, we have discussed in Section 7.4 that algorithms should search different complexity hypothesis spaces, and large Hamming distance between patterns can indicate the presence of large Rashomon ratios. Determining measures that indicate a large Rashomon set would be a possible direction for future research.

Given that large Rashomon sets have these interesting properties, it would be worth exploring methods that explicitly try to (re)shape the problem to induce large Rashomon sets. Although we are not aware of any work that has directly done this, there are some existing approaches that may be re-interpreted in this way. For example, one practical technique for producing more robust classifiers is to add noise or smoothing to the training data, e.g., applying a slight blur filter to image data before training. This can be seen as flattening the optimization landscape and potentially increasing the size of the Rashomon set. It is also possible that techniques which inject noise directly into parameter space (Hochreiter and Schmidhuber, 1997) could be interpreted as as having a similar effect. The fact that injecting noise into the data set and/or optimization potentially leads to larger Rashomon sets is a possible connection to differential privacy and other types of privacy-preserving computation. One challenge is to determine whether techniques that inject noise actually do increase the Rashomon ratio. Another future direction is to revisit older techniques like that of Hochreiter and Schmidhuber (1997), which inject noise during

optimization; theoretically, if they widen the Rashomon set, they may improve performance in practice.

The theoretical results presented herein are fairly basic and often rely upon quantities that are sometimes difficult to measure in practice. There is room for further theoretical development to establish tighter and more practical bounds that follow from large Rashomon sets or ratios. One possible direction that could strengthen the connection between the theory and the observed trends in the Rashomon curve could develop along the lines explored by Shawe-Taylor et al. (1998), which considers data-dependent hierarchies of hypothesis spaces.

The connection between Rashomon sets and interpretability of models occurs in two places. First, we provided theoretical conditions under which simpler, high performing models may exist when the Rashomon set is large. Second, we hypothesize that in cases where many different algorithms perform well, a large Rashomon set containing simpler or more interpretable models may be in play. Further experimental studies and theoretical development could strengthen and give further insight into these connections.

Kurosawa's window into human nature showed how the same event can be seen through different eyes. Decades of research on learning theory have given a variety of perspectives, such as VC theory or Rademacher complexity, on the relationship between hypothesis spaces and data sets. We have proposed Rashomon sets and ratios as another perspective on this relationship, and we have provided initial theoretical and experimental results showing that this is a unique perspective that may help explain some phenomena observed in practice.

Acknowledgments

We thank Theja Tulabandhula, Aaron Fisher, Zhi Chen, and Fulton Wang for comments on the manuscript.

Appendix A. List of Notation

Please refer to Table 4 for the list of notation and symbols used in the paper.

Appendix B. Proof of Proposition 4

We recall Proposition 4:

Proposition 4 (Approximation guarantees for the pattern Rashomon ratio) *Let D represent the size of the hypothesis space $\mathcal{F}$. For binary classification and sign performance function $\zeta(f,z) = sign(f(x))$, as $D \to \infty$, the pattern Rashomon ratio $\hat{R}^{pat}_{ratio}(\mathcal{F},\theta) \to \bar{R}^{pat} = \frac{\sum_{i\leq\lfloor\theta n\rfloor}\binom{n}{i}}{2^n}$, and for $\theta \leq 1/2$: $\frac{2^{n(H(\theta)-1)}}{\sqrt{8n\theta(1-\theta)}} \leq \bar{R}^{pat} \leq 2^{n(H(\theta)-1)}$, where n is the size of the training data set, and $H(\theta) = -\theta\log_2\theta - (1-\theta)\log_2(1-\theta)$ is the binary entropy.*

Proof Assume the model class becomes arbitrarily flexible, then at some value of D, each possible labeling of points (each pattern) will constitute a separate equivalence class. Then, the total number of all possible patterns, given that we have two classes, will be 2^n. Also, since each possible pattern is realized, there will be one pattern that achieves the best possible accuracy, 100%. Given the Rashomon parameter θ, a classification pattern should produce an accuracy of at least $1-\theta$ in order for its equivalence class of functions to be in the Rashomon set. Therefore, the Rashomon set can tolerate at most $\lfloor\theta n\rfloor$ points to be misclassified, which leads to the pattern Rashomon ratio limit $\hat{R}^{pat}_{ratio}(\mathcal{F},\theta) \to \frac{\sum_{i=0}^{\lfloor\theta n\rfloor}\binom{n}{i}}{2^n}$.

We obtain the upper bound for $\bar{R}^{pat}$ based on $\sum_{i\leq\theta n}\binom{n}{i} \leq 2^{H(\theta)n}$ for any fixed $\theta \leq 1/2$ (Galvin, 2014). The lower bound for $\bar{R}^{pat}$ follows from simple observations $\sum_{i=0}^{\lfloor\theta n\rfloor}\binom{n}{i} \geq \binom{n}{\theta n}$ and $\binom{n}{\theta n} \geq \frac{2^{nH(\theta)}}{\sqrt{8n\theta(1-\theta)}}$ (MacWilliams and Sloane, 1977). ■

Appendix C. Proofs for Generalization Results

C.1 Proof of Theorem 5

From the definition of the true anchored Rashomon set, it follows that any model in it is γ-close to the ERM. Simple observation shows that any model in the true anchored Rashomon set is also γ-close to any other model in it and is summarized is the lemma below.

Lemma 25 *For any models $f, f' \in \mathcal{F}$ that are in the true anchored Rashomon set $R^{anc}_{set}(\mathcal{F},\gamma)$ we have $|L(f) - L(f')| \leq \gamma$.*

Proof Consider two models f and f' from the true anchored Rashomon set $R^{anc}_{set}(\mathcal{F},\gamma)$. Let $L(f) = \gamma'$ and $L(f') = \gamma''$. Then if $\gamma' > \gamma''$: $L(f) - L(f') = \gamma' - \gamma'' \leq \gamma' \leq \gamma$, otherwise $L(f') - L(f) = \gamma'' - \gamma' \leq \gamma'' \leq \gamma$. Combining these inequalities, we get the statement of the lemma. ■

Now we recall and provide the proof of Theorem 5:

Theorem 5 (The advantage of a large true anchored Rashomon set I) *Consider finite hypothesis spaces $\mathcal{F}_1$ and $\mathcal{F}_2$, such that $\mathcal{F}_1 \subset \mathcal{F}_2$. Let the loss l be bounded by b, $l(f_2, z) \in [0, b] \;\; \forall f_2 \in \mathcal{F}_2, \forall z \in \mathcal{Z}$. Define an optimal function $f_2^* \in \operatorname{argmin}_{f_2 \in \mathcal{F}_2} L(f_2)$. Let us assume that the true anchored Rashomon set is large enough to include a function from $\mathcal{F}_1$, so there exists a model $\tilde{f}_1 \in \mathcal{F}_1$ such that $\tilde{f}_1 \in R^{anc}_{set}(\mathcal{F}_2, \gamma)$. In that case, for any $\epsilon > 0$ with probability at least $1 - \epsilon$ with respect to the random draw of data:*

$$|L(f_2^*) - \hat{L}(\hat{f}_1)| \leq \gamma + 2b\sqrt{\frac{\log |\mathcal{F}_1| + \log 2/\epsilon}{2n}},$$

where $\hat{f}_1 \in \operatorname{argmin}_{f_1 \in \mathcal{F}_1} \hat{L}(f_1)$.

Proof We apply the union bound and Hoeffding's inequality. The result is that with probability at least $1 - \epsilon$ for every $f_1 \in \mathcal{F}_1$ we have, for a finite hypothesis space $\mathcal{F}_1$:

$$|L(f_1) - \hat{L}(f_1)| \leq 2b\sqrt{\frac{\log |\mathcal{F}_1| + \log 2/\epsilon}{2n}}. \tag{6}$$

Combining this Occham's razor bound with the definition of $f_2^* \in \arg\min_{f \in \mathcal{F}_2} L(f)$ we get that, under the same conditions:

$$L(f_2^*) \leq L(\hat{f}_1) \leq \hat{L}(\hat{f}_1) + 2b\sqrt{\frac{\log |\mathcal{F}_1| + \log 2/\epsilon}{2n}}.$$

By assumption of the theorem, there exists a function $\tilde{f}_1 \in \mathcal{F}_1$ such that $\tilde{f}_1 \in R^{anc}_{set}(\mathcal{F}_2, \gamma)$. Since f_2^* is an optimal model, then $f_2^* \in R^{anc}_{set}(\mathcal{F}_2, \gamma)$ as well. From Lemma 25, $|L(f_2^*) - L(\tilde{f}_1)| \leq \gamma$, which implies $L(\tilde{f}_1) \leq L(f_2^*) + \gamma$. Given that $\hat{f}_1 \in \arg\min_{f \in \mathcal{F}_1} \hat{L}(f)$, and using (6), we get that with probability at least $1 - \epsilon$, we have:

$$\hat{L}(\hat{f}_1) \leq \hat{L}(\tilde{f}_1) \leq L(\tilde{f}_1) + 2b\sqrt{\frac{\log |\mathcal{F}_1| + \log 2/\epsilon}{2n}} \leq L(f_2^*) + \gamma + 2b\sqrt{\frac{\log |\mathcal{F}_1| + \log 2/\epsilon}{2n}}.$$

Combining the previous two equations together we have:

$$\left|L(f_2^*) - \hat{L}(\hat{f}_1)\right| \leq \gamma + 2b\sqrt{\frac{\log |\mathcal{F}_1| + \log 2/\epsilon}{2n}}.$$

■

C.2 Proof of Theorem 6 is in Appendix C.7

We provide the proof of Theorem 6 in Appendix C.7, as we use Proposition 11 in the proof.

C.3 Proof of Theorem 7

Theorem 7 (The advantage of a large true anchored Rashomon set II) *Consider finite hypothesis spaces $\mathcal{F}_1$ and $\mathcal{F}_2$, such that $\mathcal{F}_1 \subset \mathcal{F}_2$ and $\mathcal{F}_1$ is uniformly drawn from $\mathcal{F}_2$ without replacement. Define an optimal function $f_2^* \in \operatorname{argmin}_{f_2 \in \mathcal{F}_2} L(f_2)$. For a loss l*

bounded by b and any $\epsilon > 0$, with probability at least $(1-\epsilon)\,p$ with respect to the random draw of functions from $\mathcal{F}_2$ to form $\mathcal{F}_1$ and with respect to the random draw of data:

$$|L(f_2^*) - \hat{L}(\hat{f}_1)| \leq \gamma + 2b\sqrt{\frac{\log|\mathcal{F}_1| + \log 2/\epsilon}{2n}},$$

where $p = 1 - \frac{\binom{(1-R^{anc}_{ratio}(\mathcal{F}_2,\gamma))|\mathcal{F}_2|}{|\mathcal{F}_1|}}{\binom{|\mathcal{F}_2|}{|\mathcal{F}_1|}} = 1 - \prod_{i=1}^{|R^{anc}_{set}(\mathcal{F}_2,\gamma)|}\left(1 - \frac{|\mathcal{F}_1|}{|\mathcal{F}_2|-|R^{anc}_{set}(\mathcal{F}_2,\gamma)|+i}\right)$, *and* $\hat{f}_1 \in \operatorname{argmin}_{f_1\in\mathcal{F}_1}\hat{L}(f_1)$.

Proof The true anchored Rashomon set $R^{anc}_{set}(\mathcal{F}_2,\gamma)$ has $R^{anc}_{ratio}(\mathcal{F}_2,\gamma)|\mathcal{F}_2|$ models. The probability that at least one of these models is from the hypothesis space $\mathcal{F}_1$ is: $p = 1 - \frac{\binom{(1-R^{anc}_{ratio}(\mathcal{F}_2,\gamma))|\mathcal{F}_2|}{|\mathcal{F}_1|}}{\binom{|\mathcal{F}_2|}{|\mathcal{F}_1|}}$. In the fraction, the numerator is the number of ways we could randomly select $|\mathcal{F}_1|$ models that are outside of the Rashomon set, whereas the denominator is the total number of ways we can select $|\mathcal{F}_1|$ models from $|\mathcal{F}_2|$ at random.

Now with probability $p = 1 - \frac{\binom{(1-R^{anc}_{ratio}(\mathcal{F}_2,\gamma))|\mathcal{F}_2|}{|\mathcal{F}_1|}}{\binom{|\mathcal{F}_2|}{|\mathcal{F}_1|}}$ we can guarantee that the hypothesis space $\mathcal{F}_1$ will contain at least one model from the true anchored Rashomon set, therefore by using Theorem 5 we get the statement of Theorem 7.

Simplifying the the binomial coefficients we get that:

$$\begin{aligned}
1-p &= \frac{\binom{(1-R^{anc}_{ratio}(\mathcal{F}_2,\gamma))|\mathcal{F}_2|}{|\mathcal{F}_1|}}{\binom{|\mathcal{F}_2|}{|\mathcal{F}_1|}} = \frac{((1-R^{anc}_{ratio}(\mathcal{F}_2,\gamma))|\mathcal{F}_2|)!|\mathcal{F}_1|!(|\mathcal{F}_2|-|\mathcal{F}_1|)!}{|\mathcal{F}_1|!((1-R^{anc}_{ratio}(\mathcal{F}_2,\gamma))|\mathcal{F}_2|-|\mathcal{F}_1|)!|\mathcal{F}_2|!} \\
&= \frac{((1-R^{anc}_{ratio}(\mathcal{F}_2,\gamma))|\mathcal{F}_2|)!}{|\mathcal{F}_2|!}\frac{(|\mathcal{F}_2|-|\mathcal{F}_1|)!}{((1-R^{anc}_{ratio}(\mathcal{F}_2,\gamma))|\mathcal{F}_2|-|\mathcal{F}_1|)!} \\
&= \prod_{i=1}^{R^{anc}_{ratio}(\mathcal{F}_2,\gamma)|\mathcal{F}_2|}\frac{(1-R^{anc}_{ratio}(\mathcal{F}_2,\gamma))|\mathcal{F}_2|-|\mathcal{F}_1|+i}{(1-R^{anc}_{ratio}(\mathcal{F}_2,\gamma))|\mathcal{F}_2|+i} \\
&= \prod_{i=1}^{R^{anc}_{ratio}(\mathcal{F}_2,\gamma)|\mathcal{F}_2|}\left(1-\frac{|\mathcal{F}_1|}{(1-R^{anc}_{ratio}(\mathcal{F}_2,\gamma))|\mathcal{F}_2|+i}\right) \\
&= \prod_{i=1}^{|R^{anc}_{set}(\mathcal{F}_2,\gamma)|}\left(1-\frac{|\mathcal{F}_1|}{|\mathcal{F}_2|-|R^{anc}_{set}(\mathcal{F}_2,\gamma)|+i}\right).
\end{aligned}$$

Therefore, alternatively $p = 1-\prod_{i=1}^{|R^{anc}_{set}(\mathcal{F}_2,\gamma)|}\left(1-\frac{|\mathcal{F}_1|}{|\mathcal{F}_2|-|R^{anc}_{set}(\mathcal{F}_2,\gamma)|+i}\right)$, which is an easier expression to compute in practice, especially for large values of $|\mathcal{F}_2|$.

■

C.4 Proof of Theorem 9 via Lemma 8

Theorem 9 follows directly from Lemma 8 and Theorem 5, which guarantees that with high probability the sampled space $\mathcal{F}_1$ will contain at least one model from the true anchored Rashomon set. Now we recall and provide a proof of Lemma 8:

Lemma 8 *For a finite hypothesis space $\mathcal{F}_2$ of size $|\mathcal{F}_2|$, we will draw $|\mathcal{F}_1|$ functions uniformly without replacement from $\mathcal{F}_2$ to form $\mathcal{F}_1$. If the true anchored Rashomon ratio of the hypothesis space $\mathcal{F}_2$ is at least*

$$R^{anc}_{ratio}(\mathcal{F}_2, \gamma) \geq 1 - \epsilon^{\frac{1}{|\mathcal{F}_1|}}$$

then with probability at least $1-\epsilon$ with respect to the random draw of functions from $\mathcal{F}_2$ to form $\mathcal{F}_1$, the Rashomon set contains at least one model $\tilde{f}_1$ from $\mathcal{F}_1$.

Proof

The probability of an individual sample from $\mathcal{F}_2$ missing the true anchored Rashomon set is $1 - R^{anc}_{ratio}(\mathcal{F}_2, \gamma)$. The probability if this happening $|\mathcal{F}_1|$ times independently is $(1-R^{anc}_{ratio}(\mathcal{F}_2, \gamma))^{|\mathcal{F}_1|}$. Thus, for any $\epsilon > 0$, if the Rashomon ratio is at least $R^{anc}_{ratio}(\mathcal{F}_2, \gamma) \geq 1 - \epsilon^{\frac{1}{|\mathcal{F}_1|}}$, the probability p_w of sampling, with replacement, at least one hypothesis from $R^{anc}_{ratio}(\mathcal{F}_2, \gamma)$ is:

$$p_w = 1 - (1 - R^{anc}_{ratio}(\mathcal{F}_2, \gamma))^{|\mathcal{F}_1|} \geq 1 - \epsilon.$$

Let p_i be the probability, under sampling without replacement, that samples $1 \dots i$ have missed $R^{anc}_{ratio}(\mathcal{F}_2, \gamma)$. $p_1 = 1 - R^{anc}_{ratio}(\mathcal{F}_2, \gamma)$, and $p_i \leq (1 - R^{anc}_{ratio}(\mathcal{F}_2, \gamma))^i$. The probability, under sampling without replacement, that at least one hypothesis from $R^{anc}_{ratio}(\mathcal{F}_2, \gamma)$ in $\mathcal{F}_1$ is therefore $1 - p_{|\mathcal{F}_1|} \geq p_w$. Thus the the statement of the lemma holds with the probability at least $1 - \epsilon$. ■

C.5 Proof of Proposition 10

Proposition 10 (Empirical anchored Rashomon set is close to true) *For a loss l bounded by b and for any $\epsilon > 0$ with probability at least $1 - e^{-2n(\epsilon/b)^2}$ with respect to the random draw of training data, if $f \in \hat{R}^{anc}_{set}(\mathcal{F}, \gamma)$ then $f \in R^{anc}_{set}(\mathcal{F}, \gamma + \epsilon)$.*

Proof For a fixed $f \in \hat{R}^{anc}_{set}(\mathcal{F}, \gamma)$, by Hoeffding's inequality:

$$P\left[\hat{L}(f) - L(f) < -\epsilon\right] = P\left[\frac{1}{n}\sum_{i=1}^{n} l(f, z_i) - \mathbb{E}\left[l(f, z)\right] < -\epsilon\right] \leq e^{-2n(\epsilon/b)^2}.$$

Therefore, with probability at least $1 - e^{-2n(\epsilon/b)^2}$ with respect to the random draw of data, $L(f) - \hat{L}(f) \leq \epsilon$.

Since $f \in \hat{R}^{anc}_{set}(\mathcal{F}, \gamma)$, then by definition of the Rashomon set, $\hat{L}(f) \leq \gamma$. Combining this with Hoeffding's inequality, with probability at least $1 - e^{-2n(\epsilon/b)^2}$:

$$L(f) \leq \hat{L}(f) + \epsilon \leq \gamma + \epsilon,$$

therefore $f \in R^{anc}_{set}(\mathcal{F}, \gamma + \epsilon)$. ■

C.6 Proof of Proposition 11

Proposition 11 (True anchored Rashomon set is close to empirical) *For a loss l bounded by b and for any $\epsilon > 0$, if $f \in R^{anc}_{set}(\mathcal{F}, \gamma)$ then with probability at least $1 - e^{-2n(\epsilon/b)^2}$ with respect to the random draw of training data,*

$$f \in \hat{R}^{anc}_{set}(\mathcal{F}, \gamma + \epsilon).$$

Proof For a fixed $f \in R^{anc}_{set}(\mathcal{F}, \gamma)$ by Hoeffding's inequality:

$$P\left[\hat{L}(f) - L(f) > \epsilon\right] = P\left[\frac{1}{n}\sum_{i=1}^{n} l(f, z_i) - \mathbb{E}\left[l(f, z)\right] > \epsilon\right] \leq e^{-2n(\epsilon/b)^2}.$$

Therefore, with probability at least $1 - e^{-2n(\epsilon/b)^2}$ with respect to the random draw of data, $\hat{L}(f) - L(f) \leq \epsilon$.

Since $f \in R^{anc}_{set}(\mathcal{F}, \gamma)$, then by definition of the Rashomon set, $L(f) \leq \gamma$. Combining this with Hoeffding's inequality, we get that with probability at least $1 - e^{-2n(\epsilon/b)^2}$:

$$\hat{L}(f) \leq L(f) + \epsilon \leq \gamma + \epsilon,$$

therefore $f \in \hat{R}^{anc}_{set}(\mathcal{F}, \gamma + \epsilon)$. ■

C.7 Proof of Theorem 6

Theorem 6 (The advantage of a good approximating set) *Consider hypothesis spaces $\mathcal{F}_1$ and $\mathcal{F}_2$, such that $\mathcal{F}_1 \subset \mathcal{F}_2$. Let the loss l be bounded by b, $l(f_2, z) \in [0, b] \quad \forall f_2 \in \mathcal{F}_2, \forall z \in \mathcal{Z}$. Define an optimal function $f_2^* \in \operatorname{argmin}_{f_2 \in \mathcal{F}_2} L(f_2)$. Let us assume that the true anchored Rashomon set is large enough to include a function from $\mathcal{F}_1$, so there exists a model $\tilde{f}_1 \in \mathcal{F}_1$ such that $\tilde{f}_1 \in R^{anc}_{set}(\mathcal{F}_2, \gamma)$. In that case, for any $\epsilon > 0$ with probability at least $1 - \epsilon$ with respect to the random draw of data:*

$$|L(f_2^*) - \hat{L}(\hat{f}_1)| \leq 2\gamma + b\sqrt{\frac{\log 1/\epsilon}{2n}},$$

where as before, $\hat{f}_1 \in \operatorname{argmin}_{f_1 \in \mathcal{F}_1} \hat{L}(f_1)$.

Proof By the assumption of the theorem we have that $L(\tilde{f}_1) \leq \gamma$. Also, by the definition of an optimal model f_1^*, $L(f_1^*) \leq L(\tilde{f}_1)$. Combining these, we get that $L(f_1^*) \leq L(\tilde{f}_1) \leq \gamma$. Thus f_1^* is in the true anchored Rashomon set of $\mathcal{F}_2$, $f_1^* \in R^{anc}_{set}(\mathcal{F}_2, \gamma)$. Following Proposition 11, we have that for any $\epsilon_1 > 0$ with probability at least $1 - e^{-2n(\epsilon_1/b)^2}$ with respect to the random draw of data, f_1^* is in the slightly larger anchored Rashomon set $\hat{R}^{anc}_{set}(\mathcal{F}_2, \gamma + \epsilon_1)$, and therefore, with high probability, $\hat{L}(f_1^*) \leq \gamma + \epsilon_1$. Or alternatively, by setting $\epsilon = e^{-2n(\epsilon_1/b)^2}$ we get that for any $\epsilon > 0$ with probability at least $1 - \epsilon$, we have $\hat{L}(f_1^*) \leq \gamma + b\sqrt{\frac{\log 1/\epsilon}{2n}}$. Further, by definition of the empirical risk minimizer we get:

$$\hat{L}(\hat{f}_1) \leq \hat{L}(f_1^*) \leq \gamma + b\sqrt{\frac{\log 1/\epsilon}{2n}}. \tag{7}$$

On the other hand, from the definition of the Rashomon set, $L(f_2^*) \leq \gamma$. Combining this with (7) we have that:

$$|L(f_2^*) - \hat{L}(\hat{f}_1)| \leq L(f_2^*) + \hat{L}(\hat{f}_1) \leq 2\gamma + b\sqrt{\frac{\log 1/\epsilon}{2n}}.$$

■

C.8 Proof of Theorem 12

Theorem 12 (Existence of a simpler-but-accurate model I) *For K-Lipschitz loss l bounded by b consider hypothesis spaces $\mathcal{F}_1$ and $\mathcal{F}_2$ such that $\mathcal{F}_1 \subset \mathcal{F}_2$. With probability greater than $1-\epsilon$ w.r.t. the random draw of training data, if there exists $\bar{f}_1 \in \mathcal{F}_1$ such that $\|\hat{f}_2 - \bar{f}_1\|_p \leq \frac{\theta}{K}$, where $\hat{f}_2$ is the empirical risk minimizer within $\mathcal{F}_2$, then for a fixed parameter $\epsilon \in (0,1)$:*

1. *$\bar{f}_1$ is in the Rashomon set $\hat{R}_{set}(\mathcal{F}_2, \theta)$.*
2. *$\left|L(\bar{f}_1) - \hat{L}(\bar{f}_1)\right| \leq 2KR_n(\mathcal{F}_1) + b\sqrt{\frac{\log(2/\epsilon)}{2n}}$, where $R_n(\mathcal{F})$ is the standard Rademacher complexity of a functional space $\mathcal{F}$. (This bound arises from standard learning theory.)*

Proof The first result follows directly from the definition of the Rashomon set and Lipschitz continuity:

$$\hat{L}(\bar{f}_1) - \hat{L}(\hat{f}_2) = |\hat{L}(\bar{f}_1) - \hat{L}(\hat{f}_2)| \leq K\|\bar{f}_1 - \hat{f}_2\|_p = K\frac{\theta}{K} = \theta.$$

Using Bartlett and Mendelson's generalization bound for Lipschitz loss functions (Bartlett and Mendelson, 2002) we have that for every model $f_1 \in \mathcal{F}_1$, with probability greater that $1-\epsilon$, $|L(f_1) - \hat{L}(f_1)| \leq 2KR_n(\mathcal{F}_1) + b\sqrt{\frac{\log(2/\epsilon)}{2n}}$. Since $\bar{f}_1 \in \mathcal{F}_1$, the bound holds for it as well. ■

C.9 Proof of Theorem 13

Theorem 13 (Existence of a simpler-but-accurate model II) *For a K-Lipschitz loss l bounded by b, and hypothesis spaces $\mathcal{F}_1$ and $\mathcal{F}_2$ such that $\mathcal{F}_1 \subset \mathcal{F}_2$. With probability greater than $1-\epsilon$ w.r.t. the random draw of training data, if for every model $f_2 \in \hat{R}_{set}(\mathcal{F}_2, \theta)$ there exists a model $f_1 \in \mathcal{F}_1$ such that $\|f_2 - f_1\|_p \leq \delta$ and if the Rashomon set is large, e.g. it contains a ball of size at least δ, that is, $\hat{R}_{set}(\mathcal{F}_2, \theta) \supset B_\delta(.)$, then there exists a model $\bar{f}_1 \in \hat{R}_{set}(\mathcal{F}_2, \theta)$, such that for a fixed parameter $\epsilon \in (0,1)$:*

1. *$\bar{f}_1$ is from the simpler space $\mathcal{F}_1$.*
2. *$\left|L(\bar{f}_1) - \hat{L}(\bar{f}_1)\right| \leq 2KR_n(\mathcal{F}_1) + b\sqrt{\frac{\log(2/\epsilon)}{2n}}$, where $R_n(\mathcal{F})$ is the standard Rademacher complexity of a functional space $\mathcal{F}$. (This bound arises from standard learning theory.)*

Proof Consider a ball $B_\delta(f_2')$ of radius δ centered at f_2' that is contained within the Rashomon set. By the theorem's assumption, since $f_2' \in \hat{R}_{set}(\mathcal{F}_2, \theta)$, there exists $\bar{f}_1 \in \mathcal{F}_1$ such that $\|f_2' - \bar{f}_1\|_p \leq \delta$. Therefore $\bar{f}_1$ is inside the δ-ball $\bar{f}_1 \in B_\delta(f_2')$ and thus belongs to the Rashomon set $\hat{R}_{set}(\mathcal{F}_2, \theta)$.

The generalization bound follows Bartlett and Mendelson (2002) as before. ■

C.10 Proof of Theorem 14

Theorem 14 (Existence of multiple simpler models) *For K-Lipschitz loss l bounded by b, consider hypothesis spaces $\mathcal{F}_1$ and $\mathcal{F}_2$ such that $\mathcal{F}_1 \subset \mathcal{F}_2$. With probability greater than $1-\epsilon$ w.r.t. the random draw of training data, if for every model $f_2 \in \hat{R}_{set}(\mathcal{F}_2, \theta)$ there exists a model $f_1 \in \mathcal{F}_1$ such that $\|f_2 - f_1\|_p \leq \delta$ then there exists at least $B = \mathcal{B}(\hat{R}_{set}(\mathcal{F}_2, \theta), 2\delta)$ functions $\bar{f}_1^1, \bar{f}_1^2 ..., \bar{f}_1^B \in \hat{R}_{set}(\mathcal{F}, \theta)$ such that:*

1. *They are from a simpler space: $\bar{f}_1^1, \bar{f}_1^2 ..., \bar{f}_1^B \in \mathcal{F}_1$.*
2. *$\left|L(\bar{f}_1^i) - \hat{L}(\bar{f}_1^i)\right| \leq 2KR_n(\mathcal{F}_1) + b\sqrt{\frac{\log(2/\epsilon)}{2n}}$, for all $i \in [1, .., B]$, where $R_n(\mathcal{F})$ is the Rademacher complexity of a functional space $\mathcal{F}$. (This is from standard learning theory.)*

Proof Starting from the packing number of the Rashomon set $\mathcal{B}(\hat{R}_{set}(\mathcal{F}_2, \theta), 2\delta)$, there exists a 2δ-packing $\Xi = \{\xi_1, ..., \xi_k | \xi_i \in \hat{R}_{set}(\mathcal{F}_2, \theta)\}$ such that $\|\xi_i - \xi_j\|_p > 2\delta$. On the other hand, for each $\xi_i \in \hat{R}_{set}(\mathcal{F}_2, \theta)$ there exists $\bar{f}_1^i \in \mathcal{F}_1$ such that $\|\xi_i - \bar{f}_1^i\|_p \leq \delta$. Therefore for each ball center ξ_i in the packing there is a distinct model $\bar{f}_1^i$ from the simpler hypothesis space $\mathcal{F}_1$. Thus, the Rashomon set contains at least $B = \mathcal{B}(\hat{R}_{set}(\mathcal{F}_2, \theta), 2\delta)$ models from $\mathcal{F}_1$.

The generalization bound follows Bartlett and Mendelson (2002) as before. ■

C.11 Proof of Theorem 15

Theorem 15 (Generalization and reduced complexity of the Rashomon set) *For a K-Lipschitz loss l bounded by b consider two hypothesis spaces $\mathcal{F}_1 \subset \mathcal{F}_2$ such that for any model $f_2 \in \hat{R}_{set}(\mathcal{F}_2, \theta)$ there exists a model $f_1 \in \mathcal{F}_1$ such that $\|f_2 - f_1\|_p \leq \delta$, then for all $f_2 \in \hat{R}_{set}(\mathcal{F}_2, \theta)$ and for any $\epsilon > 0$ with probability at least $1 - \epsilon$:*

$$\left|L(f_2) - \hat{L}(f_2)\right| \leq 2K\left(\delta + R_n(\mathcal{F}_1)\right) + b\sqrt{\frac{\log(2/\epsilon)}{2n}},$$

where $R_n(\mathcal{F})$ is the standard Rademacher complexity of a functional space $\mathcal{F}$.

Proof The theorem follows from the triangle inequality, the theorem's statement, and generalization bound for K-Lipschitz loss functions from Bartlett and Mendelson (2002):

$$\begin{aligned}|L(f_2) - \hat{L}(f_2)| &\leq |L(f_2) - L(f_1)| + |\hat{L}(f_2) - \hat{L}(f_1)| + |L(f_1) - \hat{L}(f_1)| \\ &\leq K\delta + K\delta + 2KR_n(\mathcal{F}_1) + b\sqrt{\frac{\log(2/\epsilon)}{2n}}.\end{aligned}$$

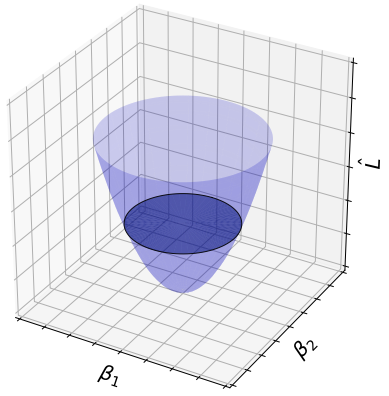


Figure 13: The Rashomon set for the two-dimensional least squares regression. The volume of a shaded ellipsoid corresponds to the Rashomon volume in a parameter space.

■

Appendix D. Approximation of the Rashomon Ratio and Volume

D.1 Rashomon Volume for Ridge Regression

D.1.1 Visualization of the Rashomon Set for Ridge Regression

Consider least squares regression, which is a corner case of ridge regression with the regularization constant $C = 0$. Figure 13 shows the plot of the empirical risk in two dimensional parameter space. Visually, the Rashomon set consists of the those parameters $\omega = [\omega_1, \omega_2]$ that produces a loss below the dark shaded ellipse on the paraboloid, and then Rashomon volume can be computed exactly as the volume of the shaded ellipsoid.

D.1.2 Proof of Theorem 16

Theorem 16 (Rashomon volume for ridge regression) *For a parametric hypothesis space of linear models $\mathcal{F}_\Omega = \{f_\omega(x) = \omega^T x, \omega \in \mathbb{R}^p\}$ and a data set $S = X \times Y$, the Rashomon set $\hat{R}_{set}(\mathcal{F}_\Omega, \theta)$ of ridge regression is an ellipsoid, containing vectors ω such that:*

$$(\omega - \hat{\omega})^T \frac{X^T X + CI_p}{\theta} (\omega - \hat{\omega}) \leq 1,$$

and the Rashomon volume can be computed as:

$$\mathcal{V}(\hat{R}_{set}(\mathcal{F}_\Omega, \theta)) = J(\theta, p) \prod_{i=1}^{p} \frac{1}{\sqrt{\sigma_i^2 + C}},$$

where σ_i are singular values of matrix X, $J(\theta, p) = \frac{\pi^{p/2} \theta^{p/2}}{\Gamma(p/2+1)}$ and $\Gamma(\cdot)$ is the gamma function.

Proof Consider all models $f_\omega \in \mathcal{F}_\Omega$ from the Rashomon set $\hat{R}_{set}(\mathcal{F}_\Omega, \theta)$. Then by Definition 1 we get:

$$\hat{L}(X, Y, \omega) \leq \hat{L}(X, Y, \hat{\omega}) + \theta. \tag{8}$$

Using $X^T Y = (X^T X + C I_p)\hat{\omega}$ from the optimal solution of the ridge regression estimator $\hat{\omega} = (X^T X + C I_p)^{-1} X^T Y$, and expanding the difference between empirical risks we have:

$$\begin{aligned}
\theta \geq & \hat{L}(X, Y, \omega) - \hat{L}(X, Y, \hat{\omega}) \\
= & (X\omega - Y)^T (X\omega - Y) + C\omega^T \omega - (X\hat{\omega} - Y)^T (X\hat{\omega} - Y) - C\omega^{*T}\hat{\omega} \\
= & \omega^T X^T X \omega - 2\omega^T X^T Y + C\omega^T \omega - \omega^{*T} X^T X \hat{\omega} + 2\omega^{*T} X^T Y - C\omega^{*T}\hat{\omega} \\
= & \omega^T X^T X \omega - 2\omega^T (X^T X + C I_p)\hat{\omega} + C\omega^T \omega - \omega^{*T} X^T X \hat{\omega} + 2\omega^{*T} (X^T X + C I_p)\hat{\omega} - C\omega^{*T}\hat{\omega} \\
= & \omega^T X^T X \omega + C\omega^T \omega - 2\omega^T (X^T X + C I_p)\hat{\omega} + \omega^{*T} X^T X \hat{\omega} + C\omega^{*T}\hat{\omega} \\
= & \omega^T (X^T X + C I_p)\omega - 2\omega^T (X^T X + C I_p)\hat{\omega} + \omega^{*T} (X^T X + C I_p)\hat{\omega} \\
= & (\omega - \hat{\omega})^T (X^T X + C I_p)(\omega - \hat{\omega}).
\end{aligned}$$

Therefore the Rashomon set is an ellipsoid centered at $\hat{\omega}$:

$$(\omega - \hat{\omega})^T \frac{X^T X + C I_p}{\theta} (\omega - \hat{\omega}) \leq 1.$$

By the formula of the volume of a p-dimensional ellipsoid the Rashomon volume can be computed as:

$$\mathcal{V}(\hat{R}_{set}(\mathcal{F}_\Omega, \theta)) = \frac{\pi^{p/2} \theta^{p/2}}{\Gamma(p/2 + 1)} \prod_{i=1}^{p} \frac{1}{\sqrt{\sigma_i^2 + C}},$$

where σ_i are singular values of X. ■

D.1.3 Rashomon Volume Lower Bounds for Ridge Regression

The results described in this section follow directly from the Theorem 16 and are basic observations of the Rashomon volume closed-form formula for ridge regression.

Corollary 26 *For a data set $S = X \times Y$ such that a Frobenius norm of the feature matrix X is bounded $\|X\|_F = \sqrt{\sum_{i,j}^{p,n} x_{ij}^2} \in [1, F]$ and for a parametric hypothesis space of linear models $\mathcal{F}_\Omega = \{f_\omega(x) = \omega^T x, \omega \in \mathbb{R}^p\}$, the Rashomon volume of ridge regression is at least*

$$\mathcal{V}(\hat{R}_{set}(\mathcal{F}, \theta)) \geq \frac{2K(\theta, p)}{F + pC}.$$

Proof

For real $a_i \geq 0$, $i = 1, .., p$ we have that $\left(\prod_{i=1}^p a_i\right)^{1/q} \leq \left(\sum_{i=1}^p a_i\right)/q$, then by by setting $q = 2$ and $a_i = \sigma_i^2 + C$ we get: $\left(\prod_{i=1}^p \left(\sigma_i^2 + C\right)\right)^{\frac{1}{2}} \leq \frac{1}{2}\left(\sum_{i=1}^p \left(\sigma_i^2 + C\right)\right) =$

$\frac{1}{2}\sum_{i=1}^{p}\left(\sigma_i^2+pC\right) = \frac{1}{2}\left(\|X\|_F+pC\right) \leq \frac{1}{2}\left(F+pC\right)$, therefore from the Theorem 16 we have that $\mathcal{V}(\hat{R}_{set}(\mathcal{F},\theta)) \geq \frac{2K(\theta,p)}{F+pC}$.

■

Corollary 27 *For a data set $S = X \times Y$ and a parametric hypothesis space of linear models $\mathcal{F}_\Omega = \{f_\omega(x) = \omega^T x, \omega \in \mathbb{R}^p\}$, if $\frac{\partial^2 \hat{L}}{\partial \omega_j^2} \leq \delta$, such that $\delta \geq 2C$, then the Rashomon volume of ridge regression is at least*

$$\mathcal{V}(\hat{R}_{set}(\mathcal{F},\theta)) \geq \frac{2K(\theta,p)}{\sqrt{p(\frac{\delta}{2}-C)}+pC}.$$

Proof

As in previous Corollary we can bound the singular values product with the Frobenius norm of the feature matrix $(\prod_{i=1}^{p}(\sigma_i^2+c))^{\frac{1}{2}} \leq \frac{1}{2}(\|X\|_F+pc)$. Given the bounded second derivative we have $\frac{\partial^2 \hat{L}}{\partial \omega_j^2} = 2\sum_i \sum_j x_{ij}^2 + 2C \leq \delta$. By the assumption $\delta \geq 2C$ we get that $\sum_i \sum_j x_{ij}^2 \leq \frac{\delta}{2} - C$ and therefore we can upper bound the Frobenius norm as follows: $\|X\|_F = \sqrt{\sum_j (\sum_i x_{ij}^2)} \leq \sqrt{\sum_j (\frac{\delta}{2}-C)} = \sqrt{p(\frac{\delta}{2}-C)}$. Taking into account the Theorem 16 $\mathcal{V}(\hat{R}_{set}(\mathcal{F},\theta)) \geq \frac{2K(\theta,p)}{\sqrt{p(\frac{\delta}{2}-C)}+pC}$.

■

Corollary 28 *For a data set $S = X \times Y$, such that x_i are on a unit sphere $\forall i : \|x_i\| = 1$ and a parametric hypothesis space of linear models $\mathcal{F}_\Omega = \{f_\omega(x) = \omega^T x, \omega \in \mathbb{R}^p\}$, the Rashomon volume of ridge regression is at least*

$$\mathcal{V}(\hat{R}_{set}(\mathcal{F},\theta)) \geq \frac{2K(\theta,p)}{\sqrt{n}+pC}$$

Proof

As in previous Corollaries we can bound the singular values product with the Frobenius norm of the feature matrix $(\prod_{i=1}^{p}(\sigma_i^2+c))^{\frac{1}{2}} \leq \frac{1}{2}(\|X\|_F+pc)$. Since $\|X\|_F = \sqrt{\sum_i (\sum_j x_{ij}^2)} = \sqrt{\sum_i 1} = \sqrt{n}$, then $\prod_{i=1}^{n}\sqrt{\sigma_i^2+c} \leq \frac{\sqrt{n}+pc}{2}$, and combined with the Theorem 16 we get that $\mathcal{V}(\hat{R}_{set}(\mathcal{F},\theta)) \geq \frac{2K(\theta,p)}{\sqrt{n}+pC}$.

■

D.2 Convex Loss

For a parametric hypothesis space where the loss $l(\omega, z)$ is convex with respect to the parameter vector ω, the Rashomon set as well as the hypothesis space, are convex as well, and we can use random walks to estimate their volumes. In particular, according to Theorem

2.1 by Kannan et al. (1997), there exists a randomized algorithm that can approximate, with high probability, the volume of a convex body $V \in \mathbb{R}^p$ within an ϵ error using approximately $O(p^5)$ calls to a separating oracle. In particular we can approximate the volume $\hat{\mathcal{V}}(V)$ such that:

$$(1-\epsilon)\hat{\mathcal{V}}(V) < \mathcal{V}(V) < (1+\epsilon)\hat{\mathcal{V}}(V). \tag{9}$$

To adapt the randomized algorithm theorem for Rashomon set estimation in the parameter space Ω, we need to construct a separating oracle (Grötschel et al., 2012): a routine that, for a given point λ and convex set Λ, tells us whether $\lambda \in \Lambda$, and if not, provides a separating hyperplane between λ and Λ. From the Rashomon set definition, given a parameter vector ω, we check if f_ω belongs to the Rashomon set according to $\hat{L}(f_\omega) \leq \hat{L}(\hat{f_\omega}) + \theta$. If f_ω is not in the Rashomon set, we construct a separating hyperplane in parameter space Ω using the perpendicular to the tangent hyperplane. In particular, since the loss is convex, a tangent hyperplane at point ω looks like:

$$\nabla l(\omega, \cdot)(\gamma - \omega) = 0.$$

Let $\omega_{pr} = PR_{\hat{R}_{set}(\mathcal{F}_\Omega, \theta)}(\omega)$ be a projection of point ω onto the Rashomon set. Then, we derive a separating hyperplane to be in the middle of ω and its projection:

$$\nabla l(\omega, \cdot)(\gamma - \omega) + (\omega - \omega_{pr})/2 = 0.$$

Applying the constructed separating oracle to the randomized algorithm theorem, with high probability, we can achieve approximation guarantees for the Rashomon volume given in (9).

D.3 Rashomon Volume Under-Approximation for SVM-1

In this section we propose an optimization procedure that allows to under-approximate the Rashomon volume in the parameter space for the Support Vector Machine (SVM)(Burges, 1998) with L-1 regularization.

Consider binary classification for the class of linear models. Directly computing the Rashomon ratio requires sampling over an infinite space of linear functions. Instead we propose a procedure that allows us to compute an L-1 ball that is contained within the Rashomon set for the SVM-1 (Cai et al., 2011; Zhu et al., 2004) learning algorithm.

Theorem 29 *For the SVM-1 with hinge loss $\phi(f(x), y) = \lfloor 1 - yf(x) \rfloor_+$, L-1 regularization, and for parameterized hypothesis space of linear models $\mathcal{F}_\Omega = \{\omega^T x,\ \omega \in \mathbb{R}^p\}$ the Rashomon volume is at least*

$$\mathcal{V}(\hat{R}_{set}(\mathcal{F}, \theta)) \geq \frac{2^p \delta^p}{p!},$$

where $\delta = \|\omega^v - \omega^c\|_1$, ω^c is an optimal solution to the 1-norm SVM problem and ω^v satisfies $\min_{\omega^v \in \hat{R}_{set}(\mathcal{F}_\Omega, \theta)} \|\omega^v - \omega^c\|_1$

Proof

Given that ω^c is an optimal solution to the 1-norm SVM problem, it satisfies:

$$\min_{\omega^c} \|\omega^c\|_1 + \sum_{i=1}^{n} \lfloor 1 - y_i f_{\omega^c}(x_i) \rfloor_+.$$

Let ω^v be a solution to the following optimization problem:

$$\min_{\omega} \|\omega - \omega^c\|_1, \quad \text{s.t.} \quad \sum_{i=1}^{n} \lfloor 1 - y_i f_{\omega}(x_i) \rfloor_+ \geq \theta + \sum_{i=1}^{n} \lfloor 1 - y_i f_{\omega^c}(x_i) \rfloor_+,$$

then ω^v is in the Rashomon set, $\omega^v \in \hat{R}_{set}(\mathcal{F}_\Omega, \theta)$, and is the closest to ω^c in the parameter space. The convexity of the optimization problem ensures that the inequality constraint will be tight since any ω^v for which the constraint is not tight can be replaced with one that improves the objective function by moving towards ω^c.

Since ω^v is the closest in L-1 norm to ω^c and has the largest tolerable loss, then all models in the cross-polytope centered in ω^c with a half-diagonal $\delta = \|\omega^v - \omega^c\|_1$ will be in the Rashomon set $\hat{R}_{set}(\mathcal{F}_\Omega, \theta)$ because they will have loss no more than that of ω^v. Therefore, the Rashomon volume is at least the volume of the cross-polytope, which is given by $\frac{2^p \delta^p}{p!}$.

■

Appendix E. Proofs for Connection to Simplicity Measures

E.1 Proof of Theorem 18

Theorem 18 (Rashomon ratio and algorithmic stability) *Consider a distribution P_X over a discrete domain $\mathcal{X} = \{x_1, ...x_N\}$ and a learning algorithm A that minimizes ridge regression's empirical risk $\hat{L}$ for a linear hypothesis space $\mathcal{F}_\Omega$, as in Equation (3). For any $\lambda > 0$ there exist joint distributions P_{X,Y_1} and P_{X,Y_2} where for $\mathbf{X}$ drawn i.i.d. from P_X, $\mathbf{Y}_1$ is drawn from $P_{Y_1|\mathbf{X}}$ over $\mathcal{Y} \,|\, \mathcal{X}$ and $\mathbf{Y}_2$ is drawn from $P_{Y_2|\mathbf{X}}$ over $\mathcal{Y} \,|\, \mathcal{X}$, such that the expected Rashomon ratios are the same:*

$$\mathbb{E}_{P_{X,Y_1}}[R_{ratio_{\mathbf{S}_1}}(\mathcal{F}_\Omega, \theta)] = \mathbb{E}_{P_{X,Y_2}}[R_{ratio_{\mathbf{S}_2}}(\mathcal{F}_\Omega, \theta)],$$

yet hypothesis stability constants are different by an arbitrarily chosen value of λ:

$$\tilde{\beta}_2 - \tilde{\beta}_1 \geq \lambda,$$

where $\mathbf{S}_1$ and $\mathbf{S}_2$ denote data sets $\mathbf{S}_1 = [\mathbf{X}, \mathbf{Y_1}]$ and $\mathbf{S}_2 = [\mathbf{X}, \mathbf{Y_2}]$, $\tilde{\beta}_1$ is the hypothesis stability coefficient of algorithm $\mathcal{A}$ for distribution P_{X,Y_1} and $\tilde{\beta}_2$ is the hypothesis stability coefficient for distribution P_{X,Y_2}.

Proof Consider the least squares regression $\min_\omega \sum_{i=1}^{n} l(\omega, \mathbf{z}_i)^2$, where $\omega \in \mathbb{R}^p$, and loss $l(\omega, \mathbf{z}) = \phi(\omega^T \mathbf{x}, \mathbf{y})$ for $\mathbf{z} = (\mathbf{x}, \mathbf{y})$. For the marginal distribution P_X and $\mathbf{X} = [\mathbf{x_1}, ..., \mathbf{x_n}]$ drawn i.i.d. from P_X we design distributions $P_{Y_1|\mathbf{X}}$ and $P_{Y_2|\mathbf{X}}$ as:

$$P_{Y_1|\mathbf{X}}(y = \mathbf{0}|\mathbf{x}) = 1 \;\; \forall \mathbf{x} \in \mathbf{X},$$

$$P_{Y_2|\mathbf{X}}(y=\mathbf{0}|\mathbf{x}\neq\mathbf{x_0})=1,\ P_{Y_2|\mathbf{X}}(y=\mathbf{0}|\mathbf{x}=\mathbf{x_0})=0.5,\ P_{Y_2|\mathbf{X}}(y=\mathbf{H}|\mathbf{x}=\mathbf{x_0})=0.5,$$

where $\mathbf{x_0}\in\{x_1,...,x_N\}$ is some fixed point with a positive probability $P_X(\mathbf{x_0})$ and we define $\mathbf{H}\in\mathbb{R}$ later.

According to the definition of algorithmic stability, for P_{X,Y_1} we have:

$$\mathbb{E}_{S_1,z}[|l(f_{\mathbf{S}_1},\mathbf{z})-l(f_{\mathbf{S}_1^{\setminus i}},\mathbf{z})|]=0=\tilde{\beta}_1,$$

and for distribution P_{X,Y_2}:

$$\begin{aligned}\mathbb{E}_{S_2,z}\left[\left|l(f_{\mathbf{S}_2},\mathbf{z})-l(f_{\mathbf{S}_2^{\setminus i}},\mathbf{z})\right|\right]&=\sum_{\mathbf{S}_2,\mathbf{z}\sim P_{X,Y_2}}P_{X,Y_2}(\mathbf{S}_2)P_{X,Y_2}(\mathbf{z})\left|l(f_{\mathbf{S}_2},\mathbf{z})-l(f_{\mathbf{S}_2^{\setminus i}},\mathbf{z})\right|\\&\geq P_{X,Y_2}(\mathbf{S}_2^s)P_{X,Y_2}(\mathbf{z}^s)\left|l(f^s_{\mathbf{S}_2},\mathbf{z}^s)-l(f_{\mathbf{S}_2^{s,\setminus i}},\mathbf{z}^s)\right|,\end{aligned}$$

where $\mathbf{S}_2^s,\mathbf{z}^s$ is a special draw such that $\mathbf{z}^s=(\mathbf{x_0},\mathbf{H})$ and $\mathbf{S}_2^s$ contains both $(\mathbf{x_0},\mathbf{H})$ and $(\mathbf{x_0},\mathbf{0})$. Since the domain $\mathcal{X}$ is discrete, the probabilities of a special draw are:

$$P_{X,Y_2}(\mathbf{z}^s)=\frac{1}{2}Bin(1,n,P_X(\mathbf{x_0})),\ \ P_{X,Y_2}(\mathbf{S}_2^s)=\frac{1}{4}Bin(1,n,P_X(\mathbf{x_0}))^2Bin(n-2,n,1-P_X(\mathbf{x_0})),$$

where $Bin(k,n,p_k)=\binom{n}{k}p_k^k(1-p_k)^{(n-k)}$ is a binomial coefficient, namely a probability of getting exactly k successes from n trials, where each trial has a probability of success p_k. Denote $P_{(\mathbf{S}_2^s,\mathbf{z}^s)}$ as the probability of getting a special draw, then $P_{(\mathbf{S}_2^s,\mathbf{z}^s)}=P_{X,Y_2}(\mathbf{S}_2^s)P_{X,Y_2}(\mathbf{z}^s)$.

If $\mathbf{S}_2^s$ contains only two points $(\mathbf{x_0},\mathbf{H})$ and $(\mathbf{x_0},\mathbf{0})$, the loss difference $|l(f^s_{\mathbf{S}_2},\mathbf{z}^s)-l(f_{\mathbf{S}_2^{s,\setminus i}},\mathbf{z}^s)|$ evaluated at $\mathbf{z}^s$ for all i will be at least $\frac{\mathbf{H}^2}{4}$. As we add more points $(\mathbf{x}_i,\mathbf{0})$ to the data set $\mathbf{S}_2^s$ the loss difference in the special draw case will only increase. Therefore for all i:

$$|l(f^s_{\mathbf{S}_2},\mathbf{z}^s)-l(f_{\mathbf{S}_2^{s,\setminus i}},\mathbf{z}^s)|\geq\frac{\mathbf{H}^2}{4}.$$

If we choose $\mathbf{H}$ such that $\mathbf{H}>2\sqrt{\lambda}\left(P_{(\mathbf{S}_2^s,\mathbf{z}^s)}\right)^{-1/2}$, then from the definition of algorithmic stability we have:

$$\tilde{\beta}_2\geq\mathbb{E}_{S_2,z}\left[\left|l(f_{\mathbf{S}_2},\mathbf{z})-l(f_{\mathbf{S}_2^{\setminus i}},\mathbf{z})\right|\right]\geq P_{(\mathbf{S}_2^s,\mathbf{z}^s)}\frac{\mathbf{H}^2}{4}>\lambda.$$

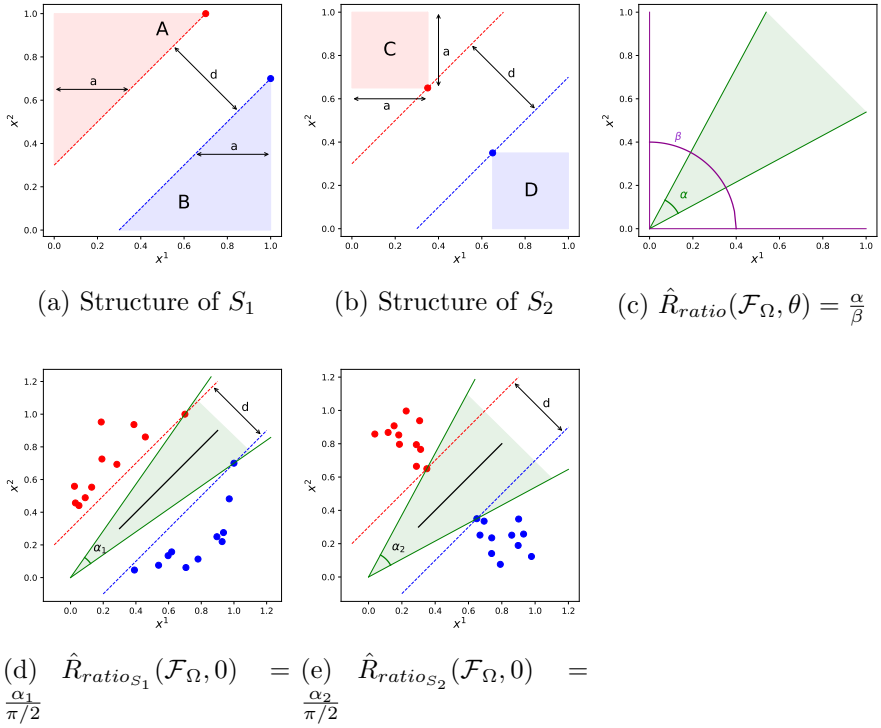
Therefore for any given λ we get that $\left|\tilde{\beta}_1-\tilde{\beta}_2\right|>\lambda$.

On the other hand, the Rashomon volume for the hypothesis space $\mathcal{F}_\Omega$ of linear models does not depend on targets and can be calculated as in (4) for both $\mathbf{S}_1$ and $\mathbf{S}_2$. Therefore the expected Rashomon volumes are the same:

$$\mathbb{E}_{S_1}\left[\mathcal{V}(R_{set_{\mathbf{S}_1}}(\mathcal{F}_\Omega,\theta))\right]=\mathbb{E}_{S_2}\left[\mathcal{V}(R_{set_{\mathbf{S}_2}}(\mathcal{F}_\Omega,\theta))\right].$$

Given the equality of the expected Rashomon volumes the the expected Rashomon ratios are the same as well as we do not change the hypothesis space when drawing different joint distribution.

■



(a) Structure of S_1 (b) Structure of S_2 (c) $\hat{R}_{ratio}(\mathcal{F}_\Omega, \theta) = \frac{\alpha}{\beta}$

(d) $\hat{R}_{ratio_{S_1}}(\mathcal{F}_\Omega, 0) = \frac{\alpha_1}{\pi/2}$ (e) $\hat{R}_{ratio_{S_2}}(\mathcal{F}_\Omega, 0) = \frac{\alpha_2}{\pi/2}$

Figure 14: An illustration of different Rashomon ratios with identical geometric margins. The black line shows the optimal model, the shaded region indicates the Rashomon set $\hat{R}_{set}(\mathcal{F}_\Omega, 0)$ with its boundaries represented by green lines, the dark color indicates boundaries of the hypothesis space. (a) and (b) show the data sets S_1 and S_2 with identical margin d. (c) shows that the Rashomon ratio can be computed as a ratio of angles α (represents the Rashomon set) and β (represents the hypothesis space). (d) and (e) illustrate that data sets S_1 and S_2 are represented by different angles α_1 and α_2 and therefore have different Rashomon ratios. Best seen in color.

E.2 Proof of Theorem 19

Theorem 19 (Rashomon ratio and geometric margin) *For any fixed $0 < \lambda < 1$, there exists a fixed hypothesis space $\mathcal{F}_\Omega$, a Rashomon parameter θ, and there exist two data sets S_1 and S_2 with the same empirical risk minimizer $\hat{f} \in \mathcal{F}_\Omega$ such that the width of the geometric margin d is the same for both data sets, yet the Rashomon ratios are different:*

$$|R_{ratio_{S_1}}(\mathcal{F}_\Omega, \theta) - R_{ratio_{S_2}}(\mathcal{F}_\Omega, \theta)| > \lambda.$$

Proof

Consider two-dimensional separable data, $\mathcal{X} \in [0,1]^2$, and a parametrized hypothesis space of origin-centered linear models: $\mathcal{F} = \{\omega^T x, \omega = (k, -1), x \in \mathbb{R}^2, k \in \mathbb{R}\}$. Consider also 0-1 loss $\phi_\omega(x, y) = \mathbb{1}_{[y=sign(\omega^T x)]}$ and an empirical risk minimizer $\hat{f} = f_{\hat{\omega}}$ that maximizes the geometric margin. Since the data are populated in a $[0,1]^2$ hypercube, as a hypothesis space we will consider all models that intersect the unit-hypercube.

For some positive constant $a \in (0,1)$ that we choose later, consider the following regions of the feature space:

$$A = \{x^1 \in [0, 1-a), x^2 > x^1 + (1-2a)\}, \quad B = \{x^1 \in (a, 1], x^2 < x^1 - (1-2a)\},$$

$$C = \{x^1 \in [0, a), x^2 \in (1-a, 1]\}, \quad D = \{x^1 \in (1-a, 1], x^2 \in [0, a)\}.$$

Construct data set S_1, such that $S_1 = (x_A, 1) \cup (x_B, -1) \cup (x_{S_1}^{s_1}, 1) \cup (x_{S_1}^{s_2}, -1)$, where $x_A \in A$ is any sample from the region A, $x_B \in B$ is any sample from the region B, $x_{S_1}^{s_1}$ and $s_{S_1}^{s^2}$ are special points for the data set S_1 such that $x_{S_1}^{s_1} = [1-2a, 1]$ and $x_{S_1}^{s_2} = [1, 1-2a]$. Please see Figure 14a for details.

Construct data set S_2, such that $S_2 = (x_C, 1) \cup (x_D, -1) \cup (x_{S_2}^{s_1}, 1) \cup (x_{S_2}^{s_2}, -1)$, where $x_C \in C$ is any sample from the region C, $x_D \in D$ is any sample from the region D, $x_{S_2}^{s_1}$ and $x_{S_2}^{s_2}$ are special points for the data set S_2 such that $x_{S_2}^{s_1} = [a, 1-a]$ and $x_{S_2}^{s_2} = [1-a, a]$. Please see Figure 14b for details.

Note that the data sets we considered have the same width for the geometrical margin $d = \sqrt{2}(2a-1)$ (see Figures 14a, 14b). Now, we are left to show that the Rashomon ratios are different.

For the functional space of origin-centered lines we have a unique parameterization and a one-to-one correspondence between an actual model and its parameterization. Therefore, if the Rashomon set is a single connected component, an angle α between the two most distant models in the Rashomon set gives us some information about the Rashomon volume. In particular, we can compute the Rashomon ratio as a ratio of the angle α that represents the Rashomon set and the angle β that corresponds to the hypothesis space as shown on Figure 14c. Since the hypothesis space is defined on the unit-hypercube, $\beta = \pi/2$ and for the Rashomon parameter $\theta = 0$ the Rashomon ratio is:

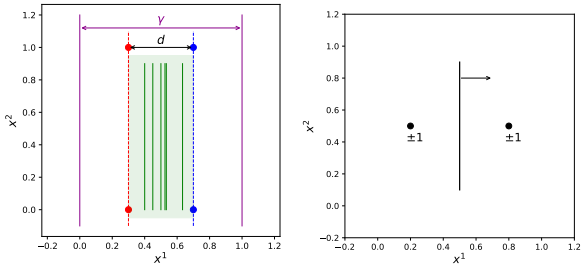
$$\hat{R}_{ratio}(\mathcal{F}, 0)) = \frac{\alpha}{\beta} = \frac{2 \max_{f \in \hat{R}_{set}(\mathcal{F}_\Omega, 0)} |\arctan(f_{\hat{\omega}}) - \arctan(f_\omega)|}{\pi/2}.$$

For data sets S_1 and S_2 Figures 14d and 14e show the Rashomon set and angles α_1 and α_2 that represent the Rashomon volume. Given the special points in the data sets we can compute α_1 and α_2 exactly: $\alpha_1 = 2\left(\arctan(1) - \arctan(1-2a)\right) = \frac{\pi}{2} - 2\arctan(1-2a)$ and $\alpha_2 = 2\left(\arctan(1) - \arctan\left(\frac{a}{1-a}\right)\right) = \frac{\pi}{2} - 2\arctan\left(\frac{a}{1-a}\right)$. Then the Rashomon ratios difference is:

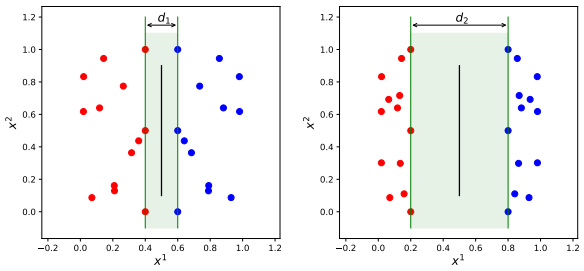
$$\begin{aligned} |R_{ratio_{S_1}}(\mathcal{F}, 0) - R_{ratio_{S_2}}(\mathcal{F}, 0)| &= \left|\frac{\alpha_1 - \alpha_2}{\pi/2}\right| = \left|\frac{4}{\pi}\left(\arctan(1-2a) - \arctan\left(\frac{a}{1-a}\right)\right)\right| \\ &= \left|\frac{4}{\pi}\arctan\left(1 - \frac{4a-2}{2a^2-1}\right)\right|. \end{aligned}$$

Now if we choose $a \in (0,1)$ and such that $\left|\frac{4}{\pi}\arctan\left(1 - \frac{4a-2}{2a^2-1}\right)\right| > \lambda$, then the Rashomon ratio difference $|R_{ratio_{S_1}}(\mathcal{F}, 0) - R_{ratio_{S_2}}(\mathcal{F}, 0)|$ is at least λ.

■



(a) $\hat{R}_{ratio}(\mathcal{F}_\Omega, \theta) = \frac{d}{\gamma}$ (b) Toy data set



(c) $\hat{R}_{ratio_{S_1}}(\mathcal{F}_\Omega, 0) = d_1$ (d) $\hat{R}_{ratio_{S_2}}(\mathcal{F}_\Omega, 0) = d_2$

Figure 15: An illustration of different Rashomon ratios with equivalent empirical local Rademacher complexities. Black line shows the optimal model, shaded region indicates the Rashomon set $\hat{R}_{set}(\mathcal{F}_\Omega, 0)$ with its models represented by green lines, the magenta color indicates boundaries of the hypothesis space. (a) The projected minimal distance d is equivalent to the Rashomon volume. (b) A toy data set that illustrates that the empirical local Rademacher complexity is zero for models in the Rashomon set. (c) Data set S_1, and (d) Data set S_2 illustrate symmetric separable data sets with different Rashomon ratios. Best seen in color.

E.3 Proof of Theorem 21

Theorem 21 (Rashomon ratio and local Rademacher complexity) *For $0 < \lambda < 1$, there exist two data sets S_1 and S_2, a hypothesis space $\mathcal{F}_\Omega$, and a Rashomon parameter θ such that the local Rademacher complexities defined on the Rashomon sets for S_1 and S_2 are the same:*

$$\hat{R}_n^{S_1}\left(\hat{R}_{set}(\mathcal{F}_\Omega, \theta)\right) = \hat{R}_n^{S_2}\left(\hat{R}_{set}(\mathcal{F}_\Omega, \theta)\right),$$

yet the Rashomon ratios are different:

$$\left| R_{ratio_{S_1}}(\mathcal{F}_\Omega, \theta) - R_{ratio_{S_2}}(\mathcal{F}_\Omega, \theta) \right| > \lambda.$$

Proof

Consider two-dimensional separable symmetric data, $\mathcal{X} \in [0,1]^2$, $\mathcal{Y} = \{0,1\}$, 0-1 loss $\phi_f(x,y) = \mathbb{1}_{[y=signf(x)]}$ with empirical risk minimizer $\hat{f}$, and a hypothesis space $\mathcal{F}_\Omega$ of decision stumps based on the first feature, where for $f \in \mathcal{F}_\Omega$: $f = 1$ if $x^1 > \omega$, $\omega \in \mathbb{R}$, $f = 0$

otherwise. We have a one-to-one correspondence between a function and its threshold parameter ω. Therefore, if the Rashomon set is a single connected component, we can compute the Rashomon volume in a parameter space by computing the difference between the largest and smallest threshold values of models within the Rashomon set, as illustrated in Figure 15a. For $\theta = 0$, the difference between the largest and the smallest threshold values will be equivalent to the minimal distance between points of opposite classes projected onto the first feature $d = \min_{x_i, x_j : y_i \neq y_j} |PR_1(x_i) - PR_1(x_j)|$, where PR_1 is the projection of point x onto first feature.

For the hypothesis space, we consider all decision stumps in the first dimension that are in the segment $[0, 1]$, where data are populated. The difference in thresholds for the hypothesis space is $\beta = 1$ and therefore $\mathcal{V}(\mathcal{F}_\Omega) = 1$. For $\theta = 0$, the Rashomon volume will be equivalent to d—the projected minimal distance between points of opposite classes, and have that $\mathcal{V}(\hat{R}_{set}(\mathcal{F}_\Omega, 0)) = d$ and $\hat{R}_{ratio}(\hat{R}_{set}(\mathcal{F}_\Omega, 0)) = \frac{d}{1} = d$. Now consider any two separable symmetric data sets S_1, S_2 with different projected minimal distances d_1 and d_2, such that $|d_1 - d_2| > \lambda$. (Please see Figure 15c and 15d for details of the data sets S_1 and S_2.) Consequently we get that:

$$\left| R_{ratio_{S_1}}(\mathcal{F}_\Omega, 0) - R_{ratio_{S_2}}(\mathcal{F}_\Omega, 0) \right| = |d_1 - d_2| > \lambda.$$

For a separable symmetric data S and 0-1 loss function, the Rashomon set $\hat{R}_{set}(\mathcal{F}_\Omega, 0)$ contains all models that separate data in the same way. Therefore the Rademacher complexity of the Rashomon set is $\hat{R}_n^S\left(\hat{R}_{set}(\mathcal{F}_\Omega\right)$ is:

$$\hat{R}_n^S\left(\hat{R}_{set}(\mathcal{F}_\Omega, 0)\right) = \frac{1}{n}\mathbb{E}_\sigma\left[\sup_{f \in \hat{R}_{set}(\mathcal{F}_\Omega, 0)} \sum_{i=1}^n \sigma_i f(x_i)\right] = \frac{1}{n}\mathbb{E}_\sigma\left[\sum_{i=1}^n \sigma_i \hat{f}(x_i)\right] = 0,$$

where in the penultimate equality we have used the fact that, in the case of separable data and $\theta = 0$, all models in the Rashomon set will perform identically on any permutation of the labels.

Equality of the empirical Rademacher complexity of the optimal model to zero follows from the symmetric data considered and symmetrical patterns of all possible target assignments. For example, for a toy data set in Figure 15b: $\hat{R}_n^S\left(\hat{R}_{set}(\mathcal{F}_\Omega, 0)\right) = \frac{1}{2}\frac{1}{4}\left(\left(\hat{f}(x_1) + \hat{f}(x_2)\right) + \left(\hat{f}(x_1) - \hat{f}(x_2)\right) + \left(-\hat{f}(x_1) + \hat{f}(x_2)\right) + \left(-\hat{f}(x_1) - \hat{f}(x_2)\right)\right) = 0$. Since both S_1 and S_2 are separable and symmetric we get that:

$$\hat{R}_n^{S_1}\left(\hat{R}_{set}(\mathcal{F}_\Omega, 0)\right) = 0 = \hat{R}_n^{S_2}\left(\hat{R}_{set}(\mathcal{F}_\Omega, 0)\right).$$

■

Appendix F. Data Set Descriptions

We provide a description of the data sets used in our experiments in Table 5. All of them we downloaded from the UCI machine learning repository. We show the number of features

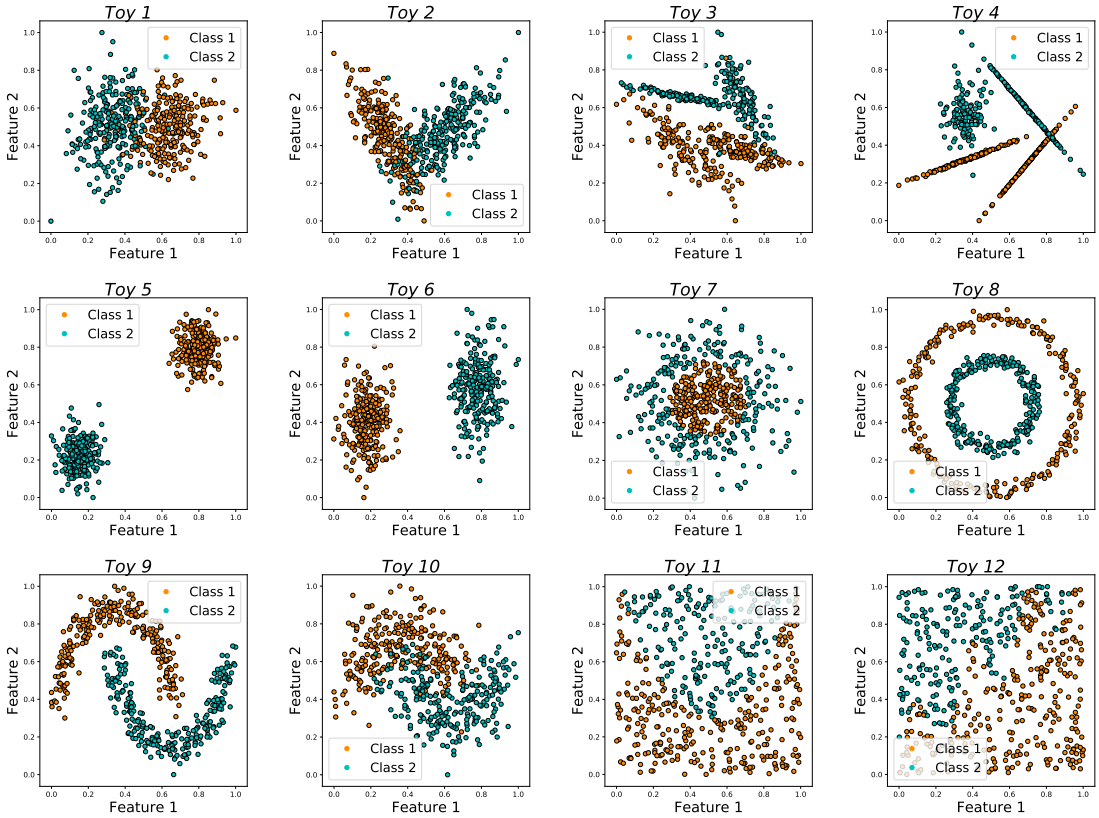
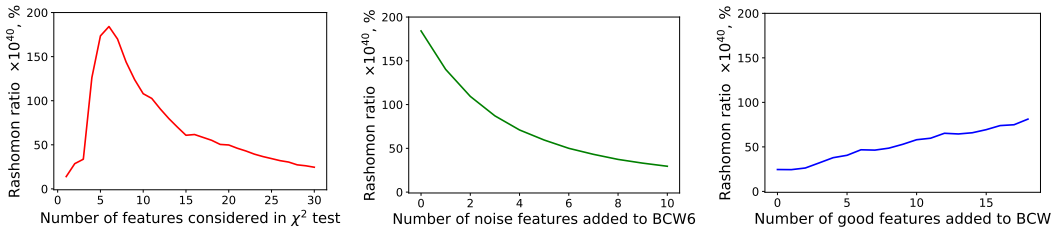


Figure 16: Synthetic two-dimensional data sets that we used in the experiments.

in each data set, sizes of the data set and any preprocessing steps that we used mainly to convert data to binary classification. For each data set, we perform cross-validation over ten folds for data sets with more than 200 points and over five folds for data sets with less than 200 points. We reserve one fold for testing, one for validation (e.g., hyper-parameter optimization) and the rest for training. All of the real-valued data sets were normalized to fit the unit-cube, and we did not standardize the data. During data processing, we omitted data records with missing values. We also omitted non-numerical features (e.g., date or text) when there was not a natural way to convert them to categorical features.

Additionally, we performed experiments on twelve synthetic binary classification data sets. These data sets have two real features and represent different geometrical concepts for two-dimensional classification (e.g., large and small margins, concentric circles, half moons, etc.) as in Figure 16. Results and implications for synthetic data sets are consistent with those on the UCI data sets.

Table 6 describes regression data sets, including characteristics and pre-processing notes, that were used in regression experiments.



(a) Reducing features (b) Adding noise (c) Adding good features

Figure 17: An illustration of the influence of feature quality on the Rashomon ratio for the BCW data set. (a) shows the Rashomon ratio for the data set with different number of significant features according to a χ^2 test. Denote the Breast Cancer Wisconsin data set (BCW) with six the most significant features as BCW6. (b) depicts the correspondence between the Rashomon ratio and different numbers of noisy features added to BCW6 data set. The noise features are sampled from normal distribution $\mathcal{N}(0, 1)$ and then standardized to be in a hypercube of volume one. (c) shows the change in the Rashomon ratio as we add more redundant features to the BCW data set. We iteratively add one out of six features from the BCW6 data set at a time. Rashomon ratios in (a)–(c) are averaged over ten folds. The Rashomon parameter θ is set to 0.05. Rashomon ratios are computed with respect to the best sampled model across all variations of the data set.

Appendix G. Quality of the Features

In our experiments, we observed a connection between the quality of the features and Rashomon ratios. The Rashomon ratio, as defined in (1) in its simplest form, is a volume fraction of models that are inside the Rashomon set compared to the models in the hypothesis space. When a data set is augmented with additional features, the size of the hypothesis space grows. If the added features are completely irrelevant (consisting, for instance, of noise) then adding these features increases the size of the hypothesis space but does not increase the size of the Rashomon set. Thus, we might predict that the Rashomon ratio could decrease as irrelevant features are added to a data set.

Additionally, if we augment a data set with features that are highly correlated or identical to features that improve performance, then not only is the size of the hypothesis space increased, but also the size of the Rashomon set is likely to increase, as there exist more relevant models (even if the set becomes redundant with models that predict equivalently). Thus, we might predict that the Rashomon ratio increases as we add copies of relevant features.

In general, these two examples of irrelevant and redundant features are corner cases, however, they do occur to a lesser degree in real world data sets, and we are interested in whether these cases have potentially influenced our experimental results in Section 7 in our observed Rashomon ratios. To investigate this, we augmented a data set with noise features, and separately, augment the same data set with copies of useful features to see whether

irrelevant or correlated features may have influenced our findings on the measurement of the Rashomon ratio. We used the Breast Cancer Wisconsin (Diagnostic) data set (shortly, BCW), which has approximately six important features. The results are shown in Figure 17. As before, our hypothesis space is decision trees of depth seven.

Irrelevant features. If the data set contains a lot of irrelevant or noisy features we expect the Rashomon set to be relatively small compared to the hypothesis space. Figure 17(a) shows how the Rashomon ratio changes as we iteratively decrease the number of features in the Wine data set, eliminating the least relevant features first, leaving the most significant ones (where relevance is determined according to a χ^2 test with the label). The Rashomon ratio grows as we first remove non-significant features, and after reaching a peak at around six features, it starts to decrease as we remove relevant features, and as models lose accuracy. Similarly, Figure 17(b) shows the influence of noisy features on the Rashomon ratio. Particularly, as we add more noisy irrelevant features, the Rashomon ratio starts to decrease. This is due to the same fact, that we artificially enlarge the hypothesis space while keeping the Rashomon set approximately the same. The noise features do not help improve the empirical risk, they only increase the size of the reasonable set.

Redundant features. As a contrast to how we increased the hypothesis space in the previous experiment, we can increase the Rashomon set by adding more redundant, good features. Figure 17(c) shows how the Rashomon ratio changes for the BCW data set as we add more copies of the four the most significant features. We observe that the Rashomon ratio increases. By adding copies of relevant features, we increased the number of tree at a given depth that could be good enough to be in the Rashomon set.

Our findings show a possible connection between the Rashomon ratio and feature analysis. In particular, in the case where different algorithms perform similarly, but the Rashomon ratio is observed to be small, it could be due to the reason that the data set contains noisy or irrelevant features. In that case, it may be possible to iteratively remove features to find those that produce the largest Rashomon ratio without changes to the empirical risk. The other extreme is less likely to be observed in practice, which is when the Rashomon ratio is extremely large due to redundant features. In that case, one could remove redundant (highly correlated) features before measuring the Rashomon ratio. The data sets with smaller numbers of features induce easier learning/optimization problems in general. As we discussed earlier, the Rashomon ratio would generally not be measured in practice, and would be inferred in other ways. Thus, these results mainly pertain to an understanding of the experiments we did in Section 7 to provide a possible explanation for cases of small observed Rashomon ratios but where all methods perform the same and all functions generalize.

Appendix H. Performance of Different Machine Learning Algorithms and Rashomon Ratio

Figure 18 and Figure 19 show a performance comparison of different machine learning algorithms with regularization for the categorical and real-valued data sets. Data sets shown in Figures 18 and 19 are shown in decreasing order of the Rashomon ratio, from the highest in Figure 18 to the Rashomon ratios that were so small that we were not able to

measure them, in Figure 19. Figure 21 and Figure 22 show a comparison of the performance of different machine learning algorithms without regularization for the categorical and real-valued data sets. Recall that algorithms with regularization and without regularization have different Rashomon sets, and can thus have different Rashomon ratios. Finally, Figure 20 and Figure 23 shows a performance comparison of different machine learning algorithms for the synthetic data sets with and without regularization.

As we mentioned before, we estimate the Rashomon ratio with importance sampling. For the proposal distribution, we generate a tree of depth D by randomly splitting on features. We assign labels to all 2^D leaves using the training data. If a leaf contains no training points, it acquires its label from the nearest ancestor that any training data pass through. The probability of sampling any tree from the proposal distribution is $p_p = p_f \times \prod_{i=1}^{2^D} 1$, where p_f is the probability of randomly sampling all of the features that comprise the splits of the tree. Our target distribution is a randomly sampled decision tree (both features and leaves) of depth D. Therefore, the probability of sampling a given tree from the target distribution is $p_t = p_f \times \prod_{i=1}^{2^D} \frac{1}{2}$, since we have two classification classes, where, as before, p_f is the probability of randomly sampling all features used within splits of the tree. Thus, for one tree of depth seven, its importance weight will be $\frac{p_t}{p_p} = \left(\frac{1}{2}\right)^{2^7} \approx 3 \times 10^{-39}$. The importance weight clearly dictates the order of magnitude of the Rashomon ratio in our experiments. The smallest possible non-zero Rashomon ratio ($\approx 1.175 \times 10^{-42}\%$) arises when we sample one model that is in the Rashomon set among 250,000 total models that were sampled. Therefore, we consider the Rashomon ratios of order $10^{-37}\%$ and $10^{-38}\%$ to be large, and Rashomon ratios of order $10^{-40}\%$, $10^{-41}\%$, etc., to be small.

If we choose another importance sampling method (for example with data assignment for only half of the leaves or with the guidance of both features and leaves) the Rashomon ratio may have different importance weights and therefore might have a different estimated size as well. This issue would be resolved if we sample a huge number of trees, which is hard to do in practice. Therefore, since our goal is to compare the Rashomon ratios across themselves, we use the same consistent method of leaf-based importance sampling across all data sets and sample a manageable number of trees (250,000 in our case).

Appendix I. Proof of Proposition 23

Proposition 23 *For a hierarchy of discrete hypothesis spaces $\mathcal{F}_1 \subset \mathcal{F}_2 \subset \cdots \subset \mathcal{F}_T$ such that there exists $f_1 \in \hat{R}_{set}(\mathcal{F}_T, \theta) \cap \mathcal{F}_1 \neq \varnothing$, if Assumption 22 holds for $B_1 = \{f_1\}$ then the Rashomon set $\hat{R}_{set}(\mathcal{F}_T, \theta)$ contains at least $\frac{C^T - 1}{C-1}$ models.*

Proof Consider $B_1 = \{f_1\}$, where $f_1 \in \hat{R}_{set}(\mathcal{F}_T, \theta) \cap \mathcal{F}_1 \neq \varnothing$. Given the assumptions, there exists at least C distinct functions f_2^i that form B_2 and are in the Rashomon set (since $\hat{L}(f_1) \geq \hat{L}(f_2^i)$). Furthermore, given B_2 there exists at least $|B_1| \times C \geq C^2$ more distinct functions in $\mathcal{F}_3$ that also belong to the Rashomon set, which now contains at least $1 + C + C^2$ models. Continuing to propagate the growth assumption through the hierarchy, at level T, the Rashomon set will have at least $1 + C + C^2 + C^3 + \cdots + C^{T-1} = \frac{C^T - 1}{C-1}$ models. ■

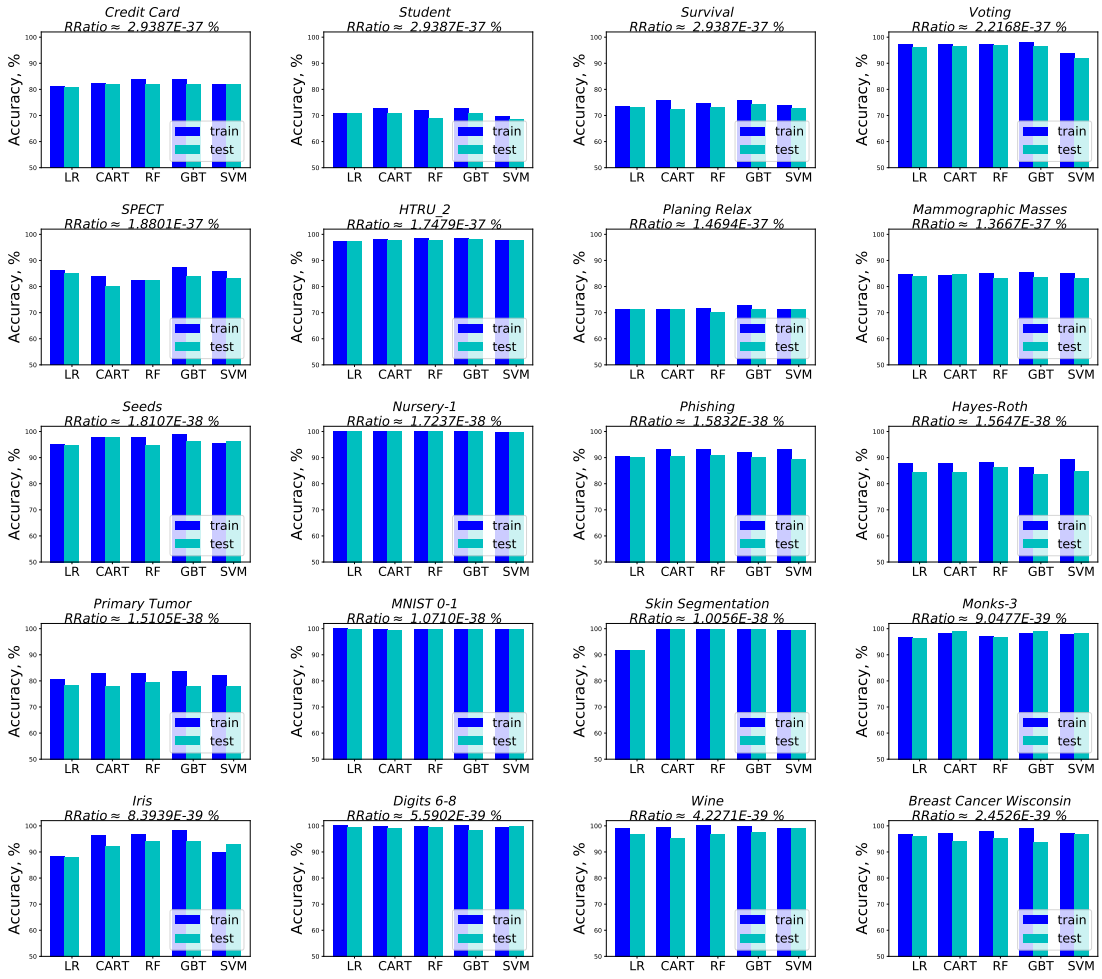


Figure 18: Performance of five machine learning algorithms with regularization for the UCI classification data sets. Data sets are listed in decreasing order of Rashomon ratio. Rashomon ratios, train and test accuracies are averaged over ten folds for data sets with more than 200 points and over five folds for data sets with less than 200 points. These plots continue in Figure 19. In these cases, test performance seems to be similar across algorithms. This will not be true in all cases as the Rashomon set becomes smaller, in Figure 19.

Appendix J. Rashomon Curve Plots for All Data Sets

Figures 24, 25 and Figures 26, 27, 28 show the Rashomon curves with generalization error for all categorical and real-valued data sets respectively. In columns (b) and (d) in the figures, we additionally show the Rashomon curves, where the reference model is the CART model if no better-performing sampled tree was found. In most of the cases, the CART model achieves better performance when the Rashomon ratio is small and we were not able to approximate it by sampling. We plot short Rashomon curves (the part we were ale to

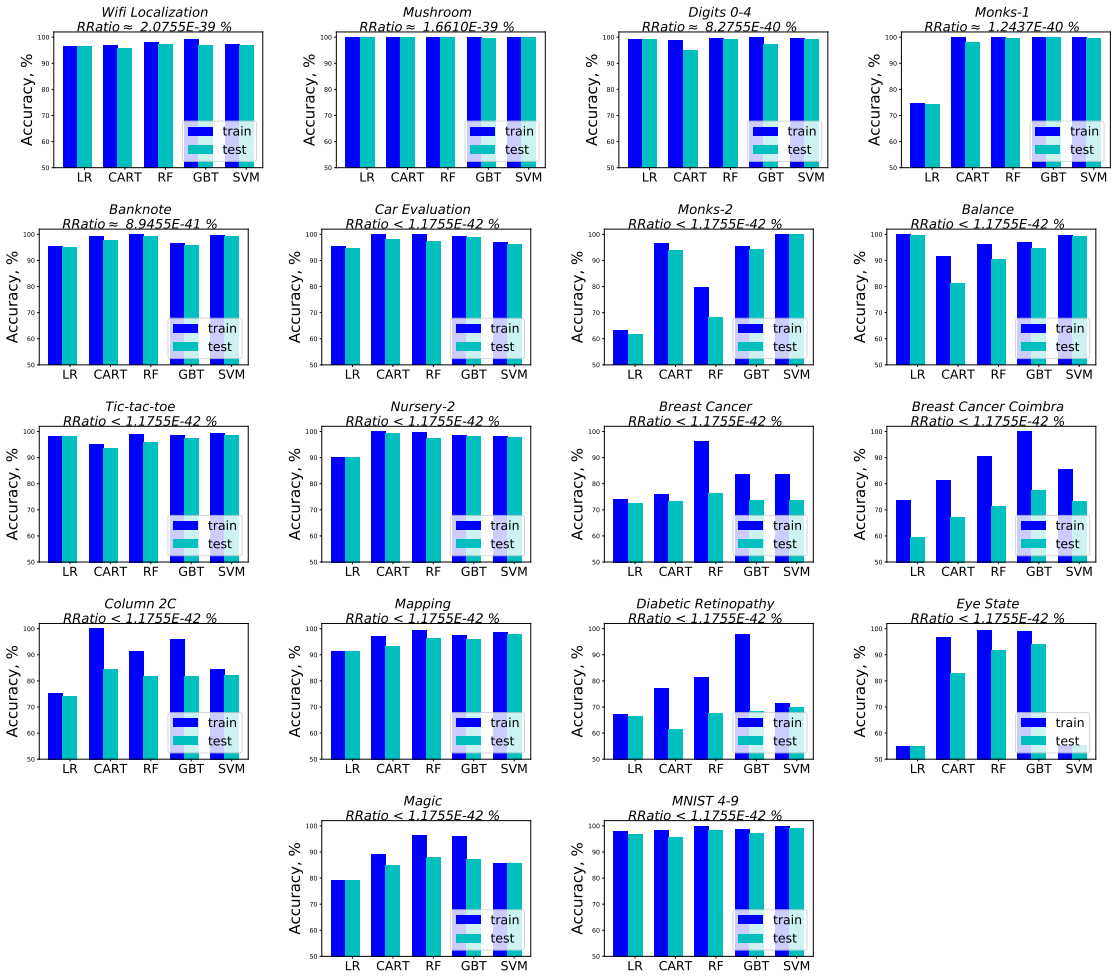


Figure 19: Performance of five machine learning algorithms with regularization for the UCI classification data sets. Data sets are listed in decreasing order of the Rashomon ratio, continued from Figure 18. Rashomon ratios, training accuracies, and test accuracies are averaged over ten folds for data sets with more than 200 points and over five folds for data sets with less than 200 points. Test performance sometimes varies across algorithms.

compute) for such datasets and indicate the training and test accuracies for the rest of the hierarchy of hypothesis spaces in a lower subplot.

Often in our experiments, the Rashomon curves that leverage CART (columns (b) and (d)) yield curves similar to the sampling-based Rashomon curves (columns(a) and (c)), especially when the Rashomon set is larger (e.g., HTRU_2 in Figure 28, Skin Segmentation in Figure 27). In other cases, CART-based Rashomon curves show a more tilted vertical trend (e.g., Car Evaluation in Figure 25) or visualize a small Rashomon ratio (e.g., Monks-2 in Figure 24).

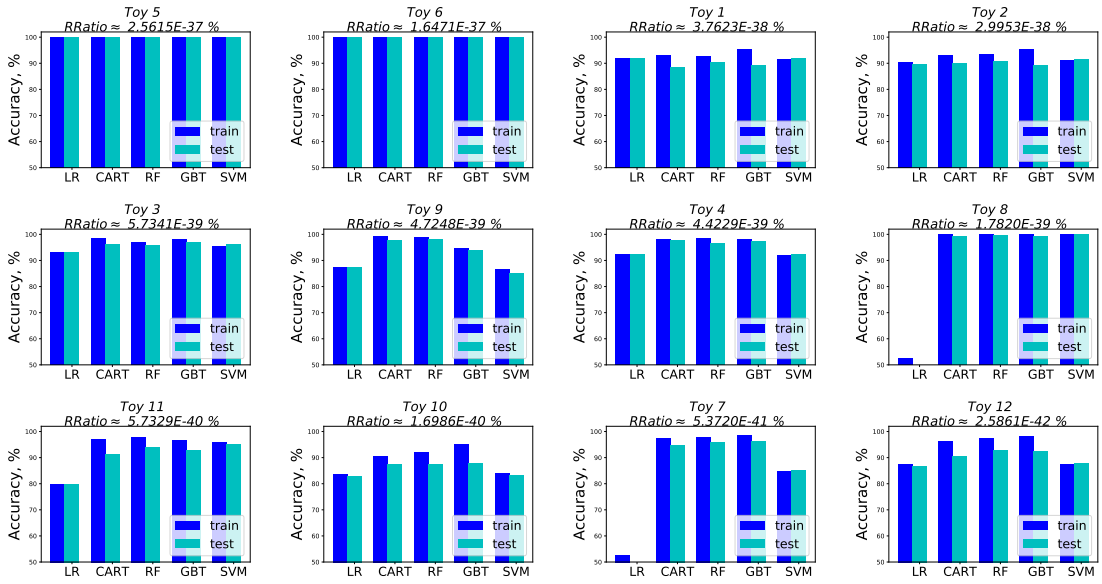


Figure 20: Performance of five machine learning algorithms with regularization for the synthetic data sets with real-valued features. Data sets are listed in decreasing order of the Rashomon ratio. Rashomon ratios, train and test accuracies are averaged over ten folds.

Figures 29 and 30 show the elbow on the Rashomon curve for two-dimensional synthetic data sets. Finally, Figure 31 shows the Rashomon trend for a hierarchy of polynomial hypothesis spaces for ridge regression.

Appendix K. Possible Ways to Compute the Rashomon Elbow

Let us create some simple ways to formalize how we might find the elbow. For a fixed θ, the elbow is the hypothesis space H_e with the highest Rashomon ratio among all model classes in the hierarchy that can approximately minimize the empirical risk as in Equation 5. The location of the Rashomon elbow can be found be solving an approximate maximization problem, where $G(\cdot, \cdot) : \mathbb{R}^2 \to \mathbb{R}$ is a practitioner-defined balance between accuracy and the Rashomon ratio. In particular, the elbow can be defined as a hypothesis space H_e such that:

$$H_e \in \underset{H_t \in H_1 \ldots H_T}{\operatorname{argmax}} \; G\left(1 - \hat{L}(H_t), \hat{R}_{ratio}(H_t, \theta_t)\right). \tag{10}$$

Here, G would be chosen by the practitioner to represent the ideal balance between accuracy and the Rashomon ratio, and the same G would be used for potentially many different problems for consistency.

As an alternative definition for the Rashomon elbow, we can use a geometric argument, based on intuition provided by Figure 10. Across all hypothesis spaces, the hypothesis space that corresponds to the Rashomon elbow has the largest distance from a line that connects

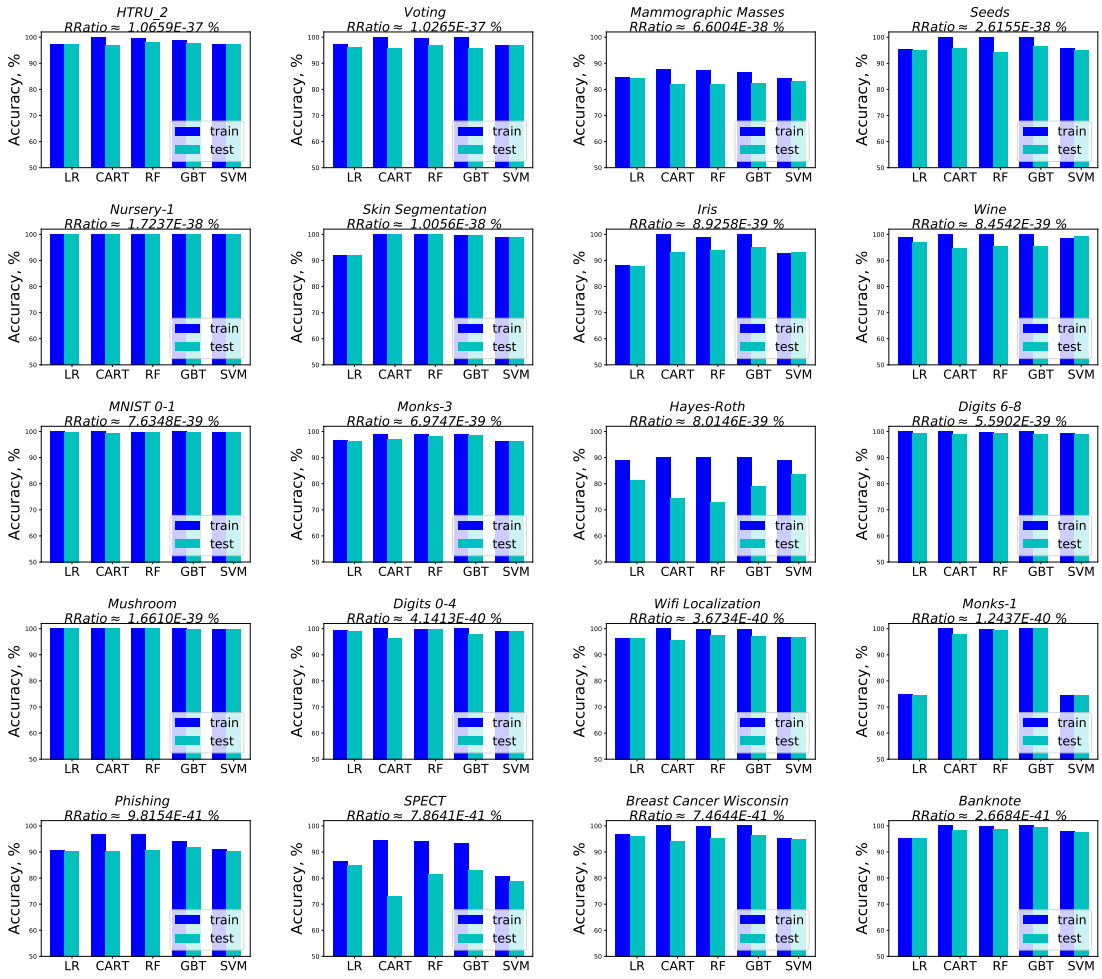


Figure 21: Performance of five machine learning algorithms without regularization for the UCI classification data sets. Data sets are listed in decreasing order of Rashomon ratio. Rashomon ratios, train and test accuracies are averaged over ten folds for data sets with more than 200 points and over five folds for data sets with less than 200 points. These plots continue in Figure 22. The datasets with larger Rashomon ratios correlate with similar performance of machine learning algorithms and good generalization.

.

points $(\hat{L}_{H_0}, \hat{R}_{ratio}(H_0, \cdot))$ and $(\hat{L}_{H_T}, \hat{R}_{ratio}(H_T, \cdot))$. This geometrically-defined Rashomon elbow will generally result in the same maximin point as in Equation (10).

References

Elaine Angelino, Nicholas Larus-Stone, Daniel Alabi, Margo Seltzer, and Cynthia Rudin. Certifiably optimal rule lists for categorical data. *Journal of Machine Learning Research*,

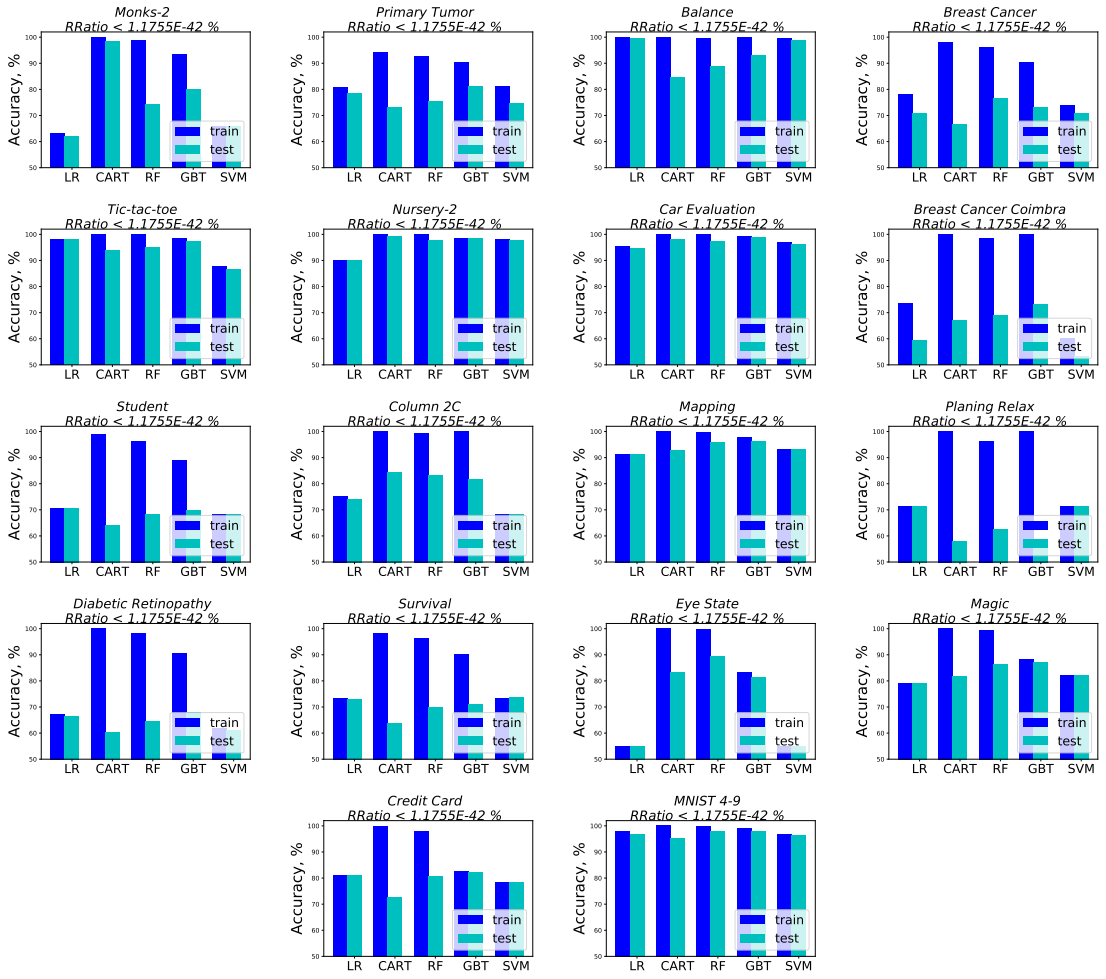


Figure 22: Performance of five machine learning algorithms without regularization for the UCI classification data sets. Data sets are listed in decreasing order of the Rashomon ratio continuing from Figure 21. Rashomon ratios, train and test accuracies are averaged over ten folds for data sets with more than 200 points and over five folds for data sets with less than 200 points

.

18:1–78, 2018.

Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. Machine bias. Available from:, May 2016.

Peter L Bartlett and Shahar Mendelson. Rademacher and Gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3(Nov):463–482, 2002.

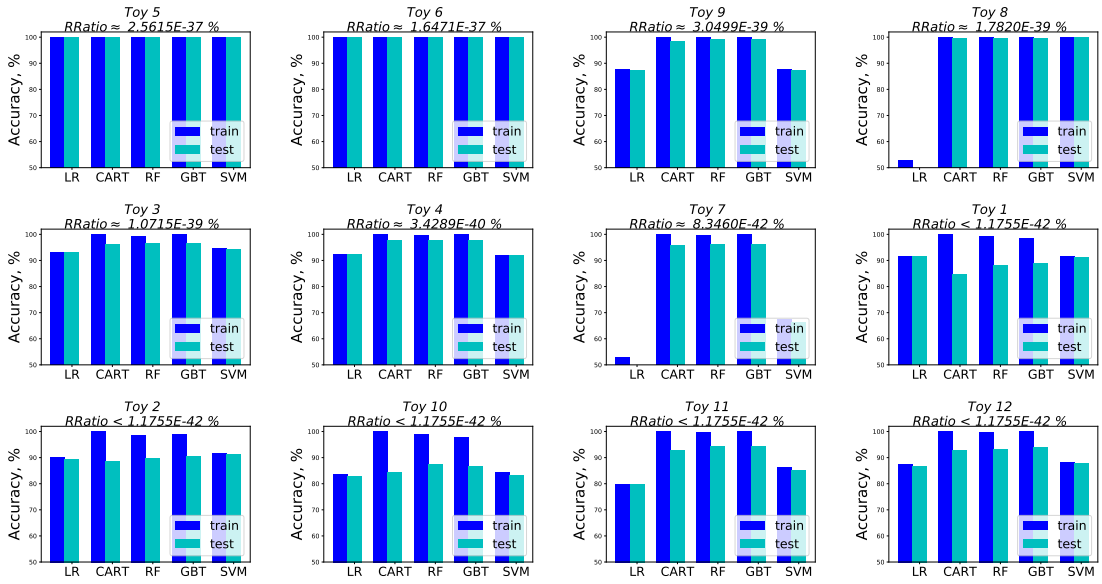


Figure 23: Performance of five machine learning algorithms without regularization for the synthetic data sets with real-valued features. Data sets are listed in decreasing order of the Rashomon ratio. Rashomon ratios, train and test accuracies are averaged over ten folds.

Peter L Bartlett, Olivier Bousquet, Shahar Mendelson, et al. Local Rademacher complexities. *The Annals of Statistics*, 33(4):1497–1537, 2005.

Olivier Bousquet and André Elisseeff. Stability and generalization. *Journal of machine learning research*, 2(Mar):499–526, 2002.

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

Christopher JC Burges. A tutorial on support vector machines for pattern recognition. *Data mining and knowledge discovery*, 2(2):121–167, 1998.

Xiao Cai, Feiping Nie, Heng Huang, and Chris Ding. Multi-class L2, 1-norm support vector machine. In *Data Mining (ICDM), 2011 IEEE 11th International Conference on*, pages 91–100. IEEE, 2011.

Feilong Cao, Tingfan Xie, and Zongben Xu. The estimate for approximation error of neural networks: A constructive approach. *Neurocomputing*, 71(4-6):626–630, 2008.

Rich Caruana, Yin Lou, Johannes Gehrke, Paul Koch, Mark Sterm, and Noémie Elhadad. Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. In *Proceedings of Knowledge Discovery in Databases (KDD)*, pages 1721–1730, 2015.

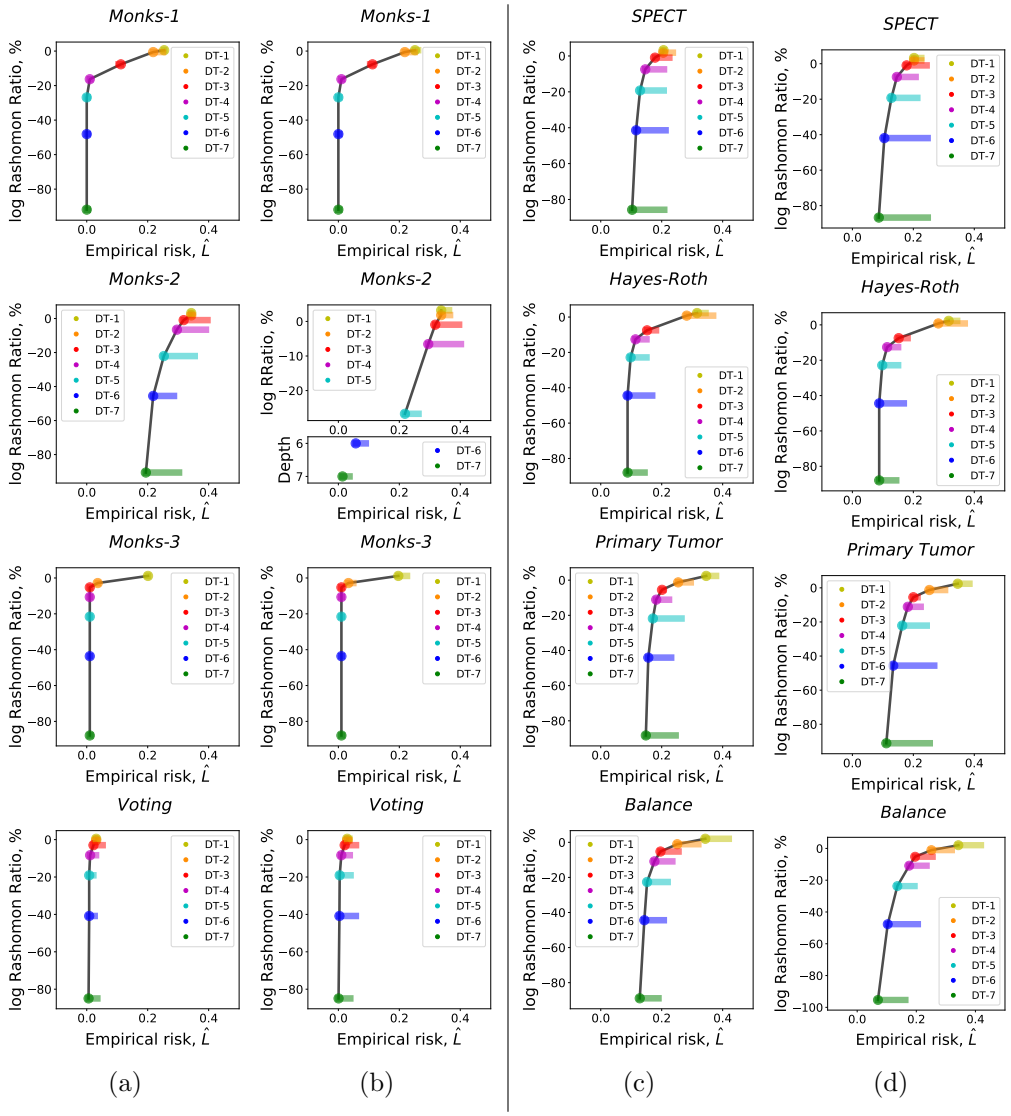


Figure 24: (a),(c)—The Rashomon curves that illustrate the generalization ability of the Rashomon elbow for the UCI classification data sets with categorical features. The hierarchy of hypothesis spaces are fully grown decision trees from depth one to seven sampled with importance sampling on leaves. (b),(d)—The Rashomon curves based on best optimal model among the sampled trees and CART model. The lower subplot shows accuracies of CART trees in cases, when we were not able to sample. The Rashomon parameter θ was set to be 0.05 for all experiments. We averaged empirical risks and Rashomon ratios over ten folds.

Pratik Chaudhari, Anna Choromanska, Stefano Soatto, Yann LeCun, Carlo Baldassi, Christian Borgs, Jennifer Chayes, Levent Sagun, and Riccardo Zecchina. Entropy-SGD: Biasing gradient descent into wide valleys. *arXiv preprint arXiv:1611.01838*, 2016.

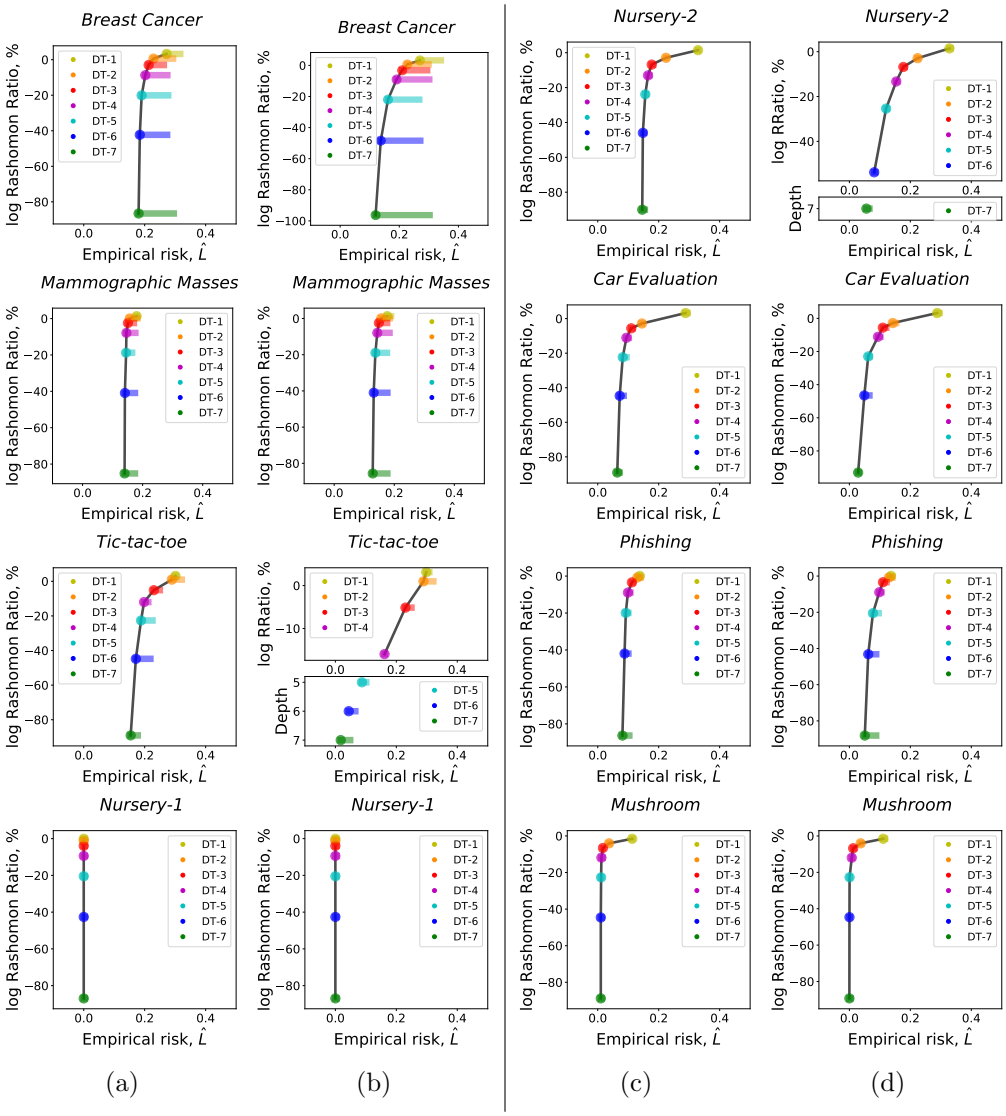


Figure 25: The Rashomon curves that illustrate the generalization ability of the Rashomon elbow for the UCI classification data sets with categorical features. The hierarchy of hypothesis spaces are fully grown decision trees from depth one to seven sampled with importance sampling on leaves. (b),(d)—The Rashomon curves based on best optimal model among the sampled trees and CART model. The lower subplot shows accuracies of CART trees in cases, when we were not able to sample. The Rashomon parameter θ was set to be 0.05 for all experiments. We averaged empirical risks and Rashomon ratios over ten folds.

Chaofan Chen, Kancheng Lin, Cynthia Rudin, Yaron Shaposhnik, Sijia Wang, and Tong Wang. An interpretable model with globally consistent explanations for credit risk. In *Proceedings of NeurIPS 2018 Workshop on Challenges and Opportunities for AI in Fi-*

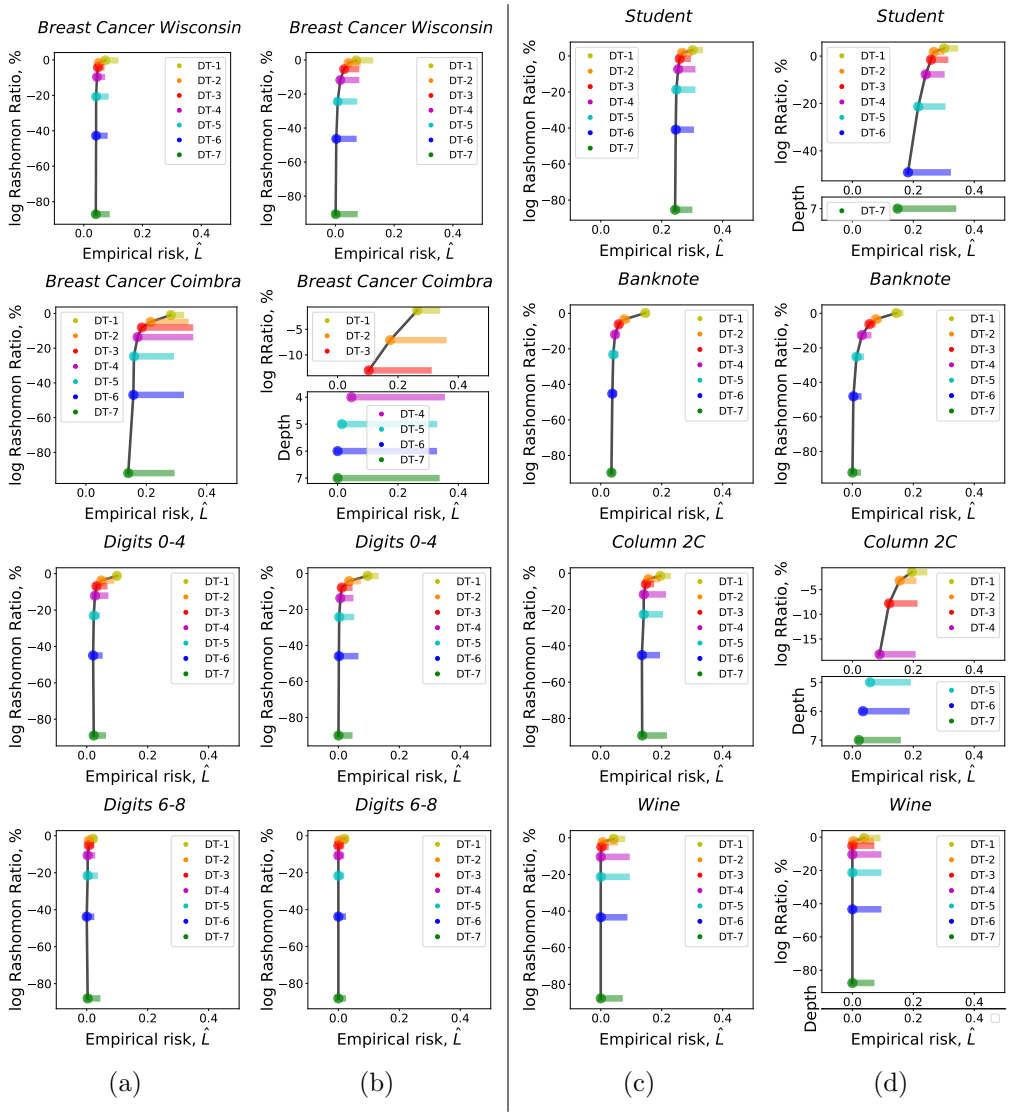


Figure 26: (a)-(c)—The generalization ability of the Rashomon elbow for the UCI classification data sets with real-valued features. The hierarchy of hypothesis spaces are fully grown decision trees from depth one to seven sampled with importance sampling on leaves. (b),(d)—The Rashomon curves based on best optimal model among the sampled trees and CART model. The lower subplot shows accuracies of CART trees in cases, when we were not able to sample. The Rashomon parameter θ was set to be 0.05 for all experiments. We averaged empirical risks and Rashomon ratios over ten folds.

nancial Services: the Impact of Fairness, Explainability, Accuracy, and Privacy, 2018.

Beau Coker, Cynthia Rudin, and Gary King. A theory of statistical inference for ensuring the robustness of scientific results. *arXiv preprint arXiv:1804.08646*, 2018.

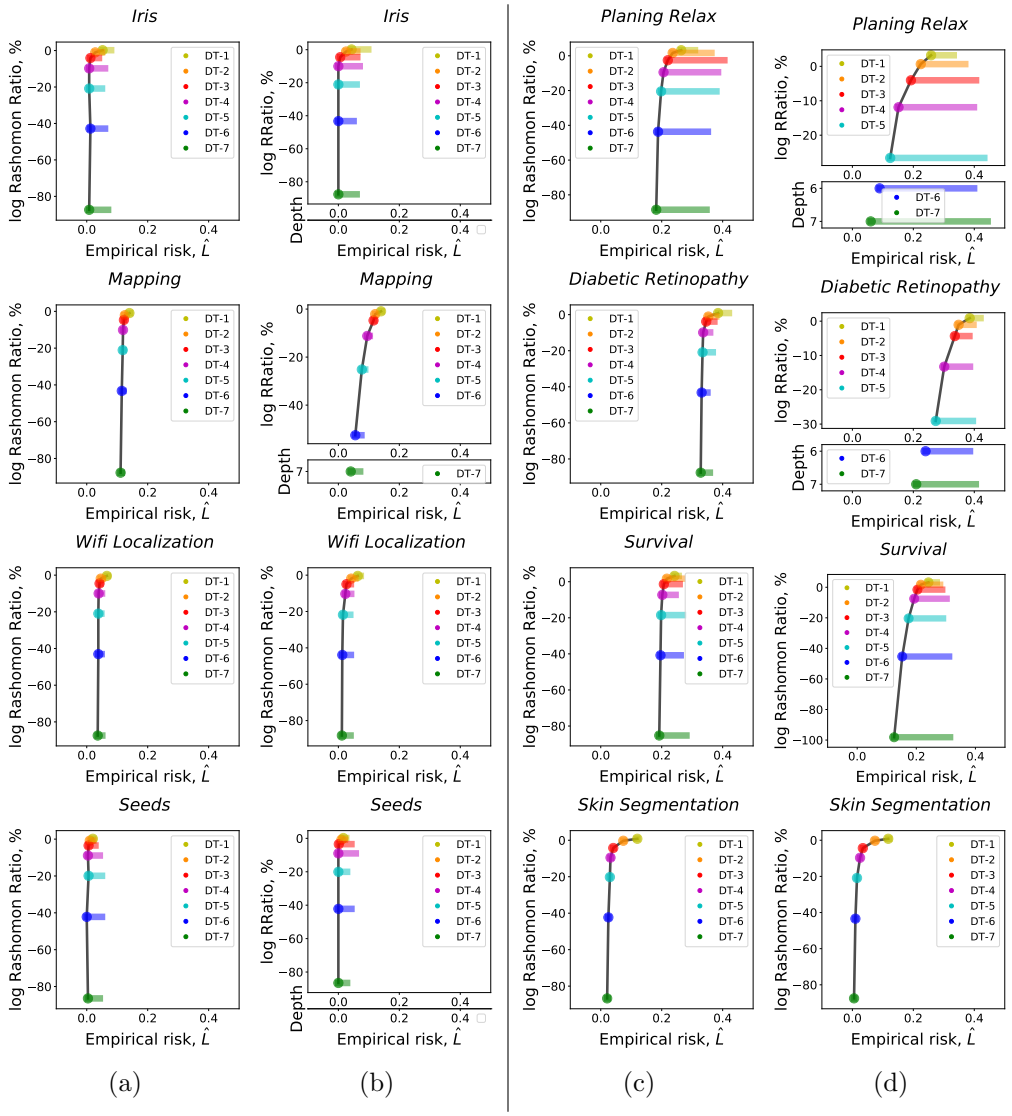


Figure 27: (a),(c)—The generalization ability of the Rashomon elbow for the UCI classification data sets with real-valued features. TThe hierarchy of hypothesis spaces are fully grown decision trees from depth one to seven sampled with importance sampling on leaves. (b),(d)—The Rashomon curves based on best optimal model among the sampled trees and CART model. The lower subplot shows accuracies of CART trees in cases, when we were not able to sample. The Rashomon parameter θ was set to be 0.05 for all experiments. We averaged empirical risks and Rashomon ratios over ten folds.

Oleg Davydov. Algorithms and error bounds for multivariate piecewise constant approximation. In *Approximation Algorithms for Complex Systems*, pages 27–45. Springer, 2011.

Ronald A DeVore. Nonlinear approximation. *Acta numerica*, 7:51–150, 1998.

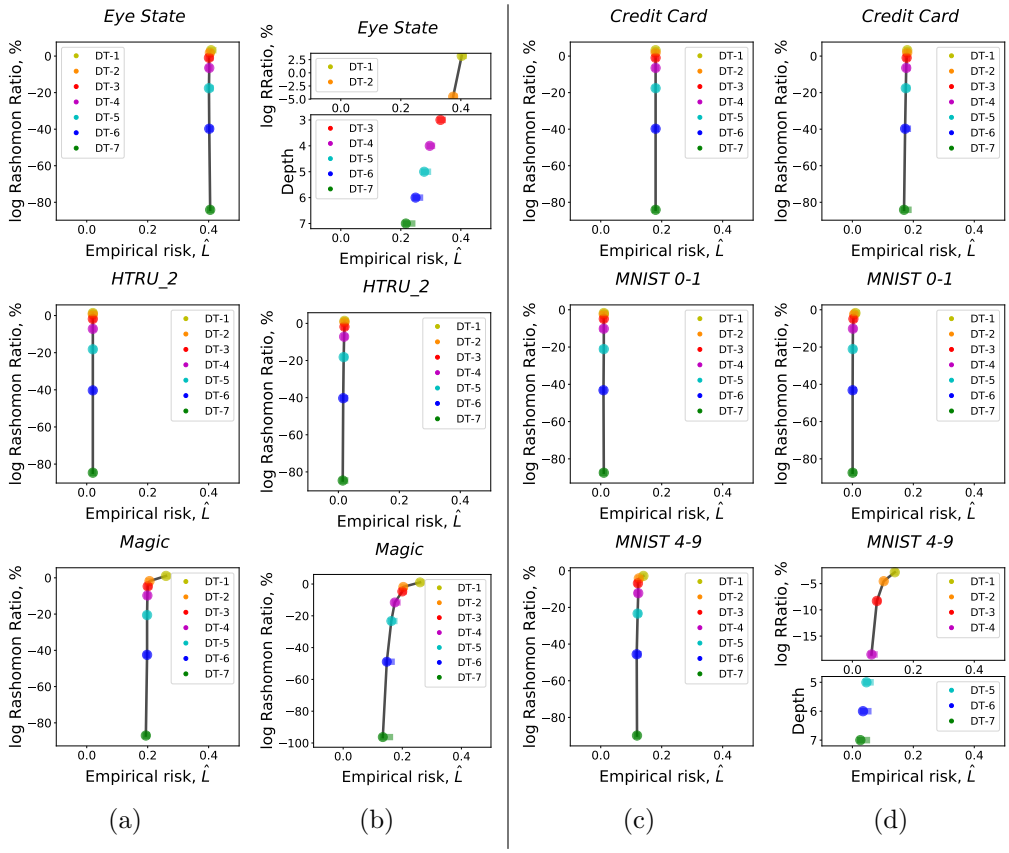


Figure 28: (a),(c)—The generalization ability of the Rashomon elbow for the UCI classification data sets with real-valued features. The hierarchy of hypothesis spaces are fully grown decision trees from depth one to seven sampled with importance sampling on leaves. (b),(d)—The Rashomon curves based on best optimal model among the sampled trees and CART model. The lower subplot shows accuracies of CART trees in cases, when we were not able to sample. The Rashomon parameter θ was set to be 0.05 for all experiments. We averaged empirical risks and Rashomon ratios over ten folds.

Laurent Dinh, Razvan Pascanu, Samy Bengio, and Yoshua Bengio. Sharp minima can generalize for deep nets. *arXiv preprint arXiv:1703.04933*, 2017.

Dheeru Dua and Casey Graff. UCI machine learning repository, 2019.

Aaron Fisher, Cynthia Rudin, and Francesca 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*, 20(177):1–81, 2019.

David Galvin. Three tutorial lectures on entropy and counting. *arXiv preprint arXiv:1406.7872*, 2014.

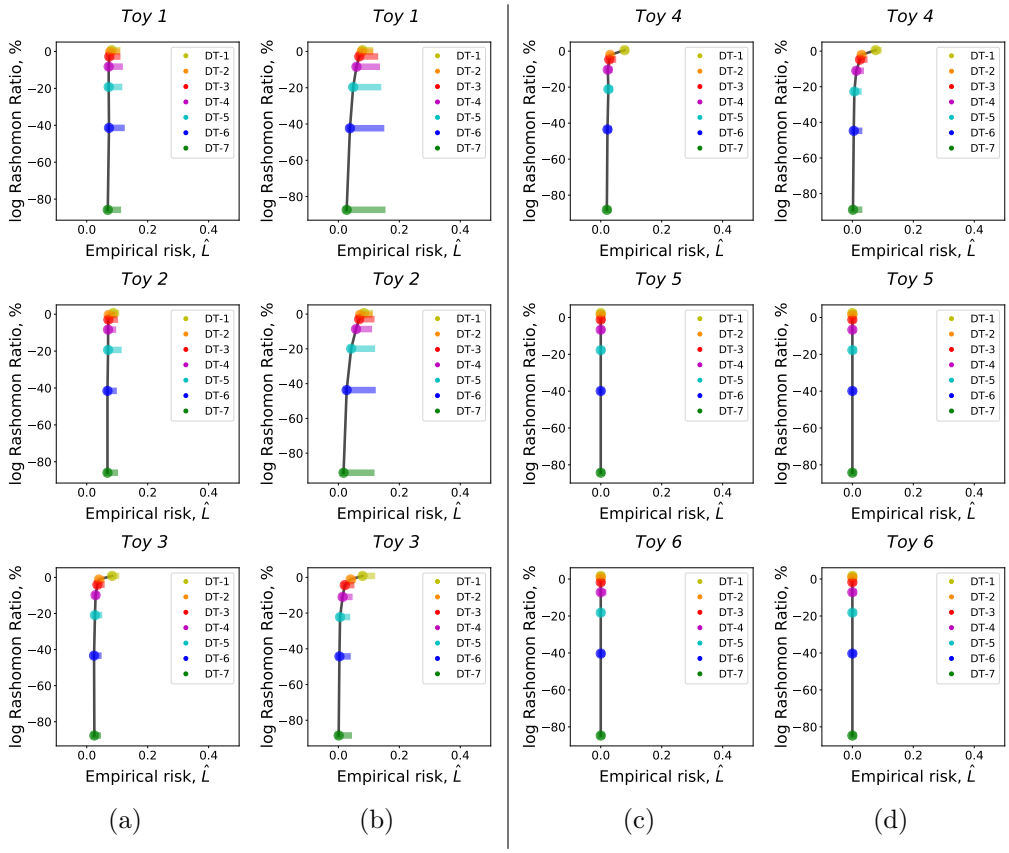


Figure 29: (a),(c)—The Rashomon curves that illustrate the generalization ability of the Rashomon elbow for synthetic data sets with two real-valued features. The hierarchy of hypothesis spaces are fully grown decision trees from depth one to seven sampled with importance sampling on leaves. (b),(d)—The Rashomon curves based on best optimal model among the sampled trees and CART model. The lower subplot shows accuracies of CART trees in cases, when we were not able to sample. The Rashomon parameter θ was set to be 0.05 for all experiments. We averaged empirical risks and Rashomon ratios over ten folds.

Martin Grötschel, László Lovász, and Alexander Schrijver. *Geometric algorithms and combinatorial optimization*, volume 2. Springer Science & Business Media, 2012.

Sepp Hochreiter and Jürgen Schmidhuber. Flat minima. *Neural Computation*, 9(1):1–42, 1997.

Sham M Kakade, Karthik Sridharan, and Ambuj Tewari. On the complexity of linear prediction: Risk bounds, margin bounds, and regularization. In *Advances in Neural Information Processing Systems*, pages 793–800, 2009.

Ravi Kannan, László Lovász, and Miklós Simonovits. Random walks and an $O^*(n^5)$ volume algorithm for convex bodies. *Random Structures & Algorithms*, 11(1):1–50, 1997.

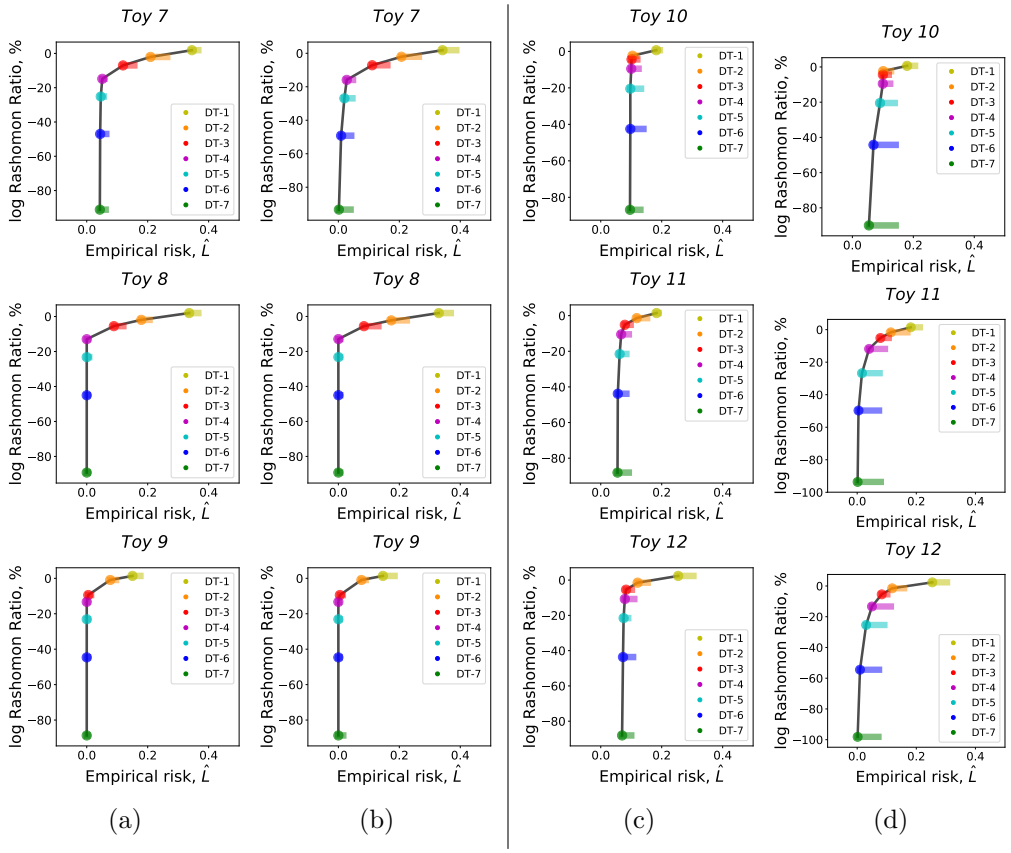


Figure 30: (a),(c)—The Rashomon curves that illustrate the generalization ability of the Rashomon elbow for synthetic data sets with two real-valued features. The hierarchy of hypothesis spaces are fully grown decision trees from depth one to seven sampled with importance sampling on leaves. (b),(d)—The Rashomon curves based on best optimal model among the sampled trees and CART model. The lower subplot shows accuracies of CART trees in cases, when we were not able to sample. The Rashomon parameter θ was set to be 0.05 for all experiments. We averaged empirical risks and Rashomon ratios over ten folds.

Michael Kearns and Dana Ron. Algorithmic stability and sanity-check bounds for leave-one-out cross-validation. *Neural Computation*, 11(6):1427–1453, 1999.

Nitish Shirish Keskar, Dheevatsa Mudigere, Jorge Nocedal, Mikhail Smelyanskiy, and Ping Tak Peter Tang. On large-batch training for deep learning: Generalization gap and sharp minima. *arXiv preprint arXiv:1609.04836*, 2016.

Vladimir Koltchinskii, Dmitry Panchenko, et al. Empirical margin distributions and bounding the generalization error of combined classifiers. *The Annals of Statistics*, 30(1):1–50, 2002.

Akira Kurosawa. *Rashomon*. Tokyo: Daiei, 1950.

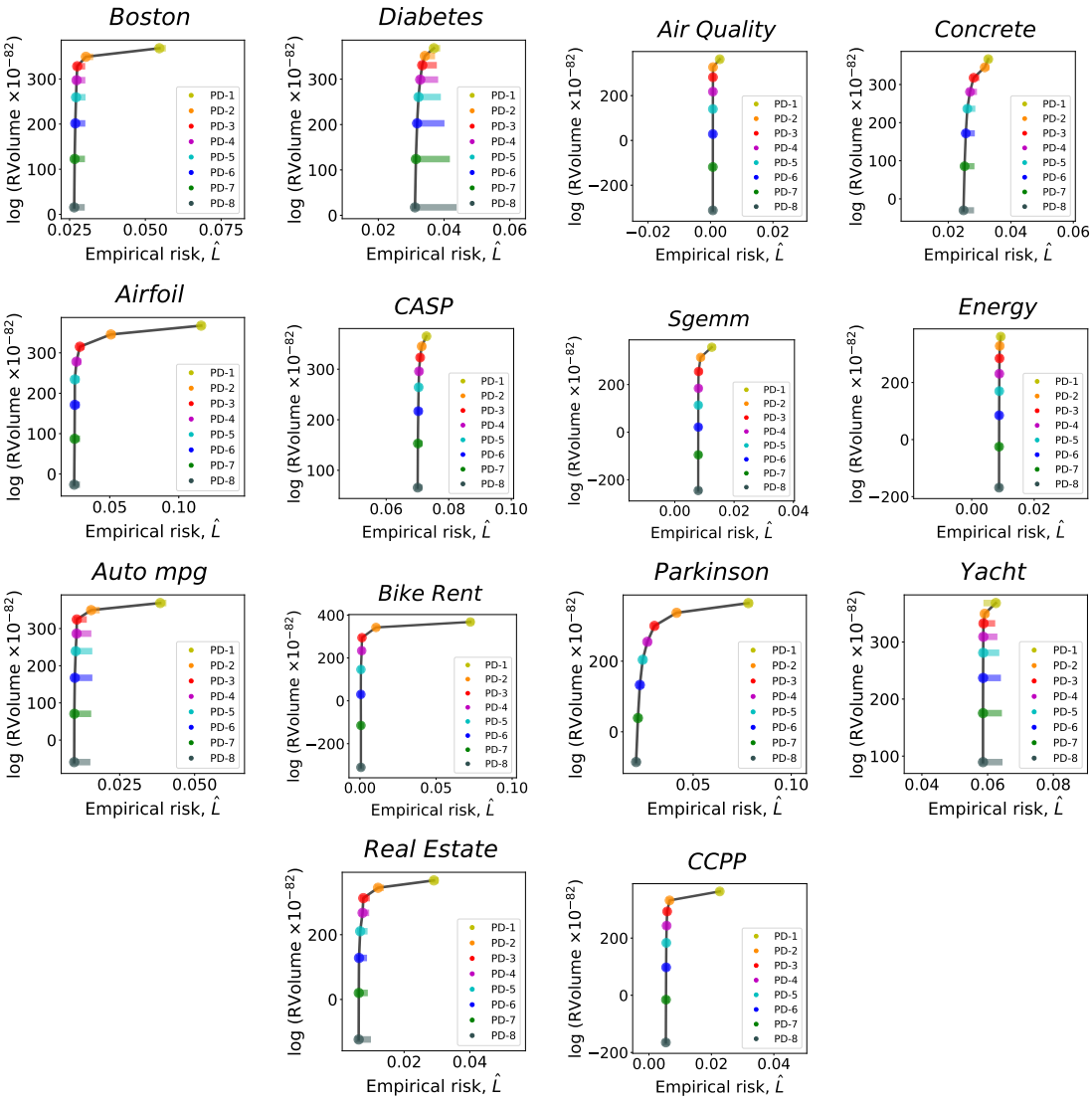


Figure 31: The generalization ability of the Rashomon elbow for the UCI regression data sets with real features. We consider only three features for every data set with the largest corresponding singular values. The hierarchy of hypothesis spaces consists of polynomials of degree (PD) from one to eight. The Rashomon parameter θ was set to be $0.1\hat{L}(\hat{f}_t)$, $t \in [1, 8]$ for all experiments. The regularization parameter for ridge regression was 0.1. We averaged over ten folds train and test least squares loss as well as Rashomon volumes.

John Langford and John Shawe-Taylor. PAC-Bayes & margins. In *Advances in Neural Information Processing Systems*, pages 439–446, 2003.

Guillaume Lecué. *Interplay between concentration, complexity and geometry in learning theory with applications to high dimensional data analysis*. PhD thesis, Université Paris-Est, 2011.

Benjamin Letham, Cynthia Rudin, Tyler H. McCormick, and David Madigan. Interpretable classifiers using rules and bayesian analysis: Building a better stroke prediction model. *Annals of Applied Statistics*, 9(3):1350–1371, 2015.

Benjamin Letham, Portia A. Letham, Cynthia Rudin, and Edward Browne. Prediction uncertainty and optimal experimental design for learning dynamical systems. *Chaos*, 26 (6), 2016.

Gábor Lugosi and Andrew B Nobel. Adaptive model selection using empirical complexities. *The Annals of Statistics*, 27(6):1830–1864, 1999.

Gábor Lugosi and Marten Wegkamp. Complexity regularization via localized random penalties. *The Annals of Statistics*, 32(4):1679–1697, 2004.

Florence Jessie MacWilliams and Neil James Alexander Sloane. *The theory of error-correcting codes*, volume 16. Elsevier, 1977.

Nicolai Meinshausen and Peter Bühlmann. Stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(4):417–473, 2010.

Shahar Mendelson. A few notes on statistical learning theory. In *Advanced lectures on machine learning*, pages 1–40. Springer, 2003.

Daniel Nevo and Ya'acov Ritov. Identifying a minimal class of models for high-dimensional data. *The Journal of Machine Learning Research*, 18(1):797–825, 2017.

DJ Newman and TJ Rivlin. Approximation of monomials by lower degree polynomials. *Aequationes Mathematicae*, 14(3):451–455, 1976.

Ramamohan Paturi. On the degree of polynomials that approximate symmetric boolean functions (preliminary version). In *Proceedings of the Twenty-Fourth Annual ACM Symposium on Theory of Computing*, pages 468–474. ACM, 1992.

Narges Razavian, Saul Blecker, Ann Marie Schmidt, Aaron Smith-McLallen, Somesh Nigam, and David Sontag. Population-level prediction of type 2 diabetes from claims data and analysis of risk factors. *Big Data*, 3(4):277–287, 2015.

William H Rogers and Terry J Wagner. A finite sample distribution-free performance bound for local discrimination rules. *The Annals of Statistics*, pages 506–514, 1978.

Cynthia Rudin, Caroline Wang, and Beau Coker. The age of secrecy and unfairness in recidivism prediction. *Harvard Data Science Review*, 2019. (accepted).

Robert E Schapire, Yoav Freund, Peter Bartlett, Wee Sun Lee, et al. Boosting the margin: A new explanation for the effectiveness of voting methods. *The Annals of Statistics*, 26 (5):1651–1686, 1998.

John Shawe-Taylor, Peter L Bartlett, Robert C Williamson, and Martin Anthony. Structural risk minimization over data-dependent hierarchies. *IEEE Transactions on Information Theory*, 44(5):1926–1940, 1998.

Nathan Srebro, Karthik Sridharan, and Ambuj Tewari. Smoothness, low noise and fast rates. In *Advances in Neural Information Processing Systems*, pages 2199–2207, 2010.

Nikolaj Tollenaar and P.G.M. van der Heijden. Which method predicts recidivism best?: a comparison of statistical, machine learning and data mining predictive models. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 176(2):565–584, 2013.

Theja Tulabandhula and Cynthia Rudin. Machine learning with operational costs. *The Journal of Machine Learning Research*, 14(1):1989–2028, 2013.

Theja Tulabandhula and Cynthia Rudin. On combining machine learning with decision making. *Machine Learning (ECML-PKDD journal track)*, 97(1-2):33–64, 2014a.

Theja Tulabandhula and Cynthia Rudin. Robust optimization using machine learning for uncertainty sets. *arXiv preprint arXiv:1407.1097*, 2014b.

Berk Ustun and Cynthia Rudin. Supersparse Linear Integer Models for Optimized Medical Scoring Systems. *Machine Learning*, 102(3):349–391, 2016.

Vladimir N Vapnik. *The Nature of Statistical Learning Theory*. Springer, 1995.

VN Vapnik and A Ya Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and its Applications*, 16(2):264, 1971.

Christopher S Wallace and David M Boulton. An information measure for classification. *The Computer Journal*, 11(2):185–194, 1968.

Jiaming Zeng, Berk Ustun, and Cynthia Rudin. Interpretable classification models for recidivism prediction. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 180(3):689–722, 2017.

Ding-Xuan Zhou. The covering number in learning theory. *Journal of Complexity*, 18(3): 739–767, 2002.

Ji Zhu, Saharon Rosset, Robert Tibshirani, and Trevor J Hastie. 1-norm support vector machines. In *Advances in Neural Information Processing Systems*, pages 49–56, 2004.

Notation	Description
n	number of points in a data set
p	number of features
S	training data set
$\mathcal{D}$	unknown data distribution
$\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$	data space; $\mathcal{X}$—input space; $\mathcal{Y}$—output space
$\mathcal{F}$	hypothesis space
$\rho(\mathcal{F})$	prior distribution over $\mathcal{F}$
$d(\cdot, \cdot)$	metric, defined on a metric space $\mathcal{F}$
$\|\cdot\|_p$	p-norm defined for elements of $\mathcal{F}$
$\phi(f(x), y)$	loss function
$l(f, z)$	loss function
$\zeta(f, z)$	perfomance function
L	true risk
f^*	optimal model
$\hat{L}$	empirical risk
$\hat{f}$	empirical risk minimizer
θ	Rashomon parameter
$\hat{R}_{set}(\mathcal{F}, \theta)$	Rashomon set
$\mathcal{V}(\hat{R}_{set}(\mathcal{F}, \theta))$	Rashomon volume
$\hat{R}_{ratio}(\mathcal{F}, \theta)$	Rashomon ratio
$\hat{R}^{pat}_{ratio}(\mathcal{F}, \theta)$	pattern Rashomon ratio
γ	anchored Rashomon parameter
$R^{anc}_{set}(\mathcal{F}, \gamma)$	true anchored Rashomon set
$\hat{R}^{anc}_{set}(\mathcal{F}, \gamma)$	anchored Rashomon set
$\hat{R}^{anc}_{ratio}(\mathcal{F}, \gamma)$	anchored Rashomon ratio
$R^{anc}_{ratio}(\mathcal{F}, \gamma)$	true anchored Rashomon ratio
K	Lipschitz constant
$B_\delta(f')$	δ-ball centered at f'
$\mathcal{B}(\mathcal{F}, \delta)$	packing number
$\mathcal{N}(\mathcal{F}, \delta)$	covering number
$R_n\mathcal{F}$	Rademacher complexity
$\hat{R}^S_n\mathcal{F}$	empirical Rademacher complexity computed on a data set S
Ω	parameter space
$\mathcal{F}_\Omega$	parameterized hypothesis space
$\hat{\omega}$	parameter in the parameter space that minimized empirical risk
C	regularization parameter
σ_i	singular values of the feature matrix
DT-D	fully grown decision tree of depth D
$\mathcal{A}$	a learning algorithm

Table 4: List of symbols and notation used in this paper

Data Set Name	Type of Features	Number of Features	Number of Data Points	Processing notes
Monks-1	Binary	15	556	
Monks-2	Binary	15	601	
Monks-3	Binary	15	554	
Voting	Binary	16	232	
SPECT	Binary	22	267	
Tic-tac-toe	Binary	27	958	
Hayes-Roth	Binary	12	160	Considered class 1 versus classes 2 and 3
Nursery-1	Binary	27	8586	Considered classes not_recom and priority
Nursery-2	Binary	27	8310	Considered classes priority and spec_prior
Mushroom	Binary	117	8124	
Breast Cancer	Binary	43	286	
Car Evaluation	Binary	21	1728	Converted to one vs all problem: class 1 versus all others
Primary Tumor	Binary	31	336	Converted to binary classification by considering classes 1, 2, 3, 4, 22, 10 versus all others
Mammographic Masses	Binary	25	830	
Phishing	Binary	23	1353	Considered classes 0 and 1 versus class -1
Balance	Binary	20	576	Considered classes L and R
Wine	Real	13	130	Considered classes 0 and 1
Iris	Real	4	100	Considered classes versicolour and viginica
Breast Cancer Wisconsin	Real	30	569	
Breast Cancer Coimbra	Real	9	116	
Digits 0-4	Real	64	363	Classes 0 and 4 considered only
Digits 6-8	Real	64	355	Classes 6 and 8 considered only
Student	Real	3	400	
Banknote	Real	4	1372	
Mapping	Real	28	10545	Converted to one vs all problem: class forest versus all others
Wifi Localization	Real	7	1000	Considered classes that represent rooms 2 and 3
Column 2C	Real	6	310	
Credit Card	Real	23	30000	
Planing Relax	Real	12	182	
Diabetic Retinopathy	Real	19	1151	
Survival	Real	3	306	
Skin Segmentation	Real	3	245057	
HTRU_2	Real	8	17898	
Magic	Real	10	19020	
Seeds	Real	7	140	Considered classes 1 and 2
Eye State	Real	14	14980	
MNIST 0-1	Real	784	13738	Considered classes 0 and 1
MNIST 4-9	Real	784	12752	Considered classes 4 and 9

Table 5: Classification data sets description and processing notes.

Data Set Name	Number of Features	Number of Data Points	Processing notes
Boston	13	506	
Diabetes	10	442	
Air quality	11	6941	Dropped features "Date", "Time", "NMHC(GT)" and entries with missing values
Concrete	8	1030	
Airfoil	5	1502	
CASP	9	45730	
Sgemm	14	241600	Dropped targets "Run2 (ms)", "Run3 (ms)", "Run4 (ms)"
Energy	24	19735	Dropped features "date", random features "rv1", "rv2", target "lights"
Auto mpg	7	392	Dropped car name feature and entries with missing values
Bike Rent	13	731	Dropped features "instant", "dteday"
Parkinson	20	5875	Dropped target "motor_UPDRS"
Yacht	6	308	
Real Estate	6	414	
CCPP	4	9568	

Table 6: Regression data sets description and processing notes.