In Pursuit of Simplicity: The Role of the Rashomon Effect for Informed Decision Making

by

Lesia Semenova

Department of Computer Science
Duke University

Defense Date: March 19, 2024

Approved:

Cynthia Rudin, Supervisor

Ronald Parr, Supervisor

Carlo Tomasi

Rong Ge

Dissertation submitted in partial fulfillment of the requirements for the degree of
Doctor of Philosophy in the Department of Computer Science
in the Graduate School of Duke University
2024

ABSTRACT

In Pursuit of Simplicity: The Role of the Rashomon Effect for Informed Decision Making

by

Lesia Semenova

Department of Computer Science
Duke University

Defense Date: March 19, 2024

Approved:

Cynthia Rudin, Supervisor

Ronald Parr, Supervisor

Carlo Tomasi

Rong Ge

An abstract of a dissertation submitted in partial fulfillment of the requirements for
the degree of Doctor of Philosophy in the Department of Computer Science
in the Graduate School of Duke University
2024

Abstract

For high-stakes decision domains, such as healthcare, lending, and criminal justice, the predictions of deployed models can have a huge impact on human lives. The understanding of why models make specific predictions is as crucial as the good performance of these models. Interpretable models, constrained to explain the reasoning behind their decisions, play a key role in enabling users' trust. They can also assist in troubleshooting and identifying errors or data biases. However, there has been a longstanding belief in the community that a trade-off exists between accuracy and interpretability. We formally show that such a trade-off does not exist for many datasets in high-stakes decision domains and that simpler models often perform as well as black-boxes.

To establish a theoretical foundation explaining the existence of simple-yet-accurate models, we leverage the Rashomon set (a set of equally well-performing models). If the Rashomon set is large, it contains numerous accurate models, and perhaps at least one of them is the simple model we desire. We formally present the Rashomon ratio as a new gauge of simplicity for a learning problem, where the Rashomon ratio is the fraction of all models in a given hypothesis space that is in the Rashomon set. Insight from studying the Rashomon ratio provides an easy way to check whether a simpler model might exist for a problem before finding it. In that sense, the Rashomon ratio is a powerful tool for understanding when an accurate-yet-simple model might exist. We further propose and study a mechanism of the data generation process, coupled with choices usually made by the analyst during the learning process, that determines the size of the Rashomon ratio. Specifically, we demonstrate that noisier datasets lead to larger Rashomon ratios through the way practitioners train models. Our results explain a key aspect of why simpler models often tend to perform as well as black box models on complex, noisier datasets.

Given that optimizing for interpretable models is known to be NP-hard and can require significant domain expertise, our foundation can help machine learning practitioners assess the feasibility of finding simple-yet-accurate models before attempting to optimize for them. We illustrate how larger Rashomon sets and noise in the data generation process explain

the natural gravitation towards simpler models based on the dataset of complex biology. We further highlight how simplicity is useful for informed decision-making by introducing sparse density trees and lists – an accurate approach to density estimation that optimizes for sparsity.

Dedication

To the people of Ukraine.

Contents

List of Tables

List of Figures

Acknowledgements

I extend my heartfelt thank you to the many individuals who have supported me throughout my Ph.D. journey. Their encouragement and friendship have been invaluable. Particularly, I am especially grateful to the following people.

I want to express my deepest gratitude to my advisors, Cynthia Rudin and Ronald Parr, for their guidance, mentorship, and support throughout my Ph.D. career. I learned a lot from them; their vision, expertise, and constructive feedback were invaluable in shaping the direction and content of my work. Thank you to Cynthia for being such an inspirational visionary in everything she does. As we worked on projects together, I learned how to sharpen ideas, write an effective introduction, and simplify challenging problems until we could solve them and then scale back up. I am especially grateful for the opportunity to coach student teams in data science competitions alongside her. Thank you to Ron for teaching me how to think about corner cases of research problems and probing solution ideas from different directions. I am very grateful to Ron for all the discussions on reinforcement learning, mentorship meetings, and constant unwavering support through all these years, especially during the most challenging times for me.

I thank my committee members, Rong Ge and Carlo Tomasi, for the insightful comments, suggestions, and valuable discussions. Thank you to Edward Browne for providing us with interesting data that started a series of collaborations and helped me learn more about effective data analysis. Thank you to Marilyn Butler for making the department feel like a family. Thank you to Susan Rodger for the mentorship during my first time as a TA at Duke and for always supporting Duke's ACM-W chapter. Thank you to Shaundra Daily for mentoring the Ph.D. accountability group.

I want to thank my collaborators who contributed in parts to my dissertation, including Harry Chen, Siong Thye Goh, Cynthia Rudin, Ronald Parr, Edward Browne, Yingfan Wang, Shane Falcinelli, David Murdoch, Alicia Volkheimer, Ethan Wu, Alexander Richardson, Manickam Ashokkumar, David Margolis, Nancie Archin, Nilu Goonetilleke, Chaofan Chen, Zhi Chen, Haiyang Huang, Chudi Zhong. Thank you to my collaborators with whom I

worked on other exciting research projects, beyond the scope of this dissertation, including Dennis Tang, Frank Willard, Ronan Tegerdine, Luke Triplett, Jon Donnelly, Luke Moffett, Alina Jade Barnett, Jin Jing, Brandon Westover, Gaurav Rajesh Parikh, Jenny Huang, Albert Sun, Chloe Qinyu Zhu, Muhang Tian, Jiachang Liu, Jack Xu, Joseph Scarpa, Alex Oesterling, Angikar Ghosal, Haoyang Yu, Rui Xin, Yasa Baig, Allan Guo, Eric Song.

A special thank you to my other labmates and friends at Duke, Zack Boner, Harsh Parikh, Marco Morucci, Tianyu Wang, Srikar Katta, Rachel Draelos, Mark Nemecek, Xiaonan Hu, Dillon Sandhu, Barrett Ames, Shuzhi Yu, Bonnie Chen, Sneha Mitra, Vincentius Martin, Shuai Yuan, Hannah Kim, Ergys Ristani, Kelsey Lieberman, Kate O'Hanlon, Alper Bozkurt, Nisarg Raval, Shweta Patwa, for research discussions, chats about life, and for making my Ph.D. journey filled with friendship. Thank you to Sneha for knowing all the great places to eat out. Thank you to Shuzhi for sharing the Ph.D. journey with me since day one. Thank you to Alper for all the wonderful coffee breaks and support. Thank you to Nisarg for being a great officemate, friend, and mentor.

I appreciate the two summer research internships I did at Pinterest Labs. I would like to thank my mentors Aditya Pal and Vishwakarma Singh for their guidance and support. I am also grateful to Chuck Rosenberg, Nikil Pancha, Pong Eksombatchai, and Nadia Fawaz for their valuable discussions and insights.

I am very grateful to my family for their love and support throughout my academic journey and life. To my husband, Ruslan, your patience, understanding, and love sustained me through all the challenges. Thank you to my mom, Oksana, whose belief in me was a constant source of motivation, and to my brother, Dmytro, who always finds the right words to make me smile. Thank you to Victoria, Georgiy, Artem, and Uliana – my family-in-law – for their continuous encouragement. A special thank you to my grandparents for being my biggest supporters. I am truly blessed to have such a supportive and caring family.

1. Introduction

The increasing availability of complex datasets, especially in high-stakes decision domains such as criminal justice, healthcare, and lending, underscores the importance of bridging the gap between data complexity and the human ability to understand these data. The quality of insights derived during the initial analysis can aid in mitigating potential biases, outliers, or errors and help in the decision-making process. Simpler or interpretable machine learning models can not only help in making these insights clearer but also can enable trust in artificial intelligence systems (Rudin, 2019).

As models that are inherently contained so that their reasoning is understandable to humans, interpretable models are easier to troubleshoot, provide truthful and complete explanations, and facilitate interactions with domain experts which can lead to a better model in the end. However, due to the belief in the trade-off between accuracy and interpretability, often the black-box complex models are used in high-stakes decision domains. In this dissertation, we show that by carefully studying data and data generation processes, machine learning practitioners can make better, more informed decisions regarding complex datasets in high-stakes decision domains.

1.1 Larger Rashomon Sets Contain Simple-yet-Accurate Machine Learning Models

Following the principle of Occam's Razor, one should use the simplest model that explains the data well. However, finding the simplest model, let alone any simple-yet-accurate model, is hard. As soon as simplicity constraints such as sparsity are introduced, the optimization problem for finding a simpler model typically becomes NP-hard. Thus, practitioners – who have no assurance of finding a simpler model that achieves the performance level of a black box – may not see a reason to attempt such potentially difficult optimization problems. Thus, sadly, what was once the holy grail of finding simpler models, has been, for the most part, abandoned in modern machine learning. Therefore, we ask a question that is essential, and potentially game-changing, for this discussion: what if we knew, before attempting a

computationally expensive search for a simpler-yet-accurate model, that one was likely to exist? Perhaps knowing this would allow us to justify the time and expense of searching for such a model. If it is true that many data sets have properties to admit simple models, then there are important implications for society – it means we may be able to use simpler or interpretable models for many high-stakes problems without losing accuracy.

Proving the existence of simpler models before aiming to find them differs from the current approach to machine learning in practice. We generally do not think about going from more complicated spaces to simpler ones; in fact, the reverse is true, where typical statistical learning theory and algorithms allowed us to maintain generalization when handling more complicated model classes (e.g., large margins for support vector machines with complex kernels or large margins for boosted trees) (Cortes & Vapnik, 1995; Schapire et al., 1998). We even build neural networks that are so complex that they can achieve zero training error, and try afterwards to determine why they generalize (Belkin et al., 2019; Nakkiran et al., 2021). However, because simple models are essential for many high-stakes decisions (Rudin, 2019), perhaps we should return to the goal of aiming directly for simpler models. We will need new ideas in order to do this.

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, 1999), or measure properties of specific algorithms (e.g., algorithmic stability, see Bousquet & Elisseeff, 2002). However, none of these theories seems to capture directly 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.*

We hypothesize that these three observations can all be explained by the same phenomenon: the "Rashomon Effect", which is the existence of many almost-equally-accurate models (Breiman, 2001). Firstly, if there is a large *Rashomon set* of almost-equally-

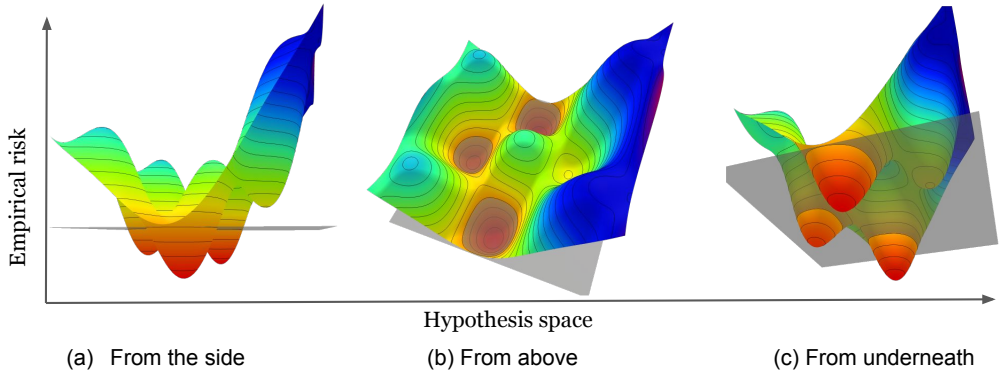

FIGURE 1.1: An illustration of a possible Rashomon set in two dimensional hypothesis space $\mathcal{F}$. Models below the gray plane belong to the Rashomon set $\hat{R}_{set}(\mathcal{F}, \theta)$, where the height of the gray plane is adjusted by the Rashomon parameter θ defined in Section 2.2.

accurate models (Figure 1.1), a simple model may also be contained in it. Secondly, if the Rashomon set is large, many different machine learning algorithms may find different but approximately-equally-well-performing models inside it. An experimenter could then observe similar performance for different types of algorithms that produce very different functions. Thirdly, if the Rashomon set is large enough to contain simpler models, those models are guaranteed to generalize well. As we will show in Chapter 3, there are mathematical assumptions that allow us to *prove* the 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.

We quantify the magnitude of the Rashomon Effect through the *Rashomon ratio*, which is the ratio of the Rashomon set's volume to the volume of the hypothesis space. An illustration of the Rashomon set is shown in Figure 1.1; it does not need to be a connected or convex set. The Rashomon ratio can serve as a gauge of simplicity for a learning problem.[1] As a property of both a data set and a hypothesis space, it differs from the VC dimension (Vapnik & Chervonenkis, 1971) (because the Rashomon ratio is specific to a data set), it

[1] Such measures are typically called "complexity" measures, but the Rashomon ratio measures simplicity, not complexity.

differs from algorithmic stability (see Kearns & Ron, 1999; Rogers & Wagner, 1978) (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 (P. L. 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, 1999) (as the maximum margin classifier can have a small minimum margin yet the Rashomon ratio can be large, 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 the Rashomon ratio disagrees with these complexity measures.The Rashomon set is not just functions within a flat minimum; it could consist of functions from many non-flat local minima as illustrated in Figure 2.1, and it applies to discrete hypothesis spaces where gradients, and thus "sharpness" (Dinh et al., 2017) do not exist. For linear regression, we derive a closed form solution for the volume of the Rashomon set in parameter space in Theorem 8 in Chapter 3.

Our theory and empirical results have implications beyond cases where the size of the Rashomon set can be estimated in practice: they suggest computationally inexpensive ways to gauge whether the Rashomon set is large without directly measuring it. *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. Thus, after running different machine learning methods and observing similar performance, our results indicate that it may be worthwhile to optimize directly for simpler models within the Rashomon set.*

1.2 There is no Simplicity-Accuracy Trade-off for a lot of High-Stakes Decision Datasets

A key question in determining the existence of simpler models is to understand why and when the Rashomon Effect happens. This is a difficult question, and there has been little study of it. The literature on the Rashomon Effect has generally been more practical,

showing either that the Rashomon Effect often exists in practice (D'Amour et al., 2022; Semenova et al., 2022; Teney et al., 2022), showing how to compute or visualize the set of good models for a given dataset (Ahanor et al., 2023; Dong & Rudin, 2020; Fisher et al., 2019; Mata et al., 2022; Wang et al., 2022; Xin et al., 2022; Yan & Zhang, 2022; Zhong et al., 2023), or trying to reduce underspecification by learning a diverse ensemble of models (Y. Lee et al., 2023; Ross et al., 2020). However, no prior works have focused on understanding what causes this phenomenon in the first place.

Our thesis is that *noise* is both a theoretical and practical motivator for the adoption of simpler models. In most of the cases, we refer to noise in the generation process that determines the labels. In noisy problems, the label is more difficult to predict. Data about humans, such as medical data or criminal justice data, are often noisy because many things worth predicting (such as whether someone will commit a crime within 2 years of release from prison, or whether someone will experience a medical condition within the next year) have inherent randomness that is tied to random processes in the world (Will the person get a new job? How will their genetics interact with their diet?). It might sound intuitive that noisy data would lead to simpler models being useful, but this is not something most machine learning practitioners have internalized – even on noisy datasets, they often use complicated, black-box models, to which post-hoc explanations are added. We show how practitioners who understand the bias-variance trade-off naturally gravitate towards more interpretable modes in the presence of noise.

We propose a *path* which begins with noise, is followed by decisions made by human analysts to compensate for that noise, and that ultimately leads to simpler models. In more detail, our path follows these steps: 1) Noise in the world leads to increased variance of the labels. 2) Higher label variance leads to worse generalization (larger differences between training and test/validation performance). 3) Poor generalization from the training set to the validation set is detected by analysts on the dataset using techniques such as cross-validation. As a result, the analyst compensates for anticipated poor test performance in a way that follows statistical learning theory. Specifically, they choose a simpler hypothesis

space, either through soft constraints (i.e., increasing regulation), hard constraints (explicit limits on model complexity, or model sparsification), or by switching to a simpler function class. Here, the analyst may lose performance on the training set but gain validation and test performance. 4) After reducing the complexity of the hypothesis space, the analyst's new hypothesis space has a larger *Rashomon ratio* than their original hypothesis space. The Rashomon ratio is the fraction of models in the function class that perform close to the empirical risk minimizer. It is the fraction of functions that performs approximately-equally-well to the best one. This set of "good" functions is called the Rashomon set, and the Rashomon ratio measures the size of the Rashomon set relative to the function class. This argument (that lower complexity function classes lead to larger Rashomon ratios) is not necessarily intuitive, but we show it empirically for 19 datasets. Additionally, we prove this holds for decision trees of various depths under natural assumptions. The argument boils down to showing that the set of non-Rashomon set models grows exponentially faster than the set of models inside the Rashomon set. As a result, since the analyst's hypothesis space now has a large Rashomon ratio, a relatively large fraction of models that are left in the simpler hypothesis are good, meaning they perform approximately as well as the best models in that hypothesis space. From that large set, the analyst may be able to find even a simpler model from a smaller space that also performs well, following the argument of Semenova et al. (2022). As a reminder, in Step 3 the analysts discovered that using a simpler model class improves test performance. This means that *these simple models attain test performance that is at least that of the more complex (often black box) models from the larger function class they used initially.*

We provide the mathematics and empirical evidence needed to establish this path in Chapter 4. Moreover, for the case of ridge regression with additive attribute noise, we prove directly that adding noise to the dataset results in an increased Rashomon ratio. Specifically, the additive noise acts as ℓ_2-regularization, thus it reduces the complexity of the hypothesis space (Step 3) and causes the Rashomon ratio to grow (Step 4).

Even if the analyst does not reduce the hypothesis space in Step 3, noise still gives us

larger Rashomon sets. We show this by introducing *pattern diversity*, the average Hamming distance between all classification patterns produced by models in the Rashomon set. We show that under increased label noise, the pattern diversity tends to increase, which implies that when there is more noise, there are more differences in model predictions, and thus, there could be more models in the Rashomon set. Hence, a much shorter version of the path also works: Noise in the world causes an increase in pattern diversity, which means there are more diverse models in the Rashomon set, including simple ones.

1.3 Data Understanding with Sparse Machine Learning Approaches

Histograms are popular piecewise constant density estimation models. They have a nice logical structure that permits interpretability, are accurate with sufficient data, and are easy to visualize in low dimensions. However, conventional histograms face limitations in higher dimensions, especially for binary or categorical data. Visualizing higher-dimensional bar plots becomes challenging, and accuracy diminishes due to insufficient data in bins. Additionally, interpretability becomes complex, obscuring important variable relationships (Goh et al., 2024). Not only do histograms become uninterpretable in high dimensions, other high-dimensional density estimation methods are also uninterpretable: flexible nonparametric approaches such as kernel density estimation simply produce a formula, and the estimated density landscape cannot be easily visualized without projecting it to one or two dimensions, in which case we would lose substantial information.

Therefore, we present sparse-density trees and rule lists, an interpretable alternative to high-dimension histograms, such as bar plots or variable bin-width histograms (e.g., see Scott, 1979; Wand, 1997). For the trees and lists methods, the leaf is comparable to a histogram bin, and the density within each leaf is estimated to be constant. In total, we present three methods: Method I – leaf-sparse density tree, Method II – branch-sparse density tree, and Method III – sparse density rule list. The Bayesian prior controls the shape of the density function with user-defined parameters. More specifically, for the leaf-sparse density tree, the user controls the number of leaves in the tree before seeing the data; for the

branch-sparse density tree – the number of branches for tree nodes; for the sparse density rule list – the number of conjectures in each node and also the length of the list.

Our methods are sparse and thus enable interpretability. They can help understand data better by providing clear and understandable representations of data distributions, making it easier to see patterns and anomalies, thereby facilitating more informed decision-making.

It is becoming increasingly common to demand interpretable models for high-stakes decision domains (criminal justice, healthcare, etc.) for *policy* reasons such as fairness or transparency. Our work is possibly the first to show that the inherent properties of many high-stakes decision domains lead to *technical* justifications for demanding such models.

1.4 Dissertation Outline

This dissertation is organized as follows. In Chapter 2, we introduce characteristics of the Rashomon set, describe their properties, and discuss methods to compute them. Chapters 3 and 4 discuss the connection between the larger Rashomon sets (Chapter 3) and noise in the data generation processes (Chapter 4) with the existence of simpler-yet-accurate models. These chapters build a theoretical foundation for the existence of simpler-yet-accurate models in high-stakes decision domains and are based on Semenova et al. (2022), Semenova, Chen, et al. (2023), Rudin et al. (2022). In Chapter 5, we illustrate how the theoretical foundation supports decisions made by human analysts (in this case, us) to find patterns in the complex biology dataset. We work with the data of people with HIV and try to understand the connection between the immune and demographic parameters and the viral reservoir. We then further illustrate how sparse models are useful in discovering patterns in data by introducing piecewise constant methods for density estimation. This chapter is based on Falcinelli et al. (2023), Goh et al. (2024), and Semenova, Wang, et al. (2023). Chapter 6 concludes the dissertation, and discusses future directions.

1.5 Summary of Contributions

We summarize the contributions of this dissertation as follows:

1. We define the Rashomon ratio, pattern Rashomon ratio, and pattern diversity as important characteristics of the Rashomon set. We study the properties of these characteristics and provide several approaches for estimating the size of the Rashomon set.
2. We demonstrate that the Rashomon ratio, as a gauge of the simplicity of a machine learning problem, is different from other known complexity measures such as VC-dimension, algorithmic stability, geometric margin, and Rademacher complexity.
3. 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. Our bound in Theorem 17 is different from standard learning theory bounds that consider the distance between the true and empirical risks for the same function.
4. We show empirically that when a large Rashomon set occurs, most machine learning methods tend to perform similarly, and also in these cases, simple or sparse (yet accurate) models exist.
5. We show that noise is the theoretical and practical motivator for the existence of simpler-yet-accurate models. More specifically, we propose a path that starts with noise, leads to an increase in variance, an increase in the generalization error, a decrease in the hypothesis space, and, finally, an increase in the Rashomon ratio. We formally prove or illustrate each step for different noise models and hypothesis spaces.
6. We show that larger Rashomon sets might occur in the presence of label or feature noise, as the Rashomon set characteristics tend to increase with noise.
7. For a dataset of patients with HIV, we assess patterns and study a connection between the viral reservoir and immune and demographic variables of the patients. We further illustrate how the choices of the machine learning models are explained by larger Rashomon sets and noise in the dataset.
8. We present sparse tree-based and rule list-based density estimation methods for categorical datasets. Our methods are interpretable, higher-dimensional analogies to

variable bin-width histograms and allow us to gain insights into datasets that would be hard to reliably obtain in other ways.