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

Abstract iv

List of Tables x

List of Figures xi

Acknowledgements xiii

1 Introduction 1

1.1 Larger Rashomon Sets Contain Simple-yet-Accurate Machine Learning Models 1

1.2 There is no Simplicity-Accuracy Trade-off for a lot of High-Stakes Decision Datasets 4

1.3 Data Understanding with Sparse Machine Learning Approaches 7

1.4 Dissertation Outline 8

1.5 Summary of Contributions 8

2 Characteristics of the Rashomon Set 11

2.1 Related Work 11

2.2 Notation and Rashomon Set Definitions 14

2.3 Rashomon Ratio 16

2.3.1 Rashomon Ratio as a Simplicity Measure 17

2.3.2 Analytical Calculation of Rashomon Ratio for Ridge Regression . . . 28

2.3.3 Sampling Methods 30

2.3.4 Rashomon Ratio for the Hypothesis Space of Sparse Decision Trees . 32

2.4 Pattern Rashomon Ratio 33

2.4.1 Definition and Properties 33

2.4.2 Branch and Bound Method to Compute the Pattern Rashomon Set . 37

2.5 Pattern Diversity 40

2.5.1 Pattern Diversity and Other Metrics of the Rashomon Set 41

2.5.2 Upper Bound on Pattern Diversity 44

2.6 Rashomon Set Characteristics for Different Datasets 49

3 When Rashomon Sets are Large, Simple-yet-Accurate Models Exist 51

3.1 Rashomon Set Models: Simplicity and Generalization 51

3.1.1 The True Rashomon Set Can Be Very Helpful 53

3.1.2 Proving the Existence of Simple-yet-Accurate Models with Good Generalization 57

3.2 Larger Rashomon Ratios Correlate with Similar Performance of Machine Learning Algorithms, and Good Generalization 60

3.2.1 Experimental Design 61

3.2.2 Experimental Results 64

3.3 Quality of the Features and Rashomon Ratio 65

4 Noise as a Theoretical and Practical Motivator for the Existence of Simple-yet-Accurate Models 71

4.1 Increase in Variance due to Noise Leads to Larger Rashomon Ratios 71

4.1.1 Step 1. Noise Increases Variance 71

4.1.2 Step 2. Higher Variance Leads to Worse Generalization 82

4.1.3 Step 3. Practitioner Chooses a Simpler Hypothesis Space 87

4.1.4 Step 4. Rashomon Ratio is Larger for Simpler Spaces 87

4.2 Rashomon Ratio for Ridge Regression Increases under Additive Attribute Noise 92

4.3 Rashomon Set Characteristics in the Presence of Noise 101

4.3.1 Margin Noise is Likely to Increase Rashomon Set 101

4.3.2 Label Noise is Likely to Increase Pattern Diversity 108

4.3.3 Experiments for Rashomon Set Characteristics and Label Noise . . . 110

5 The Role of Simplicity and the Rashomon Effect for Informed Decision Making 113

5.1 Interpretable Machine Learning Approaches to Better Understand Viral Reservoir for People with HIV 113

5.1.1 Cohort Description and Feature Analysis 114

5.1.2 Data Visualization with Sparse Decision Trees 115

5.1.3 Machine Learning Models That Predict High versus Low Reservoir . 117

5.2 Sparse Density Trees and Lists for Categorical Datasets 119

5.2.1 Background and Related Work 119

5.2.2 Methods Description and Computational Optimization 122

5.2.3 Experiments and Empirical Performance Analysis 129

5.2.4 Summary for Sparse Density Trees and Lists 136

6 Conclusions 137

6.1 Practical Guidance for a Machine Learning Researcher 137

6.2 Policy Implications 139

6.3 Future Directions 139

Appendix A Performance of Different Machine Learning Algorithms and Rashomon Ratio 142

A.1 Description of Datasets Used in Chapter 3 142

A.2 Rashomon Ratio Computation and Figures for Chapter 3 142

Appendix B Rashomon Sets in the Presence of Noise 150

B.1 Bernstein's and Hoeffding's Inequalities 150

B.2 Description of Datasets Used in Chapter 4 151

B.3 Description of Cross-Validation Process in Step 3 of the Path 152

B.4 Numerical Proof that Derivative is Negative for Conjuncture 35 153

Appendix C Additional Analysis of Sparse Density Trees and Rule Lists 155

C.1 Discussion on Run time Performance 155

C.2 Recommendations on the Choice of Algorithm and Parameters 157

Bibliography 161

Biography 173

List of Tables

2.1 Comparison of Rashomon ratio and other complexity measures. 17

2.2 Rashomon set characteristics for different classification datasets based on the hypothesis space of sparse decision trees of depth four. 50

3.1 Examples of the possible usage of Theorem 19. 57

3.2 Examples of function approximation in different hypothesis spaces. 58

5.1 Participant demographic and clinical characteristics. 115

A.1 Description of the datasets used in Chapter 3 and processing notes. 145

B.1 Description of the datasets used in Chapter 4 and processing notes. 152

C.1 Run time analysis of sparse density trees and rule lists methods for different complexity datasets. 155

C.2 Description of the datasets used in Chapter 5 and processing notes. 156

C.3 Description of the parameters for sparse density trees and lists. 158

C.4 Description of priors and the model complexities for models that maximized log-likelihood during the tuning procedure described in Appendix C.1. . . . 159

List of Figures

1.1 An illustration of a possible Rashomon set in two dimensional hypothesis space. 3

2.1 Difference between volume ϵ-flatness and the Rashomon set. 12

2.2 An illustration of different Rashomon ratios with identical geometric margins. 23

2.3 An illustration of different Rashomon ratios with equivalent empirical local Rademacher complexities. 26

2.4 Volume of the Rashomon set for the two-dimensional least squares regression. 28

2.5 The pattern Rashomon set vs. the Rashomon set. 34

2.6 Illustration of how reparameterization changes pairwise disagreement metric, but does not change pattern diversity. 44

3.1 Illustrations for Theorem 17 and Theorem 19. 55

3.2 The importance of approximation and smoothness assumptions in Theorem 20. 60

3.3 Examples of experiments showing larger and smaller Rashomon ratios. . . . 63

3.4 An illustration of the influence of feature quality on the Rashomon ratio. . . 65

4.1 An illustration of how Lemma 25 rotates each of the Gaussians. 79

4.2 The variance of losses increases with margin and additive attribute noise. . . 81

4.3 Practitioner's validation process in the presence of noise for CART. 87

4.4 Practitioner's validation process in the presence of noise for gradient-boosted trees. 88

4.5 Calculation showing that the Rashomon ratio and pattern Rashomon ratio are larger for smaller hypothesis spaces. 91

4.6 The setup for Conjecture 35. 102

4.7 Illustration of the margin noise for datasets that arise from two Gaussians. . 106

4.8 Numerical solution to the optimization problem (4.8). 107

4.9 An example that shows the increase in the Rashomon set as we move the right Gaussian away from a mean of 2 in either direction. 107

4.10 Rashomon ratios under margin noise. 108

4.11 Rashomon set characteristics tend to increase with uniform label noise for hypothesis spaces of sparse decision trees. 110

4.12 Rashomon set characteristics tend to increase with uniform label noise for hypothesis spaces of linear models. 111

4.13 The choice of the Rashomon parameter does not influence results in Chapter 4.112

5.1 Feature importance and decision tree visualization of the viral reservoir for patients with HIV. 116

5.2 Predicting HIV reservoir characteristic with machine learning. 118

5.3 A sparse density tree to represent the COCO-stuff labels 121

5.4 Examples of images that contain labels from leaf 1, 2, and 3 in Figure 5.3. . 122

5.5 A sparse density tree to represent the Titanic dataset. 130

5.6 Density rule lists with different preferred list lengths for Titanic dataset. . . 131

5.7 Performance comparison between our methods and baselines for Titanic and Crime datasets. 131

5.8 Leaf-sparse density tree representing the Crime dataset. 133

5.9 List representing the Crime dataset. 134

5.10 Algorithm run time for all datasets as a function of dataset complexity for the sparse density trees and rule lists. 135

A.1 Performance of five machine learning algorithms with regularization for the UCI classification datasets (part I). 146

A.2 Performance of five machine learning algorithms with regularization for the UCI classification datasets (part II). 147

A.3 Performance of five machine learning algorithms without regularization for the UCI classification datasets (part I). 148

A.4 Performance of five machine learning algorithms without regularization for the UCI classification datasets (part II). 149

B.1 Numerical computation that the derivative is negative for Conjuncture 35. . 154

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-

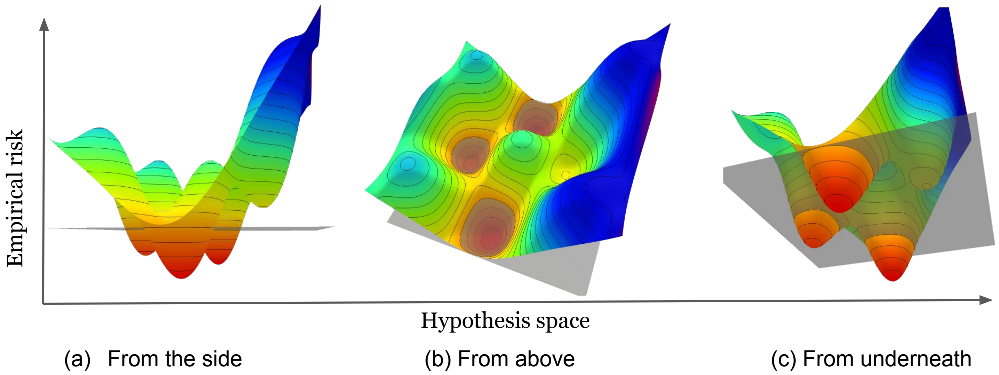


FIGURE 1.1: An illustration of a possible Rashomon set in two dimensional hypothesis space $\mathcal{F}$. Models below the gray plane belong to the Rashomon set $\hat{R}_{set}(\mathcal{F}, \theta)$, where the height of the gray plane is adjusted by the Rashomon parameter θ defined in Section 2.2.

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.

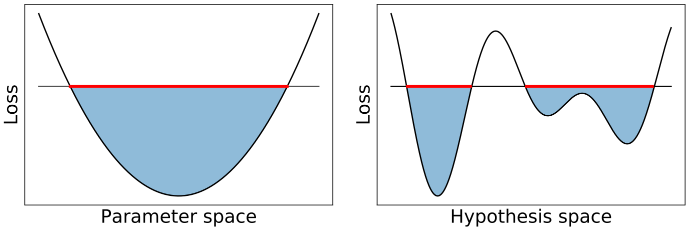
2. Characteristics of the Rashomon Set

In this chapter, we formally define the characteristics of the Rashomon set, including the Rashomon ratio, pattern Rashomon ratio, and pattern diversity. Each of these characteristics has distinct properties, which we explore in detail. Additionally, we introduce various methods and estimations for computing these characteristics, allowing us to measure the Rashomon set. The defined characteristics prove valuable for different hypothesis spaces, as discussed in Chapter 3 and Chapter 4. For instance, in classification and a discrete hypothesis space like decision trees, we use the Rashomon ratio. Conversely, for the hypothesis space of linear models, we primarily rely on the pattern Rashomon ratio. The chapter begins with a discussion related to the Rashomon set's prior work and proceeds to introduce notation and a formal definition of the Rashomon set.

2.1 Related Work

There are several bodies of relevant literature as discussed below.

Rashomon sets. The Rashomon set, named after the Rashomon Effect coined by Leo Breiman (Breiman, 2001), is based on the observation that often there are many equally good explanations of the data. When these are contradictory, the Rashomon Effect gives rise to predictive multiplicity (Black et al., 2022; Hsu & Calmon, 2022; Marx et al., 2020). Rashomon sets have been used to study variable importance (Dong & Rudin, 2020; Fisher et al., 2019; Smith et al., 2020), for characterizing fairness (Aïvodji et al., 2021; Coston et al., 2021; Shahin Shamsabadi et al., 2022), to improve robustness and generalization, especially under distributional shifts (Y. Lee et al., 2023; Ross et al., 2020), to study connections between multiplicity and counterfactual explanations (Brunet et al., 2022; Pawelczyk et al., 2020; Yan & Zhang, 2022), and to help in robust decision making (Tulabandhula & Rudin, 2014). Some works focused on trying to compute the Rashomon set for specific hypothesis spaces, such as sparse decision trees (Xin et al., 2022), generalized additive models (Zhong et al., 2023), and decision lists (Mata et al., 2022). Other works focus on near-optimality to find a diverse set of solutions to mixed integer problems (Ahanor et al., 2023), a set of targeted predictions under a Bayesian model (Kowal, 2022), or estimate the Rashomon volume via approximating model in Reproducing Kernel Hilbert Space (Mason et al., 2022). Black et al. (2022) shows that the predictive multiplicity metric defined as expected pairwise disagreement increases with expected variance over the models in the Rashomon set. On the contrary, in Chapter 5 we focus on probabilistic variance in labels in the presence of noise. 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., 2021). Large Rashomon sets can occur when the machine learning pipeline is underspecified. D'Amour et al. (2022) provides multiple examples of underspecification in computer vision, natural language processing, and healthcare domains. 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 (P. L. Bartlett et al., 2005). Our bounds do not work the same way and aim to prove a different type of result.



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

FIGURE 2.1: Difference between volume ϵ-flatness as defined in Dinh et al. (2017) and the Rashomon set. The red line represents the volume ϵ-flatness in (a), and the Rashomon set in (b). The volume of the Rashomon set is the sum of lengths of red lines in (b). The height of the shaded area represents (a) the parameter ϵ or the 2σ-sharpness, and (b) the Rashomon parameter θ. Volume 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.

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 (Chaudhari et al., 2019; Dinh et al., 2017; Hochreiter & Schmidhuber, 1997; Keskar et al., 2016). Based on a minimum-message-length argument (Wallace & Boulton, 1968), several works claim that flat loss functions lead to better generalization due to a robustness to noise around the minimum (Chaudhari et al., 2019; Hochreiter & Schmidhuber, 1997; Keskar et al., 2016). Following Hochreiter and Schmidhuber (1997), Dinh et al. (2017) define volume ϵ-flatness, which constitutes a special case of our Rashomon sets, as shown in Figure 2.1. In particular, the Rashomon set is defined over the hypothesis (functional) space, while the volume 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 volume 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 volume ϵ-flatness is relevant only for continuous loss functions. Another way of quantifying flatness is σ-sharpness (Dinh et al., 2017; Keskar et al., 2016), 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.

Metrics of the Rashomon set. To our knowledge, we were the first to propose to measure the Rashomon set. We introduced the Rashomon ratio (Semenova et al., 2022) that measures the Rashomon set as a fraction of models or predictions within the hypothesis space. Since then, multiple metrics have been proposed (Black et al., 2022; Hsu & Calmon, 2022; Marx et al., 2020; Watson-Daniels et al., 2023). Ambiguity and discrepancy (Marx et al., 2020; Watson-Daniels et al., 2023) indicate the number of samples that received conflicting estimates from models in the Rashomon set when the Rashomon capacity (Hsu & Calmon, 2022) measures the Rashomon set for probabilistic outputs. We also introduced the pattern Rashomon ratio (Rudin et al., 2022) and the pattern diversity (Semenova, Chen, et al., 2023). Pattern diversity is close to expected pairwise disagreement (as in Black et al. (2022)), however, it uses unique classification patterns (see Section 2.5.1).

2.2 Notation and Rashomon Set Definitions

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 $\mathcal{F} = \{f : \mathcal{X} \to \mathcal{Y}\}$. 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 $\phi(f(x), y)$. If ϕ is a 0-1 loss function, then for a point $z = (x, y)$ and hypothesis f, $\phi(f(x), y) = \mathbb{1}_{[f(x) \neq y]}$. 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 unknown distribution $\mathcal{D}$ and therefore is estimated with an *empirical risk*: $\hat{L}(f) = \frac{1}{n} \sum_{i=1}^{n} \phi(f(x_i), y_i)$. Let $\hat{f}$ be an empirical risk minimizer: $\hat{f} \in \arg\min_{f \in \mathcal{F}} \hat{L}(f)$. If we want to specify the dataset on which $\hat{f}$ was computed, we will indicate it by an index, $\hat{f}_S$.

The *empirical Rashomon set* (or simply *Rashomon set*) is 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, 2001; Coker et al., 2021; Fisher et al., 2019; Srebro et al., 2010; Tulabandhula & Rudin, 2014). More precisely:

Definition 1 (Rashomon set). *For dataset S, a hypothesis space $\mathcal{F}$, and a loss function ϕ, given $\theta \geq 0$, the Rashomon set $\hat{R}_{set}(\mathcal{F}, \theta)$ is:*

$$\hat{R}_{set}(\mathcal{F}, \theta) := \{f \in \mathcal{F} : \hat{L}(f) \leq \hat{L}(\hat{f}) + \theta\}, \tag{2.1}$$

where $\hat{f}$ is an empirical risk minimizer for the training data S with respect to loss function ϕ: $\hat{f} \in \arg\min_{f \in \mathcal{F}} \hat{L}(f)$, and θ is the Rashomon parameter.

If we want to specify the dataset S that is used to compute the Rashomon set, we indicate the dataset in the subscript, as $\hat{R}_{set_S}(\mathcal{F}, \theta)$.

The hypothesis space $\mathcal{F}$ can be a 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 piecewise constant functions).

Rashomon parameter θ determines the risk threshold $(\hat{L}(\hat{f}) + \theta)$, such that all models with risk lower than this threshold are inside the set. For instance, if we stay within 1% of the accuracy of the best model, then $\theta = 0.01$.

We consider *true Rashomon sets* that contain models with low true risk, relative to the optimal true risk value, with parameter $\gamma > 0$:

$$R_{set}(\mathcal{F}, \gamma) = \{f \in \mathcal{F} : L(f) \leq L(f^*) + \gamma\},$$

where $f^* \in \mathcal{F}$ minimizes the true risk.

A large true Rashomon set, as it turns out, can be a certificate of the existence of a simpler model. However, since we can never actually explore the true Rashomon set, we would never know whether it will be (or has been) useful for a particular problem. We explain this in Section 3.1, and then consider *empirical* Rashomon sets, which are easier to work with in practice.

Given a parameter $\eta \geq 0$, we call the Rashomon set with restricted empirical risk an *anchored Rashomon set*:

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

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

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

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 the properties of the other. We consider anchored Rashomon sets to show that with high probability, if a fixed model is contained within the true anchored Rashomon set, it also belongs to a slightly larger (empirical) anchored Rashomon set. The reverse statement holds as well.

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

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

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

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

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

When the hypothesis space has a parameterized representation (denote $\mathcal{F}_\Omega$; sometimes we will omit subscript Ω for convenience), we assume that we can parameterize each model $f \in \mathcal{F}_\Omega$ with a parameter vector $\omega \in \Omega$ of finite length and denote $f(z) = f_\omega(z)$.

In the next section, we define and discuss properties of the Rashomon ratio as a complexity measure.

2.3 Rashomon Ratio

Since hypothesis spaces can vary from one problem to another, we will often normalize the size of the Rashomom set via the *Rashomon ratio* $\hat{R}_{ratio}(\hat{R}_{set}(\mathcal{F}, \theta))$ which takes the Rashomon set as input and outputs a value between 0 and 1.

When considering the Rashomon ratio, we assume that the hypothesis space is bounded and that there is a prior distribution ρ over functions in $\mathcal{F}$. Given a prior, ρ, on the hypothesis

Table 2.1: Comparison of Rashomon ratio and other complexity measures. The Rashomon ratio considers the fact that there are multiple good models and is a property both of the hypothesis space and data.

Complexity measure	Property of	Depends on data	Considers set of good models
VC Dimension	hypothesis space	no	no
Algorithmic stability (Hypothesis stability (Bousquet & Elisseeff, 2002))	algorithm, hypothesis space	no	no
Empirical algorithmic stability (Algorithmic hypothesis stability (Bousquet & Elisseeff, 2002))	algorithm, hypothesis space	yes	no
Geometric margins	one function	yes	no
Empirical Local Rademacher Complexity (P. L. Bartlett et al., 2005)	hypothesis space	depends on features, not on labels	no
Rashomon ratio	hypothesis space	yes, but not always on labels (see Theorem 8)	yes

space, the Rashomon ratio measures the fraction of the hypothesis space contained in the Rashomon set. Unless explicitly specified, ρ is assumed to be uniform. For simplicity, we will denote the Rashomon ratio as $\hat{R}_{ratio}(\mathcal{F}, \theta)$. In general, the Rashomon ratio is $\hat{R}_{ratio}(\mathcal{F}, \theta) = \int_{f \in \mathcal{F}} 1_{f \in \hat{R}_{set}(\mathcal{F}, \theta)} \rho(f) d\rho$. If the hypothesis space has a uniform prior, then the Rashomon ratio is the volume of the Rashomon set divided by the volume of the hypothesis space $\hat{R}_{ratio}(\mathcal{F}, \theta) = \frac{\mathcal{V}(\hat{R}_{set}(\mathcal{F}, \theta))}{\mathcal{V}(\mathcal{F})}$, where $\mathcal{V}(\cdot) : \mathcal{F} \to \mathbb{R}_+$ is the volume function. If the hypothesis space is discrete with a uniform prior, the Rashomon ratio can be computed as $\hat{R}_{ratio}(\mathcal{F}, \theta) = \frac{|\hat{R}_{set}(\mathcal{F}, \theta)|}{|\mathcal{F}|}$, where $|A| = \sum_{f \in \mathcal{F}} \mathbb{1}_{f \in A}$.

The Rashomon ratio represents the fraction of good models (the fraction of models that fit the data about equally well). A larger Rashomon ratio implies that more models perform about equally well. The dataset S is denoted in the subscript, as $\hat{R}_{ratio_S}(\mathcal{F}, \theta)$. $R_{ratio}(\mathcal{F}, \theta)$ denotes the Rashomon ratio for the true Rashomon set.

2.3.1 Rashomon Ratio as a Simplicity Measure

The Rashomon ratio, as a *property of a dataset and a hypothesis space*, serves as a gauge of simplicity of the learning problem. If the Rashomon set is large, many different reasonable optimization procedures could lead to a model from the Rashomon set. Therefore, for large Rashomon sets, accurate models tend to be easier to find (since 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.

The Rashomon ratio can give insight into the simplicity of a learning problem, though it was designed for a fundamentally different goal than well-known complexity measures from learning theory (see Table 2.1). While those complexity measures were designed to help us understand generalization, the Rashomon ratio (with additional assumptions) helps us understand whether simpler functions might exist with the same level of accuracy as complex functions. The Rashomon ratio depends on a loss function, the hypothesis space, and a dataset, while the majority of other measures are either data-agnostic or focus on the properties of a specific model in the space. We will use demonstrations to show the differences between the Rashomon ratio and other complexity measures.

2.3.1.1 The Rashomon ratio is different from VC dimension

The VC dimension (Vapnik & Chervonenkis, 1971) shows the expressive power of a hypothesis space for *any* dataset 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 dataset, which may not be extreme.

2.3.1.2 The Rashomon ratio is different from algorithmic stability

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

Definition 3 (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 dataset S, loss $l(f_S, z) = \phi(f_S(x), y)$ for $z = (x, y)$, dataset $S = \{z_1, ...z_n\}$, and $S^{\setminus i}$ is modified from the training data by removing the i^{th} element of the dataset: $S^{\setminus i} = \{z_1, ..., z_{i-1}, z_{i+1}, ...z_n\}$.

Algorithmic stability (see Definition 3) depends on a change to a dataset, whereas Rashomon Ratio uses a fixed dataset. As we showed in Theorem 8 in Section 2.3.2, in the case of linear least squares regression, the Rashomon ratio depends on features 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. This is formalized in Theorem 4.

Theorem 4 (Rashomon ratio is not algorithmic stability)**.** *Consider a distribution P_X over a discrete domain $\mathcal{X} = \{x_1, ...x_N\}$ and a learning algorithm A that minimizes the sum of squares loss : $\|X\omega - Y\|_2^2$. for a linear hypothesis space $\mathcal{F}_\Omega$. 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, $\mathbf{Y}_1$ drawn from $P_{Y_1|\mathbf{X}}$ over $\mathcal{Y} \,|\, \mathcal{X}$, and $\mathbf{Y}_2$ drawn from $P_{Y_2|\mathbf{X}}$ over $\mathcal{Y} \,|\, \mathcal{X}$, the expected Rashomon ratios are the same:*

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

yet hypothesis stability constants are different by our arbitrarily chosen value of λ: $\tilde{\beta}_2 - \tilde{\beta}_1 \geq \lambda$, where $\mathbf{S}_1$ and $\mathbf{S}_2$ denote datasets $\mathbf{S}_1 = [\mathbf{X}, \mathbf{Y_1}]$ 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. Let us create our distribution. 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. That is, the two conditional distributions have $y = 0$ except when $\mathbf{x} = \mathbf{x}_0$ for Y_2, when it is H with probability $1/2$.

As a first part of the proof, we show that the algorithmic stability constants are different. 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}:

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

where $\mathbf{S}_2^s, \mathbf{z}^s$ is a special draw such that $\mathbf{z}^s = (\mathbf{x_0}, \mathbf{H})$, and where $\mathbf{S}_2^s$ includes one point at $(\mathbf{x_0}, \mathbf{H})$, one point at $(\mathbf{x_0}, \mathbf{0})$, and the rest at other values $(\mathbf{x}, \mathbf{0})$. Since the domain $\mathcal{X}$ is discrete, the probabilities of a special draw are:

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

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

where $Bin(k, n, p_k) = \binom{n}{k} p_k^k (1 - p_k)^{(n-k)}$ is a binomial coefficient, namely the 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 $\mathbf{z_1} = (\mathbf{x_0}, \mathbf{H})$ and $\mathbf{z_2} = (\mathbf{x_0}, \mathbf{0})$, the loss difference $|l(f_{\mathbf{S}_2^s}, \mathbf{z}^s) - l(f_{\mathbf{S}_2^{s,\setminus i}}, \mathbf{z}^s)|$ evaluated at $\mathbf{z}^s$ for all i will be at least $\frac{\mathbf{H}^2}{4}$. To see this, note that the optimal function's value at $\mathbf{x_0}$ is: $f_{\mathbf{S}_2^s}(\mathbf{x_0}) = \frac{\mathbf{H}}{2}$, the optimal function's value at $\mathbf{x_0}$ after we remove the first point is $f_{\mathbf{S}_2^{s,\setminus 1}}(\mathbf{x_0}) = \mathbf{0}$, and the optimal function's value at $\mathbf{x_0}$ after removing the second point is $f_{\mathbf{S}_2^{s,\setminus 2}}(\mathbf{x_0}) = \mathbf{H}$. Therefore, $l(f_{\mathbf{S}_2^s}, \mathbf{z}^s) = \frac{\mathbf{H}^2}{4}$, $l(f_{\mathbf{S}_2^{s,\setminus 1}}, \mathbf{z}^s) = \mathbf{H}^2$, $l(f_{\mathbf{S}_2^{s,\setminus 2}}, \mathbf{z}^s) = \mathbf{0}$. And we get that $|l(f_{\mathbf{S}_2^s}, \mathbf{z}^s) - l(f_{\mathbf{S}_2^{s,\setminus 1}}, \mathbf{z}^s)| = \frac{3\mathbf{H}^2}{4}$, $|l(f_{\mathbf{S}_2^s}, \mathbf{z}^s) - l(f_{\mathbf{S}_2^{s,\setminus 2}}, \mathbf{z}^s)| = \frac{\mathbf{H}^2}{4}$. As we add the rest of the points $(\mathbf{x}_i, \mathbf{0})$ to the dataset $\mathbf{S}_2^s$, the loss difference (from changing $f_{\mathbf{S}_2^s}(\mathbf{z}^s)$ to $f_{\mathbf{S}_2^{s,\setminus i}}(\mathbf{z}^s)$) in the special draw case will only increase. Therefore for all i:

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

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 $\left|\tilde{\beta_1} - \tilde{\beta_2}\right| > \lambda$. This proves that the hypothesis stability constants are different and completes the first part of the proof.

We now need to prove that the expected Rashomon ratios are the same, which will constitute the second part of the proof. The Rashomon ratio for the hypothesis space $\mathcal{F}_\Omega$ of linear models does not depend on targets and can be calculated as in (2.3) for both $\mathbf{S}_1$ and $\mathbf{S}_2$. Therefore the expected Rashomon ratios are the same:

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

Thus, both halves of our proof are complete. ■

2.3.1.3 The Rashomon ratio is different from geometric margins

For the parametric hypothesis space of linear models $\mathcal{F}_\Omega = \{f : f(x) = \omega^T x, \omega \in \mathbb{R}^p\}$ 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 $\frac{2}{\|\hat{\omega}\|_2}$.

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 dataset. However, the margin (i.e., the minimum margin of the maximum margin classifier) (Schapire et al., 1998) depends on points closest to the decision boundary (support vectors), while the Rashomon set does not necessarily rely on the support vectors and may depend on the full dataset. Theorem 5 summarizes this idea.

Theorem 5 (Rashomon ratio is not the 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 datasets* S_1 *and* S_2 *with the same empirical risk minimizer* $\hat{f} \in \mathcal{F}_\Omega$ *such that the geometric margin* d *is the same for both datasets, yet the Rashomon ratios are different:*

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

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.

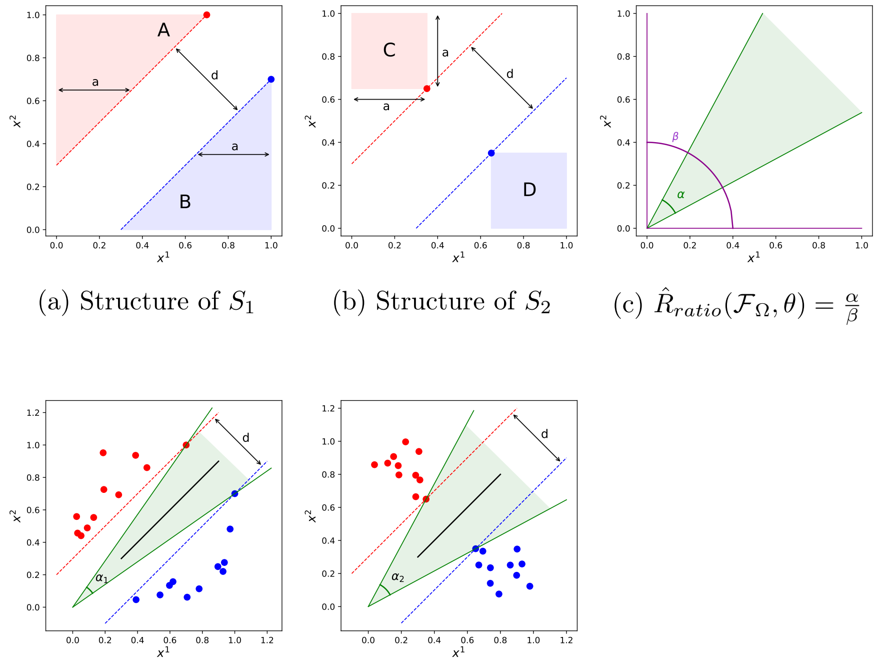
For some positive constant $a \in (0,1)$ that we choose later, consider the following regions of the feature space:

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

$$B = \{x^1 \in (a, 1], x^2 < x^1 - (1 - 2a)\},$$

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

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



(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 2.2: An illustration of different Rashomon ratios with identical geometric margins. (a) and (b) show the datasets S_1 and S_2 with identical margin d. The black line in (d) and (e) shows the optimal model, and the shaded region in (c), (d), and (e) indicates the Rashomon set $\hat{R}_{set}(\mathcal{F}_\Omega, 0)$ with its boundaries represented by green lines. The hypothesis space consists of all origin-centered linear models that intersect the zero-one hypercube, where data reside. (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 datasets S_1 and S_2 are represented by different angles α_1 and α_2 and therefore have different Rashomon ratios. The figure is best seen in color.

Construct dataset 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 dataset S_1 such that $x_{S_1}^{s_1} = [1-2a, 1]$ and $x_{S_1}^{s_2} = [1, 1-2a]$. Please see Figure 2.2a for details.

Construct dataset 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 dataset S_2 such that $x_{S_2}^{s_1} = [a, 1-a]$ and $x_{S_2}^{s_2} = [1-a, a]$. Please see Figure 2.2b for details.

Note that the datasets we considered have the same width for the geometrical margin $d = \sqrt{2}(2a-1)$ (see Figures 2.2a, 2.2b). Now, we are left to show that the Rashomon ratios are different.

For the hypothesis 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 size of the Rashomon set. 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 2.2c. 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 datasets S_1 and S_2 Figures 2.2d and 2.2e show the Rashomon set and angles α_1 and α_2 that represent the volume of the Rashomon set. Given the special points in the datasets 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 difference between the Rashomon ratios is:

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

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 $|\hat{R}_{ratio_{S_1}}(\mathcal{F}, 0) - \hat{R}_{ratio_{S_2}}(\mathcal{F}, 0)|$ is at least λ.

■

2.3.1.4 The Rashomon ratio is different from empirical local Rademacher complexity

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

Definition 6 (Empirical Rademacher complexity)**.** *Given a dataset 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 $i = 1, 2, \ldots, n$.

Based on Definition 6, empirical Rademacher complexity measures how well the hypothesis space can fit random assignments of the labels. The Rashomon ratio uses fixed labels. 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.

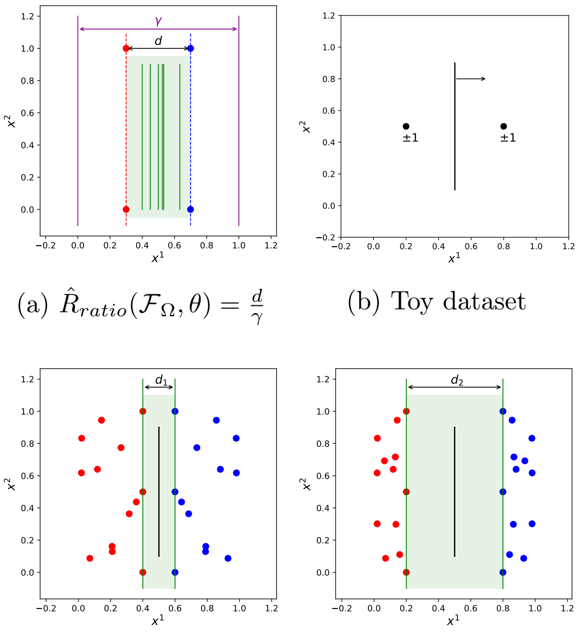
Since we are interested only in models that are inside the Rashomon set, we will consider local empirical Rademacher complexity (P. L. Bartlett et al., 2005), which is defined using the Rashomon set $\hat{R}_{set}(\mathcal{F}, \theta)$. In the following theorem, we provide a simple example to show the discrepancy between the two measures.

Theorem 7 (Rashomon ratio is not local Rademacher complexity)**.** *For $0 < \lambda < 1$, there exist two datasets 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:

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



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

(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 2.3: An illustration of different Rashomon ratios with equivalent empirical local Rademacher complexities. The 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 volume of the Rashomon set. (b) A toy dataset that illustrates that the empirical local Rademacher complexity is zero for models in the Rashomon set. (c) dataset S_1, and (d) dataset S_2 illustrate symmetric separable datasets with different Rashomon ratios. Best seen in color.

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 volume of the Rashomon set in parameter space by computing the difference between the largest and smallest threshold values of models within the Rashomon set, as illustrated in Figure 2.3a. 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 volume of the Rashomon set 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 datasets S_1, S_2 with different projected minimal distances d_1 and d_2, such that $|d_1 - d_2| > \lambda$. (Please see Figure 2.3c and 2.3d for details of the datasets S_1 and S_2.) Consequently, we get that:

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

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 the toy dataset in Figure 2.3b: $\hat{R}_n^S\left(\hat{R}_{set}(\mathcal{F}_\Omega, 0)\right) = \frac{1}{2}\frac{1}{4}\left[\left(\hat{f}(x_1) + \hat{f}(x_2)\right) + \left(\hat{f}(x_1) - \hat{f}(x_2)\right) + \left(-\hat{f}(x_1) + \hat{f}(x_2)\right) + \left(-\hat{f}(x_1) - \hat{f}(x_2)\right)\right] = 0.$

Since both S_1 and S_2 are separable and symmetric we get that:

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

■

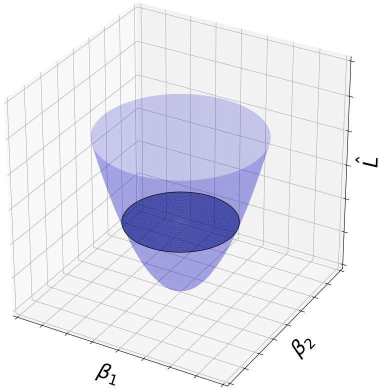


FIGURE 2.4: The Rashomon set for the two-dimensional least squares regression. The volume of a shaded ellipsoid corresponds to the volume of the Rashomon set in a parameter space.

2.3.2 Analytical Calculation of Rashomon Ratio for Ridge Regression

A special case of when the Rashomon ratio can be computed in closed-form in 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 dataset $S = [X, Y]$:

$$\min_\omega \hat{L}(\omega) = \min_\omega (X\omega - Y)^T(X\omega - Y) + C\omega^T\omega, \tag{2.2}$$

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 volume of the Rashomon set (Figure 2.4). For a hypothesis space with uniform prior and volume function $\mathcal{V}$, the Rashomon ratio is $\frac{\mathcal{V}(\hat{R}_{set}(\mathcal{F}_\Omega,\theta))}{\mathcal{V}(\mathcal{F}_\Omega)}$. Using the geometric intuition above, we compute the Rashomon ratio in the parameter space by the following theorem:

Theorem 8 (Rashomon ratio for ridge regression)**.** *For a parametric hypothesis space of linear models $\mathcal{F}_\Omega = \{f_\omega(x) = \omega^T x, \omega \in \mathbb{R}^p\}$ with uniform prior and a dataset $S = X \times Y$, 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 ratio can be computed as:

$$\hat{R}_{ratio}(\mathcal{F}_\Omega, \theta) = \frac{J(\theta, p)}{\mathcal{V}(\mathcal{F}_\Omega)} \prod_{i=1}^{p} \frac{1}{\sqrt{\sigma_i^2 + C}}, \tag{2.3}$$

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:

$$\hat{L}(X, Y, \omega) \leq \hat{L}(X, Y, \hat{\omega}) + \theta. \tag{2.4}$$

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

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

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 volume of the Rashomon set 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.

Since we assume a uniform prior on $\mathcal{F}_\Omega$, $\mathcal{V}(F_\Omega)$ is the volume of a box (or other closed region) containing the plausible values of Ω. Therefore, the Rashomon ratio is $\hat{R}_{ratio}(\mathcal{F}_\Omega, \theta) = \frac{\mathcal{V}(\hat{R}_{set}(\mathcal{F}_\Omega,\theta))}{\mathcal{V}(\mathcal{F}_\Omega)} = \frac{J(\theta,p)}{\mathcal{V}(\mathcal{F}_\Omega)} \prod_{i=1}^{p} \frac{1}{\sqrt{\sigma_i^2+C}}$, where $J(\theta, p) = \frac{\pi^{p/2}\theta^{p/2}}{\Gamma(p/2+1)}$. ■

Interestingly, from Theorem 8, it follows that for ridge regression, *the Rashomon ratio 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 ratio generally depends on the targets; in that sense, ridge regression is unusual.

2.3.3 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 the 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 can 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: $P(|\hat{P}[f \in \hat{R}_{set}(\mathcal{F},\theta)] - P[f \in \hat{R}_{set}(\mathcal{F},\theta)]| \geq t) \leq 2e^{-2kt^2}$, or alternatively $1-\alpha = P(|\hat{P}[f \in \hat{R}_{set}(\mathcal{F},\theta)] - P[f \in \hat{R}_{set}(\mathcal{F},\theta)]| < t) \leq 1 - 2e^{-2kt^2}$. 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 p_p be a target distribution over the hypothesis space $\mathcal{F}$ and p_t be a proposal distribution that is focused around the Rashomon set. We can estimate the Rashomon ratio through importance sampling as follows:

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

For the binary classification dataset and the hypothesis space of fully grown decision trees of a given depth, the target and proposal distributions are as follows. 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 passes 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.

We use the importance sampling to compute Rashomon ratios of the hypothesis space of fully grown decision trees of depth seven in Section 3.2.

2.3.4 Rashomon Ratio for the Hypothesis Space of Sparse Decision Trees

Consider the hypothesis space of sparse decision trees, where the objective is to minimize the misclassification error and a sparsity penalty on the number of leaves. For binary classification datasets with binary features, TreeFARMS (Xin et al., 2022) allows us to enumerate all sparse trees in the Rashomon set within the defined Rashomon parameter. Therefore, we can use TreeFARMS to compute the numerator of the Rashomon ratio.

To compute the denominator of the Rashomon ratio, consider all possible sparse decision trees up to a given depth d. Given the p features and the maximum depth d, let $C(d, p)$ be the number of sparse trees in the hypothesis space. Then we have the following recursive formula to compute the size of hypothesis space with p features:

$$C(d, p) = 2 + pC(d-1, p-1)^2, \tag{2.5}$$

where $C(0, \cdot) = 2$. In the base case, when we have only one leaf, there are only two possible trees: one that classifies every point as 0, or one that classifies every point as 1. Therefore, $C(0, \cdot) = 2$. Then for decision trees up to depth d with $p \geq d$ features, there are two cases. The first case is when the tree has depth 0 which produces 2 possible trees discussed above. The other case is when the tree has a depth of at least 1. In this case, there are p possible features to initially split on, and then the left and right subtrees are of depth at most $d-1$ with $p-1$ features to choose from. The left and right subtrees can be chosen independently of each other, so we have $pC(d-1,p-1)^2$ trees, which proves the overall recursive formula.

Note that for a binary dataset and decision trees of depth exactly equal to d for every tree path, the following recursive formula holds $C(d,p) = pC(d-1,p-1)^2$ which is equivalent to the closed-form solution described in the proof of Proposition 31 in Section 4.1.4.

The case we considered in this section might restrict the structure or the hypothesis space (such as enforcing sparsity), but these restrictions allow us to compute the size of the Rashomon set directly. We will use the exact computation of the Rashomon ratio for the experiments in Chapter 4.

2.4 Pattern Rashomon Ratio

The size of the Rashomon set and Rashomon ratio (when measured in parameter space) may have problems with overparameterization or reparameterization; i.e., changing parameterization of functions may inflate or deflate the values of the Rashomon ratio (Dinh et al., 2017). To mitigate problems with parameterization, we introduce a quantity that groups functions into equivalence classes, based on their predictions for each data point. Specifically, for a binary classification on a given dataset S we introduce the pattern Rashomon set and pattern Rashomon ratio.

2.4.1 Definition and Properties

Given a hypothesis f, a dataset S, and a 0-1 loss ϕ, a predictive pattern (or pattern) *pt* is the collection of outcomes from applying f to each sample from S: $pt^f = [f(x_1), ..., f(x_i), .., f(x_n)]$. We say that pattern *pt* is achievable on the Rashomon set if there exists $f \in \hat{R}_{set}(\mathcal{F}, \theta)$ such that $pt^f = pt$. Let *pattern Rashomon set* $\pi(\mathcal{F}, \theta) = \{pt^f : f \in \hat{R}_{set}(\mathcal{F}, \theta), pt^f = [f(x_i)]_{i=1}^n\}$ be all unique patterns achievable by functions from $\hat{R}_{set}(\mathcal{F}, \theta)$ on dataset S. Finally, let $\Psi(\mathcal{F})$ be the *pattern hypothesis set*, meaning that it contains all patterns achievable by models in the hypothesis space, $\Psi(\mathcal{F}) = \{pt^f : f \in \mathcal{F}, pt^f = [f(x_i)]_{i=1}^n\}$. The pattern Rashomon ratio is the ratio of patterns in the pattern Rashomon set to the

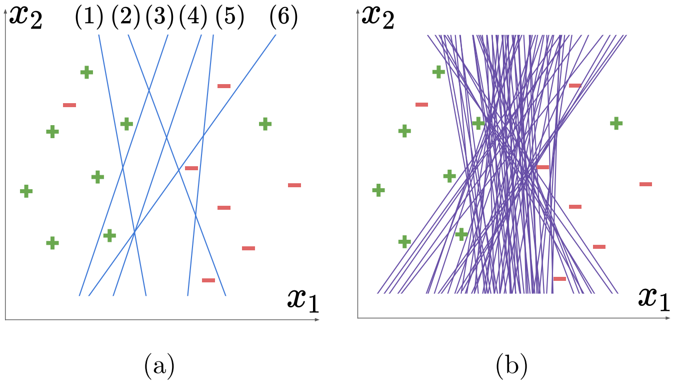


FIGURE 2.5: (a) The pattern Rashomon set. Each classifier in the figure defines a different loss pattern. The number of distinct patterns (in Figure (a) there are six patterns) created by functions in the Rashomon set comprises the numerator of the pattern Rashomon ratio. (b) The Rashomon set. Each classifier in the figure is a different model (unique hypothesis, but *not* unique loss pattern) in the Rashomon set. The fraction of hypotheses in the hypothesis space that are in the Rashomon set is the Rashomon ratio.

pattern hypothesis set: $\hat{R}^{pat}_{ratio}(\mathcal{F}, \theta) = \frac{|\pi(\mathcal{F},\theta)|}{|\Psi(\mathcal{F})|}$.

The ratio based on patterns is different from the Rashomon ratio defined in Section 2.3 as a multiplicity of models, as shown in Figure 2.5. 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 prediction properties.

The pattern Rashomon ratio not only mitigates problems with overparameterization of functions (which occurs in parameter space), it also helps in measuring the Rashomon ratio in function spaces when the hypothesis space contains many models that do not intersect with the hypercube where the data reside. Here, functions that do not intersect with the data will be grouped into two equivalence classes (one for all positive predictions, and one for all negative predictions) and will not artificially enlarge the denominator of the Rashomon ratio.

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 dataset only. This intuition is summarized in the next proposition.

Proposition 9 (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 dataset, 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 & Sloane, 1977). ■

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.

In fact, for any dataset, the maximum number of patterns in the pattern Rashomon set is bounded by the empirical risk of the empirical risk minimizer and the Rashomon parameter θ, as we show next in Proposition 10.

Proposition 10. *Given the dataset S of size n, the pattern Rashomon set $\pi(\mathcal{F}, \theta)$, the empirical risk of the empirical risk minimizer $\hat{L}(\hat{f})$, and the Rashomon parameter θ, the cardinality of the pattern Rashomon set obeys:*

$$|\pi(\mathcal{F}, \theta)| \leq \sum_{k=1}^{\lceil n\hat{L}(\hat{f})+n\theta\rceil} \binom{n}{k}.$$

Proof. For every model from the Rashomon set f, $\hat{L}(f) \leq \hat{L}(\hat{f}) + \theta$, which means that, in the worst case, the Hamming distance between pattern p^f and vector of true labels $Y = [y_i]_{i=1}^n$ is $\lceil n\hat{L}(\hat{f}) + n\theta\rceil$. Thus, patterns in the pattern Rashomon set can make one mistake, two mistakes, and so on up to $\lceil n\hat{L}(\hat{f}) + n\theta\rceil$ mistakes, which means there are at most $\sum_{k=1}^{\lceil n\hat{L}(\hat{f})+n\theta\rceil} \binom{n}{k}$ patterns in the pattern Rashomon set. ■

The pattern Rashomon ratio 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).

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 dataset. 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 dataset and do not take supremum over all datasets (as is usual for the growth function).

We can use rejection or importance sampling methods to compute the pattern Rashomon ratio as well. In this case, during a random draw, one can sample a pattern, check if the hypothesis space supports it (e.g. if there exists a model in the hypothesis space that realizes this pattern), and finally compute whether this model belongs to the Rashomon set. However, we can also shift the complexity of computing the pattern Rashomon set from sampling from the hypothesis space to enumerating all possible patterns as discussed next.

2.4.2 Branch and Bound Method to Compute the Pattern Rashomon Set

In this section, we describe a two-step method that allows us to compute all patterns in the pattern Rashomon set for the hypothesis space of linear models $\mathcal{F} = \{f = \omega^T x\}$ (although the method can be applied to other hypothesis spaces as well). In the first step, we reduce the complexity of the problem, by discarding points that have been classified similarly by models in the pattern Rashomon set. In the second step, we use a branch-and-bound approach in order to enumerate patterns and discard prefixes of those patterns that will not be in the Rashomon set based on the Rashomon parameter and the empirical risk of the empirical risk minimizer.

Given a sample z_i, let $a_i = \frac{1}{\Pi} \sum_{k=1}^{\Pi} \mathbb{1}_{[pt_k^i = y_i]}$, where pt_k^i is i^{th} index of the k^{th} pattern, denote the probability with which patterns from the pattern Rashomon set classify z_i correctly. We will call a_i *sample agreement* over the pattern Rashomon set.

Consider a dataset $S = \{z_i\}_{i=1}^n$ and a 0-1 loss ϕ. For every point z_i assume that we have sample agreement a_i. If $a_i = 0$, it means that all patterns in the pattern Rashomon set assign an incorrect label to sample z_i. On the other hand, if $a_i = 1$, all patterns assign the correct label. If we exclude all z_i such that $a_i = 0$ or $a_i = 1$ from the dataset, then the number of patterns will not change in the Rashomon set. Therefore, for a given point z_k ($k = 1..n$) we will try to answer a question: is there a model in the Rashomon set such that it classifies $\bar{z}_k = (x_k, -y_k)$ correctly and still stays in the Rashomon set. If there is no such model, then sample z_k has no influence on the pattern Rashomon set. Since it is harder to optimize for 0-1 loss, we instead consider exponential loss. If the problem is separable by 0-1 loss, then the exponential loss will converge to a separable solution exponentially fast (which is known from the convergence analysis of AdaBoost (P. Bartlett et al., 1998)). Then given hypothesis space of linear models $\mathcal{F} = \{w^T x\}$, for every z_k, we aim to solve following optimization problem:

$$\min \frac{1}{n} \sum_{i=0}^{n} e^{-y_i w^T x_i} \tag{2.6}$$

$$y_k w^T x_k \leq 0, \tag{2.7}$$

and then check if $w^T x$ is in the Rashomon set defined by 0-1 loss.

Since we optimize exponential loss, it is fast to solve the optimization problem with gradient descent. More importantly, we can run the optimization in parallel for samples z_k. After, we consider dataset S_{inside} that contains only those samples for which models were in the Rashomon set that could accommodate misclassified z_k. We formally define the discard point procedure in procedure DISCART POINTS below:

```
procedure DISCARD POINTS(dataset $S$, ERM $\hat{f}$, the Rashomon parameter $\theta$)
    Initialize $S_{dp}$.
    for every $z = (x, y) \in S$ do
        Solve optimization problem (2.6)-(2.7). Let $\bar{f}$ be a solution.
        if $\hat{L}(\bar{f}) > \hat{L}(\hat{f}) + \theta$ then
            add $z$ to $S_{dp}$   (this point has a single predicted label for the entire $\hat{R}_{set}(\mathcal{F}, \theta)$).
        end if
    end for
    return $S_{dp}$.
end procedure
```

In the second step, we build a search tree over the set of patterns that are formed by samples in S_{inside}. We use breadth-first search over subsets of data. We “bound” (i.e., exclude part of the search space) when the prefix of the pattern (which is the part of the dataset we are working with) misclassifies more samples than the threshold to stay in the Rashomon set, which is $\hat{L}(\hat{f}) + \theta$. Since not all patterns can be realized by the model class. We “bound” if the prefix or pattern can not be achieved (the pattern is achievable when

Algorithm 1 Branch and bound approach to find the pattern Rashomon set

```
**Input:** The Rashomon parameter $\theta$, dataset $S = X \times Y$, ERM $\hat{f}$, algorithm $A$.
**Output:** Pattern Rashomon set $\pi(\mathcal{F}, \theta)$.
Run DISCARD POINTS($S, \hat{f}, \theta$) to exclude points that have the same predicted label for all models in the Rashomon set. Let $S_{dp}$ be the set of discarded points, and $S_{\text{inside}}$ be the rest of the points.
Divide points in $S_{\text{inside}}$ into four categories: true positive, false positive, true negative, and false negative.
Compute the distance from the decision boundary of $\hat{f}$ to every point for every category.
Sort points in ascending order for every category.
Create a new order of the points in $S_{\text{inside}}$ by iteratively choosing points from each of the four categories until all points in $S_{\text{inside}}$ are re-ordered.
Concatenate $S_{dp}$ and $S_{\text{inside}}$ to form $S = X \times Y$ based on the new order, where discarded points are followed by the sorted points.
Initialize the prefix $pt_{init}$ of length $|S_{dp}|$ based on the labels of the samples in $S_{dp}$.
Initialize $Q$ as the queue for the breadth-first search over the prefixes.
while $i \leq |S_{\text{inside}}|$ (loop over all points in $S_{\text{inside}}$) do
    for every $elem$ in $Q$ (loop over all prefixes in $Q$) do
        for $e \in [0, 1]$ (loop over possible labels; this is a "branch" step) do
            $Y_a = Y_{S_{dp}} \cup elem \cup e$ (consider potential prefix).
            Form the training data $(X_a, Y_a)$ to check if the prefix is achievable by algorithm $A$. $X_a$ consists of the first $|S_{dp}| + i$ samples of sorted $X$.
            Fit algorithm $A$ on $(X_a, Y_a)$ and compute $accuracy$ and $loss$.
            if $accuracy = 1$ and $loss \leq \hat{L}(\hat{f}) + \theta$ then
                $Q.append(elem \cup e)$ (the prefix is achievable and the pattern has the potential to be achieved in the pattern Rashomon set, thus we add this element to the queue. This is a "bound" step).
            end if
        end for
    end for
end while
As we have now looped over all samples, $Q$ contains all the achievable patterns that are in the Rashomon set, set $\pi(\mathcal{F}, \theta) = Q$.
```

all points with labels matching the pattern are classified correctly by some model from the hypothesis space). In order to perform branch and bound more effectively, given an empirical risk minimizer (ERM), we sort points in the dataset based on their distance to the decision boundary of the ERM. More specifically, we split the points into four categories depending on whether the point is a true positive, false positive, true negative, or false negative. Then for every category, we compute the distances from each point to the decision boundary of the ERM and then sort points from least distance to greatest distance. Finally, we cyclically choose one point from each category until all samples have been considered. Conceptually, true positive and true negative samples that are closest to the decision boundary determine most of the patterns in the pattern Rashomon set. We add false positives and false negatives early to the order of points as they are more likely to be misclassified, allowing us to bound the prefixes sooner. We describe the branch and bound procedure in Algorithm 1. We use bit vectors to represent prefixes and patterns to speed up computations. Since we apply this approach to linear models, we use logistic regression without regularization to check the achievability of the patterns and their prefixes. However, the algorithm in general can be applied to other hypothesis spaces and losses (for example, hinge loss).

The branch and bound approach allows us to compute the numerator of the pattern Rashomon ratio. To compute the denominator of the Rashomon ratio, we use the following formula that gives the number of all possible patterns for the hypothesis space of linear models (Cover, 1965): if no $p-1$ points are coplanar,

$$C(n,p) = 2\sum_{i=0}^{p}\binom{n-1}{i}, \tag{2.8}$$

where $C(n,p)$ is the number of patterns in the hypothesis space of linear models for the dataset with n points and p features. We use this method to compute pattern Rashomon ratios in Chapter 5.

2.5 Pattern Diversity

We define pattern diversity as an empirical measure of differences in patterns in the Rashomon set. It computes the average distance between the patterns, which allows us not only to assess how large the Rashomon set is but also how diverse it is.

Recall that $\pi(\mathcal{F},\theta)$ is the set of unique classification patterns produced by the Rashomon set of $\mathcal{F}$ with the Rashomon parameter θ.

Definition 11 (Pattern diversity)**.** *For Rashomon set $\hat{R}_{set}(\mathcal{F},\theta)$, the pattern diversity*

$div(\hat{R}_{set}(\mathcal{F},\theta))$ is defined as:

$$div(\hat{R}_{set}(\mathcal{F},\theta)) = \frac{1}{n\,\Pi\,\Pi} \sum_{\substack{j\leq\Pi \\ pt_j\sim\pi(\mathcal{F},\theta)}} \sum_{\substack{k\leq\Pi \\ pt_k\sim\pi(\mathcal{F},\theta)}} H(pt_j, pt_k),$$

where $n = |S|$, $\Pi = |\pi(\mathcal{F},\theta)|$, and $H(pt_j, pt_k) = \sum_{i=1}^{n} \mathbb{1}_{[pt_j^i \neq pt_k^i]}$ is the Hamming distance (in our case it computes the number of samples at which predictions are different), and $|\cdot|$ denotes cardinality.

Pattern diversity measures pairwise differences between patterns of functions in the pattern Rashomon set. Pattern diversity is in the range $[0, 1)$, where it is 0 if the pattern set contains one pattern or no patterns.

2.5.1 Pattern Diversity and Other Metrics of the Rashomon Set

Among different measures of the Rashomon set, the pattern diversity is the closest to the pattern Rashomon ratio and expected pairwise disagreement (as in Black et al. (2022)).

Pattern Rashomon ratio. Pattern Rashomon ratio measures how expressive the Rashomon set is compared to the whole hypothesis space. As we discussed before, for the hypothesis space of linear models, for different datasets with the same number of samples and attributes, as long as no $p-1$ points are collinear, the denominator of the pattern Rashomon ratio is the same and equal to $2\sum_{i=0}^{p}\binom{n-1}{i}$ (Cover, 1965). If we focus only on the numerator of the pattern Rashomon ratio, it is the number of distinct predictions, whereas the pattern diversity is the average Hamming distance between distinct predictions. Intuitively, the more distinct predictions we have, the more different they are from each other, and the higher pattern diversity we should expect. However, this is not always the case, and there exist datasets such that we can have a large number of patterns with very small Hamming distance and a small number of patterns with larger Hamming distance.

In the next section, we will provide an upper bound on the pattern diversity that depends on the empirical risk of the empirical risk minimizer and the Rashomon parameter θ (see Theorem 16). Similarly to pattern diversity, we can upper-bound the number of patterns in the pattern Rashomon set by a bound that depends on the empirical risk of the empirical risk minimizer and the Rashomon parameter, as we discussed in Proposition 10.

Expected pairwise disagreement. Following (Black et al., 2022) and (Marx et al., 2020), empirical expected pairwise disagreement $I(\hat{R}_{set}(\mathcal{F},\theta))$ over the Rashomon set can be defined as $I(\hat{R}_{set}(\mathcal{F},\theta)) = \mathbb{E}_{f_1,f_2\sim\hat{R}_{set}(\mathcal{F},\theta)}\hat{\mathbb{E}}_{z\sim S}\mathbb{1}_{[f_1(x)\neq f_2(x)]}$. Expected pairwise disagreement measures the average disagreement between every two hypotheses from the Rashomon set, while pattern diversity measures the average disagreement between two patterns from the Rashomon set. The expected pairwise disagreement is equivalent to pattern diversity when every pattern is achievable with the same probability by models from the Rashomon set. However, these metrics can be very different and we can have a small expected pairwise disagreement and larger pattern diversity as we show next.

Proposition 12 (Same pattern diversity but different expected pairwise disagreement)**.** *Consider finite Rashomon set $\hat{R}_{set}(\mathcal{F},\theta)$ of size $d\geq 2$. Let $\pi(\mathcal{F},\theta)$ be the pattern set of size Π, $2\leq\Pi\leq d$. Assume that every pattern except p_1 is achievable by only one hypothesis in the Rashomon set, and thus p_1 is achievable by $d-\Pi+1$ hypotheses. Let d^* be the current value of d, then as $d\to\infty$ (for example, by replicating hypotheses that realize p_1 an infinite number of times), expected pairwise disagreement converges to zero, $I(\hat{R}_{set}(\mathcal{F}_d,\theta))\to 0$, and pattern diversity does not change, $div(\hat{R}_{set}(\mathcal{F}_d,\theta)) = div(\hat{R}_{set}(\mathcal{F}_{d^*},\theta))$.*

Proof. The proof proceeds in two steps.

Pattern diversity. As d increases, the pattern set does not change, therefore for any $d\geq d^*$, $div(\hat{R}_{set}(\mathcal{F}_d,\theta)) = div(\hat{R}_{set}(\mathcal{F}_{d^*},\theta))$.

Expected pairwise disagreement. Given a pattern $pt\in\pi(\mathcal{F},\theta)$, let $P_{f\sim\hat{R}_{set}(\mathcal{F},\theta)}\left[pt = pt^f\right]$ be a probability with which this pattern is achieved by models from the Rashomon set. Since support for all patterns except pt_1 is 1, then $P_k = P_{f\sim\hat{R}_{set}(\mathcal{F},\theta)}\left[pt_k = pt^f\right] = \frac{1}{d}$ for $k = 2..\Pi$. And for pt_1 we have $P_1 = P_{f\sim\hat{R}_{set}(\mathcal{F},\theta)}\left[pt_1 = pt^f\right] = \frac{d-\Pi+1}{d}$. Then expected pairwise disagreement:

$$
\begin{aligned}
I(\hat{R}_{set}(\mathcal{F}_d,\theta)) &= \mathbb{E}_{f_1,f_2\sim\hat{R}_{set}(\mathcal{F},\theta)}\,\hat{\mathbb{E}}_{x\sim S}\mathbb{1}_{[f_1(x)\neq f_2(x)]} \\
&= \sum_{k=1}^{\Pi}\left[P_{f\sim\hat{R}_{set}(\mathcal{F},\theta)}\left[pt_k = pt^f\right]\sum_{j=1}^{\Pi}P_{f\sim\hat{R}_{set}(\mathcal{F},\theta)}\left[pt_j = pt^f\right]\frac{1}{n}H(pt_k,pt_j)\right] \\
&= \left(P_{f\sim\hat{R}_{set}(\mathcal{F},\theta)}\left[pt_1 = pt^f\right]\right)^2\frac{1}{n}H(pt_1,pt_1) \\
&+ 2P_{f\sim\hat{R}_{set}(\mathcal{F},\theta)}\left[pt_1 = pt^f\right]\sum_{j=2}^{\Pi}P_{f\sim\hat{R}_{set}(\mathcal{F},\theta)}\left[pt_j = pt^f\right]\frac{1}{n}H(pt_1,pt_j) \\
&+ \sum_{k=2}^{\Pi}\left[P_{f\sim\hat{R}_{set}(\mathcal{F},\theta)}\left[pt_k = pt^f\right]\sum_{j=2}^{\Pi}P_{f\sim\hat{R}_{set}(\mathcal{F},\theta)}\left[pt_j = pt^f\right]\frac{1}{n}H(pt_k,pt_j)\right] \\
&= 0 + 2\frac{d-\Pi+1}{d^2}\sum_{j=2}^{\Pi}\frac{1}{n}H(pt_1,pt_j) + \frac{1}{d^2}\sum_{k=2}^{\Pi}\sum_{j=2}^{\Pi}\frac{1}{n}H(pt_k,pt_j) \\
&= \frac{1}{d}\left(2\left(1-\frac{\Pi-1}{d}\right)\sum_{j=2}^{\Pi}\frac{1}{n}H(pt_1,pt_j) + \frac{1}{d}\sum_{k=2}^{\Pi}\sum_{j=2}^{\Pi}\frac{1}{n}H(pt_k,pt_j)\right).
\end{aligned}
$$

Therefore, as $d \to \infty$, $I(\hat{R}_{set}(\mathcal{F}_d,\theta)) \to 0$. ■

As we see from Proposition 12, we can change expected pairwise disagreement, for example, by adding multiple copies of the same functions f to the hypothesis space. Expected pairwise disagreement measures predictive multiplicity (Black et al., 2022), but it has the issue we showed above that it can depend on the weighting of hypotheses in the hypothesis space. In the case we described in Proposition 12, the multiplicity is small because one subset of hypotheses (which produce the same pattern) is weighted very heavily. Thus, expected pairwise disagreement can be influenced by overparameterization or poor choice of parameter space. We further illustrate the effect of re-parameterization on pairwise disagreement on a simple one-dimensional example in Figure 2.6. Pattern diversity does not depend on the parameter space and is computed in the pattern space. It is not impacted by any probability distribution or weighting on the hypotheses. Moreover, we can compute the pattern diversity by enumerating all possible patterns of the given finite dataset as described in Appendix 2.4. We cannot do the same for the pairwise disagreement metric without additional assumptions on the patterns' support.

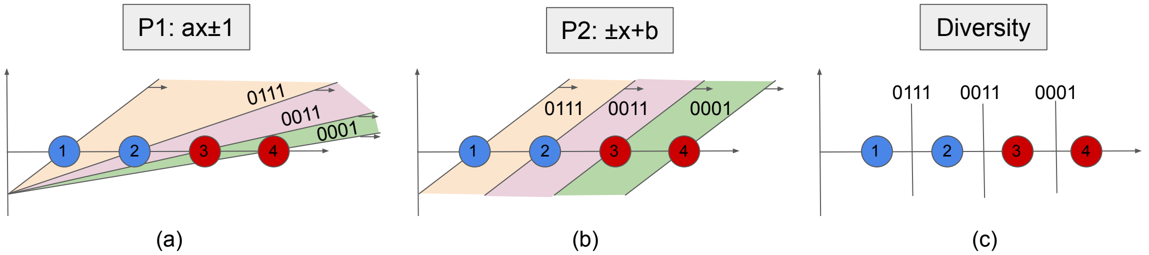


FIGURE 2.6: Illustration of how reparameterization changes pairwise disagreement metric. Consider a separable dataset of four data points with a real-valued feature in one dimension: $S = \{(1, 0), (2, 0), (3, 1), (4, 1)\}$ and a hypothesis space of linear models. Let the Rashomon parameter be $\theta = 0.25$. There are three patterns in the Rashomon set: 0111, 0011, and 0001. The pattern diversity (c) is 0.444. Consider two different parameterizations for the hypothesis space of linear models: $ax \pm 1$ and $\pm x + b$. These two parameterizations produce the same decision boundaries for the dataset S. For the parameterization $\pm x + b$ (b), each pattern is achieved with the same number of models. For the parameterization $ax \pm 1$ (a), more models will support patterns that are closer to the origin. The support of each pattern is shown in a different color. The pairwise disagreement metric is 0.321 for $ax \pm 1$ and 0.444 for $\pm x + b$. (For the parameterization $ax \pm 1$, we see that the pattern 0001 occurs when $a \in (1, \frac{1}{2})$, the pattern 0011 occurs when $a \in (\frac{1}{2}, \frac{1}{3})$, and the pattern 0111 occurs when $a \in (\frac{1}{3}, \frac{1}{4})$. Therefore, the pattern 0001 has probability $w_1 = \frac{1-\frac{1}{2}}{\frac{3}{4}} = 0.666$, the pattern 0011 has probability $w_2 = \frac{\frac{1}{2}-\frac{1}{3}}{\frac{3}{4}} = 0.222$, and the pattern 0111 has probability $w_3 = \frac{\frac{1}{3}-\frac{1}{4}}{\frac{3}{4}} = 0.111$. Recall that $H(cdot, \cdot)$ is the Hamming distance, then the pairwise disagreement metric is $w_1w_2H(0001, 0011) + w_1w_3H(0001, 0111) + w_2w_3H(0011, 0111) = w_1w_2 + 2w_1w_3 + w_2w_3 = 0.321$. For the parameterization $\pm x + b$, each pattern has equal probability $\frac{1}{3}$. We can then similarly calculate that the pairwise disagreement metric is 0.444). Note that if the data points are shifted together to the left, the difference in pairwise disagreement metrics for parameterizations in (a) and (b) will only grow.

2.5.2 Upper Bound on Pattern Diversity

Recall that a_i is the sample agreement over the pattern Rashomon set. When $a_i = 1$, then all patterns agreed and correctly classified z_i. If $a_i = \frac{1}{2}$, only half of the models were able to correctly predict the label. As we will show, when more samples have sample agreement near $\frac{1}{2}$, we have higher pattern diversity. We can compute pattern diversity using average sample agreements instead of the Hamming distance according to the theorem below.

Theorem 13 (Pattern diversity via sample agreement)**.** *For 0-1 loss, dataset S, and pattern Rashomon set $\pi(\mathcal{F}, \theta)$, pattern diversity can be computed as $div(\hat{R}_{set}(\mathcal{F}, \theta)) = \frac{2}{n}\sum_{i=1}^{n} a_i(1 - a_i)$, where $a_i = \frac{1}{\Pi}\sum_{k=1}^{\Pi} \mathbb{1}_{[pt_k^i = y_i]}$ is sample agreement over the pattern Rashomon set.*

Proof. Let $y \in \{0, 1\}$. We can transform $y \in \{-1, 1\}$ to $\{0, 1\}$, simply by adding one and dividing by two.

Recall that Hamming distance $H(pt_j, pt_k) = \sum_{i=1}^{n} \mathbb{1}_{[pt_j^i \neq pt_k^i]}$. Alternatively, we can rewrite logical XOR as $\mathbb{1}_{[pt_j^i \neq pt_k^i]} = pt_j^i(1 - pt_k^i) + pt_k^i(1 - pt_j^i)$. Denote $b_i = \frac{1}{\Pi}\sum_{j=1}^{\Pi} pt_j^i$, then from the pattern diversity definition:

$$div(\hat{R}_{set}(\mathcal{F}, \theta)) = \frac{1}{n\Pi\Pi}\sum_{j=1}^{\Pi}\sum_{k=1}^{\Pi}\sum_{i=1}^{n} \mathbb{1}_{[pt_j^i \neq pt_k^i]} =$$

$$= \frac{1}{n\Pi\Pi}\sum_{j=1}^{\Pi}\sum_{k=1}^{\Pi}\sum_{i=1}^{n} \left[pt_j^i(1 - pt_k^i) + pt_k^i(1 - pt_j^i)\right]$$

$$= \frac{1}{n\Pi\Pi}\sum_{j=1}^{\Pi}\sum_{k=1}^{\Pi}\sum_{i=1}^{n} \left[pt_j^i + pt_k^i - 2pt_k^i pt_j^i\right]$$

$$= \frac{1}{n\Pi}\sum_{i=1}^{n}\sum_{j=1}^{\Pi} \left[\frac{1}{\Pi}\sum_{k=1}^{\Pi} pt_j^i + \frac{1}{\Pi}\sum_{k=1}^{\Pi} pt_k^i - 2\frac{1}{\Pi}\sum_{k=1}^{\Pi} pt_k^i pt_j^i\right]$$

$$= \frac{1}{n\Pi}\sum_{i=1}^{n}\sum_{j=1}^{\Pi} \left[pt_j^i + b_i - 2b_i pt_j^i\right]$$

$$= \frac{1}{n}\sum_{i=1}^{n} \left[\frac{1}{\Pi}\sum_{j=1}^{\Pi} pt_j^i + \frac{1}{\Pi}\sum_{j=1}^{\Pi} b_i - 2\frac{1}{\Pi}\sum_{j=1}^{\Pi} b_i pt_j^i\right]$$

$$= \frac{1}{n}\sum_{i=1}^{n} \left[b_i + b_i - 2b_i^2\right]$$

$$= \frac{2}{n}\sum_{i=1}^{n} b_i(1 - b_i).$$

On the other hand, according to logical XNOR, we have that $\mathbb{1}_{[pt_k^i=y_i]} = pt_k^i y_i + (1 - pt_k^i)(1 - y_i)$, therefore we can rewrite a_i as:

$$
\begin{aligned}
a_i &= \frac{1}{\Pi}\sum_{k=1}^{\Pi} \mathbb{1}_{[pt_k^i=y_i]} \\
&= \frac{1}{\Pi}\sum_{k=1}^{\Pi} \left[pt_k^i y_i + (1 - pt_k^i)(1 - y_i)\right] \\
&= \frac{1}{\Pi}\sum_{k=1}^{\Pi} \left[2pt_k^i y_i + 1 - y_i - pt_k^i\right] \\
&= 2y_i\frac{1}{\Pi}\sum_{k=1}^{\Pi} pt_k^i + 1 - y_i - \frac{1}{\Pi}\sum_{k=1}^{\Pi} pt_k^i \\
&= 2y_i b_i + 1 - y_i - b_i.
\end{aligned}
$$

Since $y_i \in \{0,1\}$, then $y_i^2 = y_i$ and we have that:

$$
\begin{aligned}
\frac{2}{n}\sum_{i=1}^{n} a_i(1-a_i) &= \frac{2}{n}\sum_{i=1}^{n}(2y_ib_i + 1 - y_i - b_i)(-2y_ib_i + y_i + b_i) \\
&= \frac{2}{n}\sum_{i=1}^{n}(-4y_ib_i^2 + 2y_ib_i + 2y_ib_i^2 - 2y_ib_i + y_i + b_i \\
&+ 2y_ib_i - y_i - y_ib_i + 2y_ib_i^2 - y_ib_i - b_i^2) \\
&= \frac{2}{n}\sum_{i=1}^{n}(b_i - b_i^2) \\
&= \frac{2}{n}\sum_{i=1}^{n} b_i(1-b_i).
\end{aligned}
$$

Therefore we get:

$$
div(\hat{R}_{set}(\mathcal{F},\theta)) = \frac{2}{n}\sum_{i=1}^{n} b_i(1-b_i) = \frac{2}{n}\sum_{i=1}^{n} a_i(1-a_i).
$$

■

We next show that average sample agreement (over all samples z_i) over hypotheses that realize patterns in the pattern Rashomon set is negatively proportional to the average loss of these hypotheses. We use this intuition to derive an upper bound for average sample agreement and then discuss the upper bound for pattern diversity.

Let *hypothesis pattern set* $\mathcal{H}_{\pi(\mathcal{F},\theta)} \subset \hat{R}_{set}(\mathcal{F},\theta)$ be a set of unique hypotheses corresponding to each pattern[1] in $\pi(\mathcal{F},\theta)$, meaning that there is no $f_1^\pi, f_2^\pi \in \mathcal{H}_{\pi(\mathcal{F},\theta)}$, such that $f_1^\pi \neq f_2^\pi$, yet $pt^{f_1^\pi} = pt^{f_2^\pi}$.

Theorem 14. *Average sample agreement over the pattern Rashomon set is negatively proportional to the average loss of models in the hypothesis pattern Rashomon set $\mathcal{H}_\pi(\mathcal{F},\theta)$,*

$$\frac{1}{n}\sum_{i=1}^{n} a_i = 1 - \hat{L}_{avg}(\mathcal{H}_\pi(\mathcal{F},\theta)),$$

where $\hat{L}_{avg}(\mathcal{H}_\pi(\mathcal{F},\theta)) = \frac{1}{\Pi}\sum_{k=1}^{\Pi} \hat{L}(f_k^\pi)$. Moreover, when the Rashomon parameter $\theta = 0$, then

$$\frac{1}{n}\sum_{i=1}^{n} a_i = 1 - \hat{L}(\hat{f}).$$

Proof. For a given (x_i, y_i), when hypothesis f_k^π realizes pattern $pt^{f_k^\pi} = pt_k$, we have that $pt_k^i = f_k^\pi(x_i)$. Consider average sample agreement:

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n} a_i &= \frac{1}{n}\sum_{i=1}^{n}\frac{1}{\Pi}\sum_{k=1}^{\Pi}\mathbb{1}_{[pt_k^i = y_i]} \\
&= \frac{1}{n}\sum_{i=1}^{n}\left(1 - \frac{1}{\Pi}\sum_{k=1}^{\Pi}\mathbb{1}_{[pt_k^i \neq y_i]}\right) \\
&= 1 - \frac{1}{n}\sum_{i=1}^{n}\frac{1}{\Pi}\sum_{k=1}^{\Pi}\mathbb{1}_{[pt_k^i \neq y_i]} \\
&= 1 - \frac{1}{\Pi}\sum_{k=1}^{\Pi}\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}_{[f_k^\pi(x_i) \neq y_i]}
\end{aligned}$$

[1] Since there could be many hypotheses that achieve the same pattern, $\mathcal{H}_{\pi(\mathcal{F},\theta)}$ is not unique. We can work with any of them, as $\mathcal{H}_{\pi(\mathcal{F},\theta)}$ is simply a representation of the pattern set in the hypothesis space.

$$= 1 - \frac{1}{\Pi}\sum_{k=1}^{\Pi} \hat{L}(f_k^{\pi})$$

$$= 1 - \hat{L}_{avg}(\mathcal{H}_{\pi}(\mathcal{F}, \theta)).$$

When $\theta = 0$, for any k, $\hat{L}(\hat{f}) = \hat{L}(f_k^{\pi})$, therefore $\frac{1}{n}\sum_{i=1}^{n} a_i = 1 - \hat{L}(\hat{f})$. ■

Given the definition of models in the Rashomon set, we can derive an upper bound on average sample agreement in Corollary 15.

Corollary 15. *For any parameter $\theta > 0$, average sample agreement is upper and lower bounded by the empirical loss of the empirical risk minimizer,*

$$1 - \hat{L}(\hat{f}) - \theta \leq \frac{1}{n}\sum_{i=1}^{n} a_i \leq 1 - \hat{L}(\hat{f}).$$

Proof. Proof follows directly from Theorem 14 and the fact that for every model f from the Rashomon set, $\hat{L}(\hat{f}) \leq \hat{L}(f) \leq \hat{L}(\hat{f}) + \theta$. ■

Next, we upper bound the pattern diversity by the empirical risk of the empirical risk minimizer and the Rashomon parameter θ.

Theorem 16 (Upper bound on pattern diversity)**.** *Consider hypothesis space $\mathcal{F}$, 0-1 loss, and empirical risk minimizer $\hat{f}$. For any $\theta \geq 0$, pattern diversity can be upper bounded by*

$$div(\hat{R}_{set}(\mathcal{F}, \theta)) \leq 2(\hat{L}(\hat{f}) + \theta)(1 - (\hat{L}(\hat{f}) + \theta)) + 2\theta. \tag{2.9}$$

Proof. From the Cauchy–Schwarz inequality, we have that

$$\left(\sum_{i=1}^{n} a_i\right)^2 \leq \sum_{i=1}^{n} 1^2 \sum_{i=1}^{n} a_i^2 = n\sum_{i=1}^{n} a_i^2.$$

Given this and from the definition of pattern diversity and Corollary 15 we get that:

$$div(\hat{R}_{set}(\mathcal{F}, \theta)) = \frac{2}{n}\sum_{i=1}^{n} a_i(1 - a_i) = \frac{2}{n}\sum_{i=1}^{n} a_i - \frac{2}{n}\sum_{i=1}^{n} a_i^2$$

$$\leq \frac{2}{n}\sum_{i=1}^{n} a_i - \frac{2}{n^2}\left(\sum_{i=1}^{n} a_i\right)^2 = 2\left(\frac{1}{n}\sum_{i=1}^{n} a_i\right) - 2\left(\frac{1}{n}\sum_{i=1}^{n} a_i\right)^2$$

$$\leq 2(1-\hat{L}(\hat{f})) - 2(1-\hat{L}(\hat{f})-\theta)^2$$

$$= 2 - 2\hat{L}(\hat{f}) - 2 + 4(\hat{L}(\hat{f})+\theta) - 2(\hat{L}(\hat{f})+\theta)^2$$

$$= 2(\hat{L}(\hat{f}) + \theta - (\hat{L}(\hat{f})+\theta)^2 + \theta)$$

$$= 2(\hat{L}(\hat{f})+\theta)(1-(\hat{L}(\hat{f})+\theta)) + 2\theta.$$

When $\theta = 0$, then $div(\hat{R}_{set}(\mathcal{F}, 0)) \leq 2\hat{L}(\hat{f})(1-\hat{L}(\hat{f}))$. ■

The bound (2.9) emphasizes how important the performance of the empirical risk minimizer is for understanding pattern diversity. If dataset is well separated so that the empirical risk is small, then pattern diversity will also be small, as there are not many different ways to misclassify points and stay within the Rashomon set. As the dataset becomes noisier, on average, we expect the empirical risk to increase and thus pattern diversity as well. We will formally show this in Chapter 5.

2.6 Rashomon Set Characteristics for Different Datasets

Different characteristics of the Rashomon set describe various properties of the set. The Rashomon ratio measures how many equally performing models there are, while the pattern Rashomon ratio focuses on how many different predictions these models produce on the dataset. On the other hand, pattern diversity emphasizes how distinct these predictions are from each other. Each characteristic provides additional useful information to the machine learning practitioner, aiding in estimating the Rashomon Effect and understanding the properties of the learning problem at hand. To provide specific examples, we computed these characteristics for different datasets in Table 2.2 based on the hypothesis space of sparse decision trees of depth four. We set the Rashomon parameter to 5%. Note that it is very computationally expensive to compute the number of patterns in the hypothesis space of decision trees of depth 4, therefore we provide the number of patterns only (the numerator of the pattern Rashomon ratio) in Table 2.2.

Table 2.2: Rashomon set characteristics for different classification datasets based on the hypothesis space of sparse decision trees of depth four.

Dataset	Number of models in the Rashomon set	Number of patterns in the Rashomon set	Rashomon ratio	Pattern diversity
Car Evaluation	498	108	1.518e-19	0.131
Monks 2	1089	369	8.443e-17	0.246
Monks 1	33	10	2.559e-18	0.081
Monks 3	80	29	6.202e-18	0.206
Bar 7 (Coupon)	3939	1538	4.004e-18	0.068
COMPAS	19859	6193	3.144e-16	0.148
Breast Cancer Wisconsin	23200	6987	1.065e-14	0.068
Carryout Takeaway	18368	3699	5.599e-18	0.038
Cheap Restaurant	113189	30654	3.450e-17	0.081
Bar	13950	6524	4.252e-18	0.127
Coffee House	3483	1168	1.062e-18	0.234

In Table 2.2, datasets with the highest pattern diversity (Monks 2, Monks 3, Coffee House; see the description of the datasets in Table B.2) have different numbers of patterns. The larger number of patterns in the Coffee House dataset is most likely caused by a larger number of data points (3816 as compared to 169 for Monks 2). Conversely, Monks 2 is originally a noisier dataset than Monks 3, indicated by its lower empirical risk, likely resulting in a higher number of patterns than Monks 3. The Breast Cancer Wisconsin dataset has a larger Rashomon ratio, containing a relatively large number of models in the Rashomon set. One possible reason why the Cheap Restaurant dataset contains more models than Breast Cancer Wisconsin could be its higher number of features, providing more possibilities to construct different trees in the Rashomon set.

Now that we have introduced different characteristics of the Rashomon set and established the Rashomon ratio as a simplicity measure, we can shift our focus to proving simplicity and generalization properties of models in the Rashomon set. This is critical to our thesis that simple-yet-accurate models exist.

3. When Rashomon Sets are Large, Simple-yet-Accurate Models Exist

Numerous works provide generalization bounds based on different complexity measures under different assumptions. Some discuss Rademacher (Kakade et al., 2008; Srebro et al., 2010) and Gaussian complexities (Kakade et al., 2008), PAC-Bayes theorems (Langford & Shawe-Taylor, 2002), covering numbers bounds (Zhou, 2002), and margin bounds (Koltchinskii & Panchenko, 2002; Schapire et al., 1998; Vapnik & Chervonenkis, 1971). In contrast, under assumptions elaborated in Section 3.1, the Rashomon ratio provides a certificate of the existence of a simpler model that generalizes. Is some of our bounds we use an approximating set, which is also used throughout the literature on learning theory (Lecué, 2011; Lugosi & Wegkamp, 2004; Mendelson, 2003; Schapire et al., 1998). An 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. In this Chapter, we assume that we will have two hypothesis spaces, a simple and a more complex one with good approximating properties.

3.1 Rashomon Set Models: Simplicity and Generalization

Consider two hypothesis (functional) spaces with different levels of complexity, where the lower-complexity space serves as a good *approximating set* (i.e., a good *cover*) for the higher-complexity space. The hypothesis 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 hypothesis 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 & Rudin, 2016), which are sparse linear models, with at most d' nonzero integer coefficients, $d' \ll d$. Another example is if the more complex space $\mathcal{F}_2$ consists of boosted decision trees, and $\mathcal{F}_1$ consists of single trees. 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 setup, we have several questions to answer:

1. What if the higher-complexity hypothesis space we chose were more complex than necessary for modeling the data? In that case, if we had instead used the simpler model class $\mathcal{F}_1$, would we still get a model that is (almost) as good as we could have obtained using the more complex class $\mathcal{F}_2$? If so, perhaps we can leverage the complexity of the simpler model class $\mathcal{F}_1$ for generalization bounds on our model rather than the more complex class $\mathcal{F}_2$. We answer this question in Section 3.1.1, where a property on the complex space that will help us is that **the true Rashomon set of $\mathcal{F}_2$ is large enough to admit a simpler model**. We do not need to know what this model is and we may never discover it (we would likely discover a different model using data).
2. Under what conditions on the complex and simpler model classes does the property we mentioned above (that the Rashomon set includes simpler models) hold? Does it hold often? As it turns out, under natural conditions on the function class and loss function, a large Rashomon set in the complex class does imply the existence of simple-yet-accurate models. We identify these conditions in Section 3.1.2, namely that the loss function is smooth, and that $\mathcal{F}_1$ serves as a cover for $\mathcal{F}_2$. Thus, **under these natural conditions that occur in practice, a large Rashomon set for a complex class of functions implies the existence of a simple-yet-accurate model**.

The bounds we present in Section 3.1.1 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 bounding train loss of one function (a *simpler* function) with test loss of another (the optimal model in a *more complex* function class). Standard learning theory analysis handles the single function case nicely; we are concerned with other questions here.

3.1.1 The True Rashomon Set Can Be Very Helpful

As in classic Occam's razor bounds, we start with finite hypothesis spaces. 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 a guarantee that a model we produce using a simpler function class $\mathcal{F}_1$ on our data could be approximately as good as the test performance of the best model from $\mathcal{F}_2$? In the following theorem, we will make a key assumption that allows us to do this: we assume that the Rashomon set of $\mathcal{F}_2$ includes a member of the simpler class of functions, $\mathcal{F}_1$, even if we do not know which function it is. Later, in Section 3.1.2, we show conditions under which simple models from $\mathcal{F}_1$ are proven to exist in the Rashomon set of $\mathcal{F}_2$, which depends on the size of $\mathcal{F}_2$'s Rashomon set. Here, $|\mathcal{F}|$ denotes the cardinality of the finite space $\mathcal{F}$. These bounds can be generalized to infinite hypothesis spaces with a simple extension to covering numbers, but they are designed for intuition, which works nicely with finite hypothesis spaces. Again, this is different from a regular learning theory bound as it does not consider the generalization of just one function.

Theorem 17 (The advantage of a true Rashomon set)**.** *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 $f_2^* \in \text{argmin}_{f_2 \in \mathcal{F}_2} L(f_2)$. Assume that the true Rashomon set includes a function from $\mathcal{F}_1$, so there exists a model $\tilde{f}_1 \in \mathcal{F}_1$ such that $\tilde{f}_1 \in R_{set}(\mathcal{F}_2, \gamma)$. (Note that we do not know $\tilde{f}_1$.) In that case, for any $\epsilon > 0$ with probability at least $1 - \epsilon$ with respect to the random draw of data:*

$$L(f_2^*) - b\sqrt{\frac{\log|\mathcal{F}_1| + \log 2/\epsilon}{2n}} \leq \hat{L}(\hat{f}_1) \leq L(f_2^*) + \gamma + b\sqrt{\frac{\log 1/\epsilon}{2n}}, \tag{3.1}$$

where $\hat{f}_1 \in \text{argmin}_{f_1 \in \mathcal{F}_1} \hat{L}(f_1)$. (Unlike $\tilde{f}_1$, we do know $\hat{f}_1$ because we can calculate it.)

Proof. **Lower bound.** 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 finite hypothesis space $\mathcal{F}_1$:

$$L(f_1) \leq \hat{L}(f_1) + b\sqrt{\frac{\log|\mathcal{F}_1| + \log 2/\epsilon}{2n}}. \tag{3.2}$$

Combining this Occam'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:

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

Upper bound. By the assumption of the theorem, we have that $L(\tilde{f}_1) \le L(f_2^*) + \gamma$. Also, by the definition of an optimal model f_1^*, $L(f_1^*) \le L(\tilde{f}_1)$. Combining these, we get that $L(f_1^*) \le L(\tilde{f}_1) \le L(f_2^*) + \gamma$. Thus f_1^* is in the true Rashomon set of $\mathcal{F}_2$ with parameter γ. Alternatively, f_1^* is in the true anchored Rashomon set of $\mathcal{F}_2$ with parameter $\eta = L(f_2^*) + \gamma$, $f_1^* \in R_{set}^{anc}(\mathcal{F}_2, \eta)$. Following Proposition 2, 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 $\hat{R}_{set}^{anc}(\mathcal{F}_2, \eta+\epsilon_1)$, and therefore, with high probability, $\hat{L}(f_1^*) \le \eta+\epsilon_1$. 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 $\hat{L}(f_1^*) \le \eta + b\sqrt{\frac{\log 1/\epsilon}{2n}}$. Further, by definition of the empirical risk minimizer and given that $\eta = L(f_2^*) + \gamma$ we get:

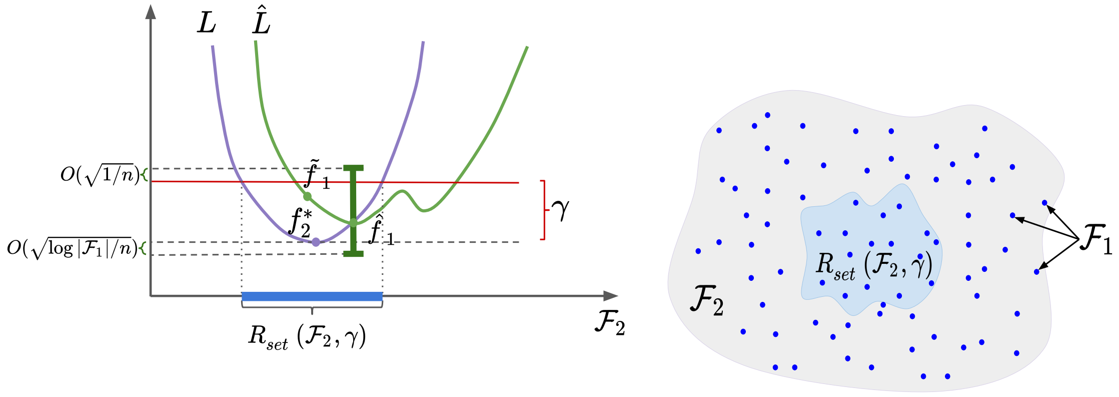
$$\hat{L}(\hat{f}_1) \le \hat{L}(f_1^*) \le L(f_2^*) + \gamma + b\sqrt{\frac{\log 1/\epsilon}{2n}}.$$

Combining the previous two equations yields the statement of the theorem.

■

That is, we can bound the best empirical model from $\mathcal{F}_1$ with the true risk of the best model within $\mathcal{F}_2$ (Figure 3.1 (a)). Thus, if the Rashomon set is large enough to include a single model from $\mathcal{F}_1$, we can work with the simpler class $\mathcal{F}_1$ in practice and achieve strong performance guarantees.

The main assumption in Theorem 17 is about the population, and does not rely on the sample. It relies only on the existence of one special function in the true Rashomon set. There are no smoothness assumptions on the loss function. If the main assumption of this theorem holds, then we gain the benefit of guarantees on $\mathcal{F}_2$ from looking only at $\mathcal{F}_1$



(a)

(b)

FIGURE 3.1: (a) For $\mathcal{F}_1 \subset \mathcal{F}_2$, the empirical risk of $\mathcal{F}_1$ is bounded by the true risk of $\mathcal{F}_2$ and γ if there exists a model $\tilde{f}_1$ in the intersection of $\mathcal{F}_1$ and the Rashomon set of $\mathcal{F}_2$ as shown in Theorem 17. (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 a high probability there will be a model from $\mathcal{F}_1$ that will be within the Rashomon set of $\mathcal{F}_2$.

empirically. We cannot check whether the assumption holds since it involves the true risk, but practitioners can reap the benefits of it anyway: The possibility of a large Rashomon set may embolden the practitioner to minimize over $\mathcal{F}_1$, achieving test error close to the best of $\mathcal{F}_2$ if the conditions of Theorem 17 are indeed satisfied.

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 $\mathcal{F}_2$ (as illustrated in Figure 3.1(b)) 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 within $\mathcal{F}_2$, which, arguably, could approximate the way some simpler spaces are embedded in more complex spaces.

If $\mathcal{F}_1$ is a random sample of functions from $\mathcal{F}_2$, and if $\mathcal{F}_2$ has a large true Rashomon set, then the true Rashomon set is likely to include at least one model from $\mathcal{F}_1$. This is formalized below.

Lemma 18. *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 Rashomon ratio of the hypothesis space $\mathcal{F}_2$ is at least*

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

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

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

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

In the case of Lemma 18, Theorem 17 applies, and therefore we have Theorem 19.

Theorem 19 (Example of the advantage of a large true Rashomon set)**.** *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 loss l bounded by b, if the Rashomon ratio is at least*

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

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, the assumptions of Theorem 17 hold and thus the bound (3.1) *holds.*

Table 3.1: Examples of the possible usage of Theorem 19.

If $\lvert F_1 \rvert = 100000$ then to get the bound (3.1) to hold with probability at least 99% the Rashomon ratio should be $R_{ratio}(\mathcal{F}_2, \gamma) \geq 0.0053\%$.
If $\lvert F_1 \rvert = 10000$ then to get the bound (3.1) to hold with probability at least 99% the Rashomon ratio should be $R_{ratio}(\mathcal{F}_2, \gamma) \geq 0.053\%$.
If $\lvert F_1 \rvert = 1000$ then to get the bound (3.1) to hold with probability at least 99% the Rashomon ratio should be $R_{ratio}(\mathcal{F}_2, \gamma) \geq 0.53\%$.

Proof. According to the Lemma 18, for any $\epsilon > 0$ with probability at least $1-\epsilon$ with respect to the random draw of functions, if the Rashomon set it at least $R_{ratio}(\mathcal{F}_2, \gamma) \geq 1 - \epsilon^{\frac{1}{|\mathcal{F}_1|}}$, then the Rashomon set contains at least one model $\tilde{f}$ from $\mathcal{F}_1$. In that case, according to Theorem 17 with probability at least $1-\epsilon$ with respect to the random draw of data, the bound (3.1) holds. Therefore with probability at least $(1-\epsilon)^2$ we get the statement of the theorem. ■

Table 3.1 shows possible values of the lower bound on the Rashomon ratio, given $|\mathcal{F}_1|$ and ϵ. For example, the first line of the table states that if at least a tiny fraction (0.0053%) of the complex function space $\mathcal{F}_2$ consists of good models, and there exists at least 100,000 simple functions in $\mathcal{F}_1$, then the chance that we will find an accurate-but-simple model on our dataset is over 99%.

The intuition for Theorem 19 holds beyond the case when $\mathcal{F}_1$ is randomly sampled from $\mathcal{F}_2$, it holds whenever $\mathcal{F}_1$ covers $\mathcal{F}_2$ sufficiently well. This intuition is that as the true 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 $\mathcal{F}_2$.

3.1.2 Proving the Existence of Simple-yet-Accurate Models with Good Generalization

Theorems 17 and 19 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 $\mathcal{F}_1$. Thus, the next analysis makes two improvements: (1) it studies empirical Rashomon sets instead of true Rashomon sets, (2) it substitutes the unrealistic random draw assumption for a realistic smoothness assumption. We now assume smoothness

Table 3.2: Examples of function approximation in different hypothesis spaces: a function from space $\mathcal{F}_1$ approximates a function in space $\mathcal{F}_2$ with given guarantee δ.

$\mathcal{F}_2$	$\mathcal{F}_1$	δ (depends on parameters in bounds below)	Source
$f \in L_\infty(\Omega)$, $\|f\|_\infty \in [m, M]$	$s_N \in S(\Omega)$, s_N—piecewise constant, N—number of constants	$\|f - s_N\|_\infty \le \frac{M-m}{2N}$	Davydov (2011) and DeVore (1998)
$f \in W^1_p(\Omega)$, $1 \le p \le \infty$, where W^1_p 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	$\|f - s_\Delta(f)\|_p \le CN^{-1/d}\|f\|_{W^1_p\Omega}$	Davydov (2011)
$f \in \{x^k,\, k \in N\}$	$P(n)$—polynomials of degree at most $n \in N$	$\|f - P(n)\|_\infty \le \frac{1}{2^{k-1}}\sum_{j>(n+k)/2}\binom{k}{j}$	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	$\|f - P(d)\|_\infty \le \mathcal{O}(\sqrt{n(n-\Gamma(f))})$	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, $n \in \mathbb{N}$	$\sup_{x\in[a,b]} \|f(x) - N_n(x)\| \le \frac{5M}{2}\left(\frac{b-a}{n}\right)^\alpha$	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	$\inf_{p\in P_r}\|f - p\|_{L_p(I)} \le C\|I\|^{r/d}\|f\|_{W^r(L_p(I))}$	DeVore (1998)

of the loss over the function space.

The field of Approximation Theory provides general conditions under which classes of functions can approximate each other. Given a target function from one class, we want to know whether a sequence of functions from another class can converge to the target. Table 3.2 shows classes of functions $\mathcal{F}_2$ that can be approximated by classes $\mathcal{F}_1$. For instance, piecewise constant functions, such as decision trees, can approximate smooth functions.

For a hypothesis space $\mathcal{F}$ and some $f' \in \mathcal{F}$, define the δ-ball of functions centered at f' as $B_\delta(f') = \{f \in \mathcal{F} : \|f' - f\|_p \le \delta\}$. A loss $l : \mathcal{F} \times \mathcal{X} \to \mathcal{Y}$ is said to be K-*Lipschitz*, $K \ge 0$, if for all $f_1, f_2 \in \mathcal{F}$ and for all $z \in \mathcal{Z}$: $|l(f_1, z) - l(f_2, z)| \le K\|f_1 - f_2\|_p$. 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 $\mathcal{X}$. Define a δ-packing as a finite set $\Xi = \{\xi_1, ..., \xi_k | \xi_i \in \mathcal{F}\}$ such that $\|\xi_i - \xi_j\|_p > \delta$, meaning that $B_{\delta/2}(\xi_i) \cap B_{\delta/2}(\xi_j) = \varnothing$ for all $i \neq j$. The *packing number* $\mathcal{B}(\mathcal{F}, \delta)$ is the largest δ-packing.

Theorem 20 below uses the approximating set argument from the previous subsection, but now requires the Rashomon set to be large enough to include balls of functions rather than using the random draw assumption. As long as the set of simpler functions is distributed well among the full hypothesis space, each ball contains at least one function from the simpler class.

Theorem 20 (Existence of multiple simpler models)**.** *For K-Lipschitz loss l bounded by b, consider hypothesis spaces $\mathcal{F}_1$ and $\mathcal{F}_2$, $\mathcal{F}_1 \subset \mathcal{F}_2$. 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 $f_1 \in \mathcal{F}_1$ 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 the 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 hypothesis 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 $\|\xi_i - \xi_j\|_p > 2\delta$ for all $i \neq j$. 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 $\|\xi_i - \bar{f}_1^i\|_p \leq \delta$ (this is the assumption that $\mathcal{F}_1$ serves as a good cover for $\mathcal{F}_2$). 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 P. L. Bartlett and Mendelson (2002). ∎

From Theorem 20, we see that since larger Rashomon sets have larger packing numbers, they contain more simpler models with good generalization guarantees. Figure 3.2(a)-(b)

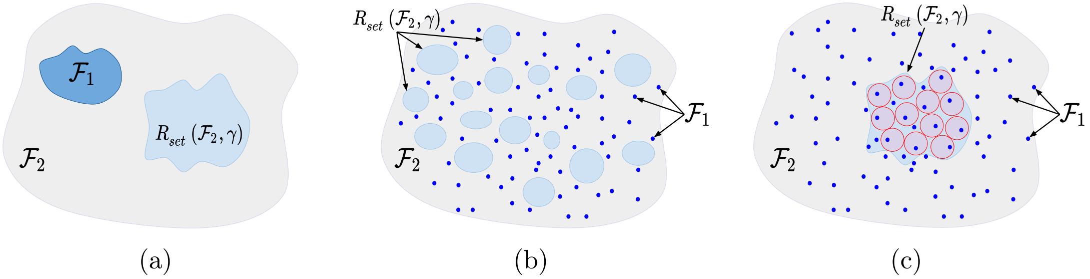


FIGURE 3.2: (a)-(b) An illustration that shows why both approximation and smoothness assumptions in Theorem 20 are important. If the approximation assumption does not hold, $\mathcal{F}_1$ can be concentrated in $\mathcal{F}_2$ and thus does not necessarily intersect with the Rashomon set of $\mathcal{F}_2$ (a). If the smoothness assumption does not hold, $\mathcal{F}_1$ is discontinuous and thus does not necessarily intersect with the Rashomon set of $\mathcal{F}_2$ even though $\mathcal{F}_1$ is dense in $\mathcal{F}_2$ (b). (c) An illustration where the Rashomon set contains at least its packing number of simple models when both assumptions hold.

illustrate why both approximation and smoothness assumptions are important for Theorem 20 to show the existence of simpler models in the empirical Rashomon set. When both of them are satisfied, the Rashomon set contains multiple simple models with good generalization guarantees as schematically shown in Figure 3.2 (c). Note that in Theorem 20, other complexity measures from learning theory could be used. We chose Rademacher complexity as it provides the tightest bound among standard complexity measures.

Theorem 20 has practical implications. *If the Rashomon set is large, and the smoothness conditions are obeyed, Theorem 20 shows that many simple-yet-accurate models would exist, prior to actually finding them.* Knowledge that simple models exist implies it will be worthwhile to actually solve the difficult optimization problem to find a simple model.

Thus, *if* the Rashomon set is large, we have a guarantee. But how will we know when is the Rashomon set large? This is what we answer in the next section.

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

We expect that in many real-world applications of machine learning, properties similar to the assumptions behind our theorems hold, i.e., that large enough Rashomon sets intersect simpler hypothesis spaces in ways that lead to or explain good performance. This conjecture is difficult to verify theoretically because it is not a mathematical conjecture about the structure of two specific function spaces, but a statement about many function spaces, and how they interact with commonly occurring datasets. Thus, we consider this question empirically.

Our experiments will demonstrate that, in the case where Rashomon sets are large, two conclusions follow that are consistent with our theoretical development. First, *training* performance in *simpler* hypothesis spaces is correlated with *test* performance in the more *complex* hypothesis spaces (Theorem 17), and second, that good *training* performance in a *simpler* space $\mathcal{F}_1$ correlates with *good generalization* performance of other models in the more *complex* space $\mathcal{F}_2$. Most importantly, our experiments suggest an intriguing alternative to the often difficult computational problem of directly estimating the size of the Rashomon set, namely that *similar performance across a range of algorithms with different hypothesis spaces is strongly correlated with a large Rashomon set.*

Now we will describe our experimental setup for arriving at these conclusions.

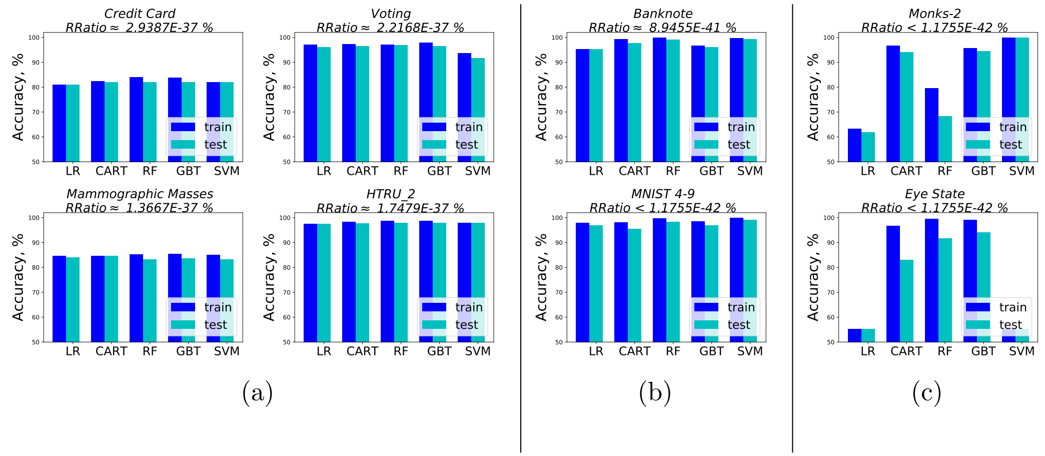
3.2.1 Experimental Design

Datasets. We used 38 machine learning classification datasets from the UCI Machine Learning Repository (Dua & Graff, 2019), among which 16 have categorical features and 22 have real-valued features. The majority of the datasets are binary classification datasets and we adapted the rest to binary classification (as shown in Table A.1 in Appendix A.1) to make importance sampling easier (as discussed in Appendix A.2). The number of features varies from 3 to 784, with the majority of the datasets being in the 15–25 feature range. Appendix A.1 contains a description of the datasets we considered.

Definition of complex hypothesis space. For these experiments, we will consider $\mathcal{F}_2$ to be the union ($\mathcal{F}_{union}$) of the hypothesis spaces of five popular machine learning algorithms: logistic regression (LR), CART, random forests (RF), gradient boosted trees (GBT), and support vector machines with RBF kernels (SVM). CART, RF, and GBT 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 trees in the ensemble. SVMs were tuned by varying the regularization parameter and the kernel coefficient and LR by varying the regularization parameter. Appendix A.2 discusses the effect of regularization on the model class. We chose algorithms that search hypothesis spaces of different complexity to ensure that these algorithms produce diverse models. The notion of $\mathcal{F}_2$ as a union of hypothesis spaces may seem surprising at first, but it is consistent with how many machine learning practitioners approach problems by running a collection of machine learning techniques in parallel and comparing the results, creating a *de facto* union space. Our experiment has three steps, as follows.

Step 1: Run all machine learning algorithms. We obtain training and generalization performance from all algorithms (logistic regression, CART, random forests, gradient-boosted trees, and SVM with RBF kernels) on all datasets.

Step 2: Estimate the size of the Rashomon set. It is not possible to measure the Rashomon set of such a complex model space, so we will estimate its size by sampling from an approximating set, which is decision trees of bounded depth. Decision trees are easy to sample and can refine an input space arbitrarily finely as tree depth increases. With sufficient depth, they can approximate many other types of hypothesis spaces, including those used by other machine learning methods. Thus, we will measure the size of the Rashomon set and Rashomon ratio in decision trees of depth seven as a surrogate for measuring these quantities in $\mathcal{F}_2$. The suitability of these trees for this role is an empirical observation about the datasets we have used; they may not be a suitable surrogate for some other datasets, e.g., imagery data. We measure the size of the empirical Rashomon ratio as a surrogate for the true Rashomon ratio when referring to Theorem 17. To estimate the Rashomon ratio of depth seven decision trees, we used importance sampling. 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 dataset 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



(a) (b) (c)

FIGURE 3.3: (a) Examples of experiments on four datasets 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. We show all 38 datasets in the Appendix A.2.

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 an 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 further discuss experimental setup in Appendix A.2.

Step 3: See if a large Rashomon Set in Step 2 correlates with performance differences in Step 1. By construction, the hypothesis spaces of each of the machine learning algorithms we consider are embedded in $\mathcal{F}_2$. RF and GBT both enjoy extremely rich hypothesis spaces that are likely close in size to $\mathcal{F}_2$ itself. LR and CART are less expressive than these others, so we will view LR and CART as simpler, $\mathcal{F}_1$ type, hypothesis spaces. Our question to answer is whether a large Rashomon set measured in Step 2 correlates with the functions from $\mathcal{F}_1$ (CART, LR) having performance as good as that of $\mathcal{F}_2$ (GBT, RF, SVM) as our theory predicts it will.

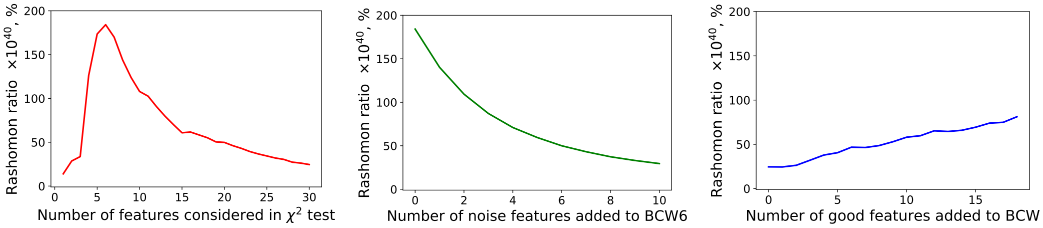
3.2.2 Experimental Results

Figure 3.3(a) shows the performance of the five machine learning algorithms on datasets for which the Rashomon ratio was largest, as measured in the space of decision trees of depth 7. Performance for all datasets is shown in Figures A.1 and A.2 in Appendix A.2. Across the 38 datasets considered, we observe ***larger Rashomon ratios led to approximately similar training results across all algorithms*** (within $\sim 5\%$ difference between algorithms). Here, large Rashomon ratios are on the order of $10^{-37}\%$ or $10^{-38}\%$, whereas small Rashomon ratios are $10^{-40}\%$ or less[1]. Moreover, all of the models chosen by the algorithms, including simpler $\mathcal{F}_1$ type models, generalized well (the differences between training and test errors are within $\sim 5\%$). These results are consistent with our thesis that larger Rashomon sets lead to the existence of accurate-yet-simpler models (in agreement with the theory in Section 3.1.1), and that larger Rashomon sets lead to better generalization. *The results also imply that large Rashomon sets do occur in many datasets, with the Rashomon Effect being large enough to include simpler models in practice (in agreement with Section 3.1.2).*

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 (see Figure 3.3(b)). This observation could be explained in several different ways. Mainly, the Rashomon ratio is not the only driver of good generalization performance. The amount of data is one obvious additional driver. Appendix A.2 discusses this further. Quality of features is another driver, as discussed in Section 3.3.

Our second main result is that in ***all*** cases where large Rashomon ratios were observed, ***test performance was consistent with training performance across algorithms of varying complexity.*** This correlation between the size of the Rashomon ratio and consistent generalization performance suggests an indirect means of assessing the size of the Rashomon ratio as an alternative to the computationally intensive approach of sampling.

[1] For other datasets and other metrics of measuring the Rashomon set, the results might be different.



(a) Reducing features

(b) Adding noise

(c) Adding good features

FIGURE 3.4: An illustration of the influence of feature quality on the Rashomon ratio for the Breast Cancer Wisconsin dataset (BCW). (a) shows the Rashomon ratio for the dataset with different numbers of significant features according to a χ^2 test. Denote the BCW with the six most significant features as BCW6. (b) depicts the correspondence between the Rashomon ratio and different numbers of noisy features added to the BCW6 dataset. The noise features are sampled from the 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 dataset. We iteratively add one out of six features from the BCW6 dataset 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 dataset.

When consistent training and test performance across algorithms is observed, this *may* indicate a large Rashomon ratio.

One thing we notably did not observe were cases where algorithms did not generalize, performance differed across algorithms, and the Rashomon set was large. Across all 38 datasets, we did not observe cases where the Rashomon set was large and performance differed among algorithms.

Figure 3.3(c) shows small Rashomon sets, where we observe wildly different performance across algorithms, where sometimes the models generalize and sometimes they do not. We show one example of each of these cases in Figure 3.3(c). Our theory does not apply to the case of small Rashomon sets, and thus there is no guarantee for such datasets.

3.3 Quality of the Features and Rashomon Ratio

In our experiments, we observed a connection between the quality of the features and Rashomon ratios. The Rashomon ratio in its simplest form, under uniform prior on the hypothesis space, is the fraction of models that are inside the Rashomon set compared to the models in the hypothesis space. When a dataset 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 dataset.

Additionally, if we augment a dataset 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 datasets, and we are interested in whether these cases have potentially influenced our experimental results in Section 3.2 in our observed Rashomon ratios. To investigate this, we augmented a dataset with noise features, and separately, augmented the same dataset 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) dataset (shortly, BCW), which has approximately six important features. The results are shown in Figure 3.4. As before, our hypothesis space is decision trees of depth seven.

Irrelevant features. If the dataset contains a lot of irrelevant or noisy features, we expect the Rashomon set to be relatively small compared to the hypothesis space. Figure 3.4(a) shows how the Rashomon ratio changes as we iteratively decrease the number of features in the Wine dataset, 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 3.4(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 3.4(c) shows how the Rashomon ratio changes for the BCW dataset as we add more copies of the six most significant features. We observe that the Rashomon ratio increases. By adding copies of relevant features, we increased the number of trees at a given depth that could be good enough to be in the Rashomon set.

For the binary dataset and the hypothesis space of decision trees, we further formalize intuition of the influence of redundant and irrelevant features on the Rashomon ratio in the theorem below:

Theorem 21 (Rashomon ratio increases with more good features and decreases with more bad features.)**.** *For a dataset $S = X \times Y$ with binary feature matrix $X \in \{0,1\}^{n\times p}$, consider a hypothesis space $\mathcal{F}_d^p$ of fully grown trees of depth d. Assume that there is a set of $p_{bad} \geq d$ "bad" features such that if a model is not in the Rashomon set, it is using only the bad features. Then:*

1. *As we remove p^{rem}* ***good*** *features from the dataset, $p^{rem} \in [1, p-|p_{bad}|)$, the Rashomon ratio* ***decreases****,*

$$\hat{R}_{ratio}(\mathcal{F}_d^p, \theta) > \hat{R}_{ratio}(\mathcal{F}_d^{p-p^{rem}}, \theta).$$

2. *As we remove p^{rem}* ***bad*** *features from the dataset, $p^{rem} \in [1, |p_{bad}|)$, the Rashomon ratio* ***increases****,*

$$\hat{R}_{ratio}(\mathcal{F}_d^p, \theta) < \hat{R}_{ratio}(\mathcal{F}_d^{p-p^{rem}}, \theta).$$

Proof. The hypothesis space of fully-grown trees of depth d contains

$$|\mathcal{F}_d^p| = 2^{2^d} \prod_{k=1}^{d} (p-k+1)^{2^{k-1}}$$

trees, where 2 is the number of label options each leaf can have, 2^d is the number of leaves we have, $\prod_{k=1}^{d}$ is the product over all depth levels in a tree, 2^{k-1} is the number of nodes we have at that level, and $p-(k-1)$ is the number of options we have to choose from given that the previous features were used in the path from the root. We do not count symmetric trees, meaning that we always assume that split $=0$ is on the left and $=1$ is on the right.

Now let's compute the size of the Rashomon set. If a tree is not in the Rashomon set, it has only features from the set p_{bad}. Therefore, trees in the Rashomon set must have at least one "good" feature at some node, where good means that the feature is not in p_{bad}. The cardinality of the Rashomon set is:

$$\hat{R}_{set}(\mathcal{F}_d^p, \theta) = 2^{2^d} \prod_{k=1}^{d} (p-k+1)^{2^{k-1}} - 2^{2^d} \prod_{k=1}^{d} (|p_{bad}|-k+1)^{2^{k-1}},$$

meaning that among all models, we do not consider those that consist of bad features only (since $p_{bad} \geq d$, there exists at least one such tree). Then the Rashomon ratio is:

$$\begin{aligned}\hat{R}_{ratio}(\mathcal{F}_d^p, \theta) = \frac{|\hat{R}_{set}(\mathcal{F}_d^p, \theta)|}{|\mathcal{F}_d^p|} &= \frac{2^{2^d} \prod_{k=1}^{d} (p-k+1)^{2^{k-1}} - 2^{2^d} \prod_{k=1}^{d} (|p_{bad}|-k+1)^{2^{k-1}}}{2^{2^d} \prod_{k=1}^{d} (p-k+1)^{2^{k-1}}} \\ &= 1 - \frac{\prod_{k=1}^{d} (|p_{bad}|-k+1)^{2^{k-1}}}{\prod_{k=1}^{d} (p-k+1)^{2^{k-1}}} \\ &= 1 - \prod_{k=1}^{d} \left(\frac{|p_{bad}|-k+1}{p-k+1} \right)^{2^{k-1}}.\end{aligned}$$

1. Assume that we have $p^{before} = p$ features and we removed p^{rem} **good** features. Then the number of features after removal is $p^{after} = p^{before} - p^{rem} < p^{before}$, and number of bad features did not change $p_{bad}^{before} = p_{bad}^{after} = |p_{bad}|$. Therefore, we have that:

$$\begin{aligned}\hat{R}_{ratio}(\mathcal{F}_d^p, \theta) &= 1 - \prod_{k=1}^{d} \left(\frac{p_{bad}^{before} - k + 1}{p^{before} - k + 1} \right)^{2^{k-1}} \\ &= 1 - \prod_{k=1}^{d} \left(\frac{p_{bad}^{after} - k + 1}{p^{before} - k + 1} \right)^{2^{k-1}}\end{aligned}$$

$$> 1 - \prod_{k=1}^{d} \left(\frac{p_{bad}^{after} - k + 1}{p^{after} - k + 1} \right)^{2^{k-1}} = \hat{R}_{ratio}(\mathcal{F}_d^{p-p^{rem}}, \theta).$$

Thus, we showed that the Rashomon ratio decreases as we remove good features from the dataset.

2. Assume that we have $p^{before} = p$ features and we removed p^{rem} **bad** features. Then the number of features after removal is $p^{after} = p^{before} - p^{rem} < p^{before}$, and number of good features did not change $p_{good}^{before} = p_{good}^{after} = p - |p_{bad}|$. Therefore, we have that:

$$\begin{aligned}
\hat{R}_{ratio}(\mathcal{F}_d^p, \theta) &= 1 - \prod_{k=1}^{d} \left(\frac{p_{bad}^{before} - k + 1}{p^{before} - k + 1} \right)^{2^{k-1}} \\
&= 1 - \prod_{k=1}^{d} \left(\frac{p^{before} - p_{good}^{before} - k + 1}{p^{before} - k + 1} \right)^{2^{k-1}} \\
&= 1 - \prod_{k=1}^{d} \left(1 - \frac{p_{good}^{before}}{p^{before} - k + 1} \right)^{2^{k-1}} \\
&= 1 - \prod_{k=1}^{d} \left(1 - \frac{p_{good}^{after}}{p^{before} - k + 1} \right)^{2^{k-1}} \\
&< 1 - \prod_{k=1}^{d} \left(1 - \frac{p_{good}^{after}}{p^{after} - k + 1} \right)^{2^{k-1}} = \\
&< 1 - \prod_{k=1}^{d} \left(\frac{p_{bad}^{after} - k + 1}{p^{after} - k + 1} \right)^{2^{k-1}} = \hat{R}_{ratio}(\mathcal{F}_d^{p-p^{rem}}, \theta).
\end{aligned}$$

Thus, we showed that the Rashomon ratio increases as we remove bad features from the dataset.

■

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 dataset 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 datasets with smaller numbers of features induce easier learning/optimization problems in general.

4. Noise as a Theoretical and Practical Motivator for the Existence of Simple-yet-Accurate Models

In this chapter, we study why simpler models are often suitable for noisier datasets from the perspective of the Rashomon Effect. More specifically, we walk along the steps of the path that starts with the uncertainty in the data generation processes and ends with the existence of simple-yet-accurate models. Rather than trying to prove these points for every possible situation (which would be volumes beyond what we can handle here), we aim to find at least some way to illustrate that each step is reasonable in a natural setting.

Overall, learning with noisy labels has been extensively studied (Natarajan et al., 2013), especially for linear regression (Bishop, 1995) and, more recently, for neural networks (Song et al., 2022) to understand and model the effects of noise. Stochastic gradient descent with label noise acts as an implicit regularizer (Damian et al., 2021) and noise has been added to hidden units (Noh et al., 2017), labels (Shallue et al., 2018), or covariances (Wen et al., 2019) to prevent overfitting in deep learning. When the labels are noisy, constructing robust loss (Ghosh et al., 2017), adding a slack variable for each training sample (Hu et al., 2020), or early stopping (M. Li et al., 2020) also helps to improve generalization. However, none of these prior works consider the effect of noise on the Rashomon set or ratio as we do here.

4.1 Increase in Variance due to Noise Leads to Larger Rashomon Ratios

4.1.1 Step 1. Noise Increases Variance

One would think that something as simple as uniform label noise would not really affect anything in the learning process. In fact, we would expect that adding such noise would just uniformly increase the losses of all functions, and the Rashomon set would stay the same. However, this conclusion is (surprisingly) not true. Instead, noise adds variance to the loss, which, in turn, prevents us from generalizing.

For infinite data distribution, $\mathcal{D}$ consider uniform label noise, where each label is flipped independently with probability $\rho < \frac{1}{2}$. If $\tilde{y}$ is a flipped label, $P(\tilde{y} \neq y) = \rho$. If the empirical risk of f is over $\frac{1}{2}$ after adding noise, we transform f to $-f$. For a given model $f \in \mathcal{F}$ let

$\sigma^2(f, \mathcal{D})$ be the variance of the loss, meaning that $\sigma^2(f, \mathcal{D}) = \operatorname{Var}_{z \sim \mathcal{D}} l(f, z)$. We show in the following theorem that, for a given $f \in \mathcal{F}$, label noise increases the variance of the loss.

Theorem 22 (Variance increases with label noise)**.** *Consider infinite true data distribution $\mathcal{D}$, and uniform label noise, where each label is flipped independently with probability ρ. Let $\mathcal{D}_\rho$ denote the noisy version of $\mathcal{D}$. Consider 0-1 loss l, and assume that there exists at least one function $\bar{f} \in \mathcal{F}$ such that $L_{\mathcal{D}}(\bar{f}) < \frac{1}{2} - \gamma$. For a fixed $f \in \mathcal{F}$, let $\sigma^2(f, \mathcal{D}_\rho)$ be the variance of the loss, $\sigma^2(f, \mathcal{D}_\rho) = Var_{z \sim \mathcal{D}_\rho} l(f, z)$ on data distribution $\mathcal{D}_\rho$. For any $0 < \rho_1 < \rho_2 < \frac{1}{2}$,*

$$\sigma^2(f, \mathcal{D}_{\rho_1}) < \sigma^2(f, \mathcal{D}_{\rho_2}).$$

Proof. Recall that the true risk for 0-1 loss $L_{\mathcal{D}}(f) = \mathbb{E}_{z=(x,y)\sim\mathcal{D}}[l(f,z)] = \mathbb{E}_{(x,y)\sim\mathcal{D}}[\mathbb{1}_{[f(x)\neq y]}]$. Without loss of generality, let $y \in \{0, 1\}$. Drawing from $z_\rho \sim \mathcal{D}_\rho$ is equivalent to drawing $z \sim \mathcal{D}$ and changing label y to $1 - y$ with probability ρ. More explicitly, let $\eta \sim \text{Bernoulli}(\rho)$, then the flipped label is $XOR(y, \eta) = \mathbb{1}_{[y \neq \eta]}$. For any given $f \in \mathcal{F}$ we have that:

$$\begin{aligned}
L_{\mathcal{D}_\rho}(f) &= \mathbb{E}_{\eta\sim Ber(\rho)}\,\mathbb{E}_{(x,y)\sim\mathcal{D}}\left[\mathbb{1}_{[f(x)\neq XOR(y,\eta)]}\right] \\
&= \mathbb{E}_{\eta\sim Ber(\rho)}\,\mathbb{E}_{(x,y)\sim\mathcal{D}}\left[\mathbb{1}_{[f(x)\neq y]}(1-\eta)\right] + \mathbb{E}_{\eta\sim Ber(\rho)}\,\mathbb{E}_{(x,y)\sim\mathcal{D}}\left[\mathbb{1}_{[f(x)=y]}\eta\right] \\
&= \mathbb{E}_{(x,y)\sim\mathcal{D}}\,\mathbb{E}_{\eta\sim Ber(\rho)}\left[\mathbb{1}_{[f(x)\neq y]}(1-\eta)\right] + \mathbb{E}_{(x,y)\sim\mathcal{D}}\,\mathbb{E}_{\eta\sim Ber(\rho)}\left[\mathbb{1}_{[f(x)=y]}\eta\right] \\
&= \mathbb{E}_{(x,y)\sim\mathcal{D}}\left[\mathbb{1}_{[f(x)\neq y]}\,\mathbb{E}_{\eta\sim Ber(\rho)}\left[(1-\eta)\right]\right] + \mathbb{E}_{(x,y)\sim\mathcal{D}}\left[\mathbb{1}_{[f(x)=y]}\,\mathbb{E}_{\eta\sim Ber(\rho)}\left[\eta\right]\right] \\
&= (1-\rho)\,\mathbb{E}_{(x,y)\sim\mathcal{D}}\left[\mathbb{1}_{[f(x)\neq y]}\right] + \rho\,\mathbb{E}_{(x,y)\sim\mathcal{D}}\left[\mathbb{1}_{[f(x)=y]}\right] \\
&= (1-\rho)\,\mathbb{E}_{(x,y)\sim\mathcal{D}}\left[\mathbb{1}_{[f(x)\neq y]}\right] + \rho\left(1-\mathbb{E}_{(x,y)\sim\mathcal{D}}\left[\mathbb{1}_{[f(x)\neq y]}\right]\right) \\
&= (1-\rho)L_{\mathcal{D}}(f) + \rho\,(1-L_{\mathcal{D}}(f)) \\
&= (1-2\rho)L_{\mathcal{D}}(f) + \rho.
\end{aligned}$$

Note, following the technique above, a similar statement is true about dataset S instead of true distribution $\mathcal{D}$, meaning that for a given $f \in \mathcal{F}$,

$$\mathbb{E}_{S_\rho}\hat{L}_{S_\rho}(f) = (1-2\rho)\hat{L}_S(f) + \rho. \tag{4.1}$$

Recall that we take expectation with respect to different ways of adding noise to labels, therefore S_ρ and S have the same x, but different y. We do not use (4.1) for the proof of Theorem 22, but use it for proof of Theorem 36.

For true distribution $\mathcal{D}$, since l is 0-1 loss, then for a given model f, $l(f,z)$ is Bernoulli distributed with mean $p_{Ber} = \mathbb{E}_{z\sim\mathcal{D}}\, l(f,z) = L_\mathcal{D}(f)$ and variance $\sigma_f^2 = p_{Ber}(1-p_{Ber}) = L_\mathcal{D}(f)(1-L_\mathcal{D}(f))$. Therefore, the expected variance for a given model $f \in R_{set}(\mathcal{F},\gamma)$ on distribution $\mathcal{D}_\rho$ is:

$$\begin{aligned}
Var_{z\sim\mathcal{D}_\rho}\left[l(f,z)\right] &= \mathbb{E}_{\mathcal{D}_\rho} L_{\mathcal{D}_\rho}(f)(1-L_{\mathcal{D}_\rho}(f)) \\
&= \mathbb{E}_{\mathcal{D}_\rho} L_{\mathcal{D}_\rho}(f) - \mathbb{E}_{\mathcal{D}_\rho}(L_{\mathcal{D}_\rho}(f))^2 \\
&= \mathbb{E}_{\mathcal{D}_\rho} L_{\mathcal{D}_\rho}(f) - \mathbb{E}_{\mathcal{D}_\rho}(L_{\mathcal{D}_\rho}(f))^2 \\
&= \mathbb{E}_{\mathcal{D}_\rho} L_{\mathcal{D}_\rho}(f) - (\mathbb{E}_{\mathcal{D}_\rho} L_{\mathcal{D}_\rho}(f))^2 - Var_{\mathcal{D}_\rho}[L_{\mathcal{D}_{\rho(f)}}] \\
&= L_\mathcal{D}(f)(1-2\rho) + \rho - \left(L_\mathcal{D}(f)(1-2\rho)+\rho\right)^2 - Var_{\mathcal{D}_\rho}[L_{\mathcal{D}_{\rho(f)}}] \\
&= L_\mathcal{D}(f)\left((1-2\rho) - 2\rho(1-2\rho)\right) - L^2_\mathcal{D}(f)(1-2\rho)^2 + \rho - \rho^2 - Var_{\mathcal{D}_\rho}[L_{\mathcal{D}_{\rho(f)}}] \\
&= (1-2\rho)^2(L_\mathcal{D}(f) - L^2_\mathcal{D}(f)) + \rho - \rho^2 - Var_{\mathcal{D}_\rho}[L_{\mathcal{D}_{\rho(f)}}] \\
&= (1-2\rho)^2\left(L_\mathcal{D}(f)(1-L_\mathcal{D}(f))\right) + \rho(1-\rho) - Var_{\mathcal{D}_\rho}[L_{\mathcal{D}_{\rho(f)}}] \\
&= (1-2\rho)^2\left(L_\mathcal{D}(f)(1-L_\mathcal{D}(f))\right) + \rho(1-\rho),
\end{aligned}$$

Note that, by our assumption, there exists $\bar{f}$ such that $L_\mathcal{D}(\bar{f}) < \frac{1}{2}-\gamma$, so $L_\mathcal{D}(f^*) < \frac{1}{2}-\gamma$, where f^* is optimal model. Then for any fixed $f \in \mathcal{F}$, we get $L_\mathcal{D}(f) \leq L_\mathcal{D}(f^*) + \gamma < \frac{1}{2}$ which implies that $L_\mathcal{D}(f)(1-L_\mathcal{D}(f)) < \frac{1}{4}$.

For $\rho \in (0, \frac{1}{2})$, $Var_{z\sim\mathcal{D}_\rho}\left[l(f,z)\right]$ is monotonically increasing in ρ, since:

$$\begin{aligned}
\frac{\partial}{\partial\rho}\left[Var_{z\sim\mathcal{D}_\rho}\left[l(f,z)\right]\right] &= \frac{\partial}{\partial\rho}\left[(1-2\rho)^2\left(L_\mathcal{D}(f)(1-L_\mathcal{D}(f))\right) + \rho(1-\rho)\right] \\
&= -4(1-2\rho)\left(L_\mathcal{D}(f)(1-L_\mathcal{D}(f))\right) + (1-2\rho) \\
&= (1-2\rho)\left(1-4L_\mathcal{D}(f)(1-L_\mathcal{D}(f))\right) \\
&> \left(1-2\times\frac{1}{2}\right)\left(1-4\times\frac{1}{4}\right) = 0.
\end{aligned}$$

Consider $\rho_1 < \rho_2$. Since $Var_{z \sim \mathcal{D}_\rho}\left[l(f,z)\right]$ is monotonically increasing in ρ for a fixed f, then $\sigma^2(f, \mathcal{D}_{\rho_1}) < \sigma^2(f, \mathcal{D}_{\rho_2})$, and we proved that variance increases with random uniform label noise.

■

This covers the uniform noise case, but variance increases more generally, and we prove this for several other common cases. More specifically, we show that the variance increases with other types of label noise, such as non-uniform label noise (see Theorem 23) and margin noise (see Theorem 26). For non-uniform label noise, for a sample $z = (x, y)$, each label y is flipped independently with probability ρ_x, meaning that noise can depend on x. This noise model is more realistic than uniform label noise and allows the modeling of cases when one sub-population has much more noise than another. The variance of the loss increases under non-uniform label noise as we show below:

Theorem 23 (Variance increases with non-uniform label noise)**.** *Consider 0-1 loss l, infinite true data distribution $\mathcal{D}$, and a hypothesis space $\mathcal{F}$. Assume that there exists at least one function $\bar{f} \in \mathcal{F}$ such that $L_{\mathcal{D}}(\bar{f}) < \frac{1}{2} - \gamma$. For a fixed $f \in \mathcal{F}$, let $\sigma^2(f, \mathcal{D})$ be the variance of the loss: $\sigma^2(f, \mathcal{D}) = Var_{z \sim \mathcal{D}} l(f, z)$ on data distribution $\mathcal{D}$. Consider non-uniform label noise, where each label y is flipped independently with probability ρ_x, $(x, y) \sim \mathcal{D}$. Let $\mathcal{D}_\rho$ denote the noisy version of $\mathcal{D}$. For any $\delta > 0$, let $\mathcal{D}_{\rho^\delta}$ be a noisier data distribution than $\mathcal{D}_\rho$, meaning that for every sample (x, y) the probabilities of labels being flipped are higher by δ: $\rho_x^\delta = \rho_x + \delta$. If for a fixed model $f \in \mathcal{F}$, $L_{\mathcal{D}_{\rho^\delta}}(f) < 0.5$, then*

$$\sigma^2(f, \mathcal{D}_\rho) < \sigma^2(f, \mathcal{D}_{\rho^\delta}).$$

Proof. Recall that the true risk for 0-1 loss $L_{\mathcal{D}}(f) = \mathbb{E}_{z=(x,y) \sim \mathcal{D}}[l(f, z)] = \mathbb{E}_{(x,y) \sim \mathcal{D}}[\mathbb{1}_{[f(x) \neq y]}]$. Without loss of generality, let $y \in \{0, 1\}$. Drawing from $z_\rho \sim \mathcal{D}_\rho$ is equivalent to drawing $z \sim \mathcal{D}$ and changing label y to $1-y$ with probability ρ_x. More explicitly, let $\eta \sim \text{Bernoulli}(\rho_x)$, then the flipped label is $XOR(y, \eta) = \mathbb{1}_{[y \neq \eta]}$. For any given $f \in \mathcal{F}$ we have that:

$L_{\mathcal{D}_\rho}(f) = \mathbb{E}_{(x,y) \sim \mathcal{D}}\, \mathbb{E}_{\eta \sim Ber(\rho_x)}\left[\mathbb{1}_{[f(x) \neq XOR(y,\eta)]}\right]$

$$
\begin{aligned}
&= \mathbb{E}_{(x,y)\sim\mathcal{D}}\, \mathbb{E}_{\eta\sim Ber(\rho_x)} \left[\mathbb{1}_{[f(x)\neq y]}(1-\eta)\right] + \mathbb{E}_{(x,y)\sim\mathcal{D}}\, \mathbb{E}_{\eta\sim Ber(\rho_x)} \left[\mathbb{1}_{[f(x)=y]}\eta\right] \\
&= \mathbb{E}_{(x,y)\sim\mathcal{D}} \left[\mathbb{1}_{[f(x)\neq y]}\, \mathbb{E}_{\eta\sim Ber(\rho_x)} \left[(1-\eta)\right]\right] + \mathbb{E}_{(x,y)\sim\mathcal{D}} \left[\mathbb{1}_{[f(x)=y]}\, \mathbb{E}_{\eta\sim Ber(\rho_x)} \left[\eta\right]\right] \\
&= \mathbb{E}_{(x,y)\sim\mathcal{D}} \left[\mathbb{1}_{[f(x)\neq y]}\, \mathbb{E}_{\eta\sim Ber(\rho_x)} \left[(1-\eta)\right]\right] + \mathbb{E}_{(x,y)\sim\mathcal{D}} \left[\left(1-\mathbb{1}_{[f(x)\neq y]}\right) \mathbb{E}_{\eta\sim Ber(\rho_x)} \left[\eta\right]\right] \\
&= \mathbb{E}_{(x,y)\sim\mathcal{D}} \left[\mathbb{1}_{[f(x)\neq y]}\, \mathbb{E}_{\eta\sim Ber(\rho_x)} \left[(1-\eta)\right]\right] + \mathbb{E}_{(x,y)\sim\mathcal{D}} \left[\mathbb{E}_{\eta\sim Ber(\rho_x)} \left[\eta\right] - \mathbb{1}_{[f(x)\neq y]}\, \mathbb{E}_{\eta\sim Ber(\rho_x)} \left[\eta\right]\right] \\
&= \mathbb{E}_{(x,y)\sim\mathcal{D}} \left[\mathbb{1}_{[f(x)\neq y]}\, \mathbb{E}_{\eta\sim Ber(\rho_x)} \left[(1-2\eta)\right]\right] + \mathbb{E}_{(x,y)\sim\mathcal{D}} \left[\mathbb{E}_{\eta\sim Ber(\rho_x)} \left[\eta\right]\right] \\
&= \mathbb{E}_{(x,y)\sim\mathcal{D}} \left[\mathbb{1}_{[f(x)\neq y]} \left(1-2\rho_x\right)\right] + \mathbb{E}_{(x,y)\sim\mathcal{D}}\, \rho_x.
\end{aligned}
$$

Now we will show that $L_{\mathcal{D}_{\rho^\delta}}(f) > L_{\mathcal{D}_\rho}(f)$:

$$
\begin{aligned}
L_{\mathcal{D}_{\rho^\delta}}(f) - L_{\mathcal{D}_\rho}(f) &= \mathbb{E}_{(x,y)\sim\mathcal{D}} \left[\mathbb{1}_{[f(x)\neq y]} \left(1-2\rho_x^\delta\right)\right] + \mathbb{E}_{(x,y)\sim\mathcal{D}}\, \rho_x^\delta \\
&\quad - \mathbb{E}_{(x,y)\sim\mathcal{D}} \left[\mathbb{1}_{[f(x)\neq y]} \left(1-2\rho_x\right)\right] + \mathbb{E}_{(x,y)\sim\mathcal{D}}\, \rho_x \\
&= \mathbb{E}_{(x,y)\sim\mathcal{D}} \left[\mathbb{1}_{[f(x)\neq y]} \left(-2\rho_x^\delta + 2\rho_x\right)\right] + \mathbb{E}_{(x,y)\sim\mathcal{D}} \left(\rho_x^\delta - \rho_x\right) \\
&= \mathbb{E}_{(x,y)\sim\mathcal{D}} \left[\mathbb{1}_{[f(x)\neq y]} \left(-2\delta\right)\right] + \mathbb{E}_{(x,y)\sim\mathcal{D}} \left(\delta\right) \\
&= \left(-2\delta\right) \mathbb{E}_{(x,y)\sim\mathcal{D}} \left[\mathbb{1}_{[f(x)\neq y]}\right] + \delta \\
&= \delta(1-2L_{\mathcal{D}}(f)) > 0.
\end{aligned}
$$

Note that, by our assumption, there exists $\bar{f}$ such that $L_{\mathcal{D}}(\bar{f}) < \frac{1}{2}-\gamma$, so $L_{\mathcal{D}}(f^*) < \frac{1}{2}-\gamma$, where f^* is an optimal model. Then for any fixed $f \in R_{set}(\mathcal{F},\gamma)$, we get $L_{\mathcal{D}}(f) \leq L_{\mathcal{D}}(f^*) + \gamma < \frac{1}{2}$, and then $1 - 2L_{\mathcal{D}}(f) > 0$. Since $\delta > 0$, we have shown that $L_{\mathcal{D}_{\rho^\delta}}(f) > L_{\mathcal{D}_\rho}(f)$.

For true distribution $\mathcal{D}$, since l is 0-1 loss, then for a given model f, $l(f,z)$ is Bernoulli distributed with mean $p_{Ber} = \mathbb{E}_{z\sim\mathcal{D}}\, l(f,z) = L_{\mathcal{D}}(f)$ and variance $\sigma_f^2 = p_{Ber}(1-p_{Ber}) = L_{\mathcal{D}}(f)(1-L_{\mathcal{D}}(f))$. Therefore, the expected variance for a given model $f \in \mathcal{F}$ on distributions $\mathcal{D}_\rho$ and $\mathcal{D}_{\rho^\delta}$ is:

$$
\begin{aligned}
Var_{z\sim\mathcal{D}_\rho}\left[l(f,z)\right] &= L_{\mathcal{D}_\rho}(f)(1-L_{\mathcal{D}_\rho}(f)) \\
&< L_{\mathcal{D}_{\rho^\delta}}(f)(1-L_{\mathcal{D}_{\rho^\delta}}(f)) \\
&= Var_{z\sim\mathcal{D}_{\rho^\delta}}\left[l(f,z)\right],
\end{aligned}
$$

where the inequality arises from the fact that the parabola $x(1-x)$ is monotonic along the interval $x \in [0, 0.5]$. This implies that $\sigma^2(f, \mathcal{D}_\rho) < \sigma^2(f, \mathcal{D}_{\rho^\delta})$.

■

We model margin noise such as that which arises from high-dimensional Gaussians by moving two Gaussians closer together along the vector that connects the two means (as in Figure 4.2(a)). Because of the central limit theorem, data often follow Gaussian distributions, therefore this noise is realistic and models mistakes near the decision boundary. Before stating and proving the theorem we discuss two lemmas (Lemma 24 and Lemma 25) that are helpful for the proof.

Lemma 24. *Consider distribution $\mathcal{X} \in \mathbb{R}^p$ and a linear model $f = \omega^T x + b$, where $\omega \in \mathbb{R}^p$, $\omega \neq \bar{0}$ and $b \in \mathbb{R}$. Let $x \mapsto Ax + c$ be a bijective affine transformation, where $A \in \mathbb{R}^{p \times p}$ and $c \in \mathbb{R}^p$. For the linear model $g(x) = f(A^{-1}(x-c))$ and the distribution $\mathcal{Z} = A\mathcal{X} + c$, we have that:*

$$P_{x \sim \mathcal{X}}(f(x) > 0) = P_{z \sim \mathcal{Z}}(g(z) > 0).$$

Proof. The proof follows from the lemma's statement and the assumption that A is a bijective affine transformation, and thus is invertible:

$$P_{z \sim \mathcal{Z}}(g(z) > 0) = P_{x \sim \mathcal{X}}(g(Ax+c) > 0) = P_{x \sim \mathcal{X}}(f(A^{-1}(Ax+c-c)) > 0) = P_{x \sim \mathcal{X}}(f(x) > 0).$$

■

Lemma 25. *Consider a Gaussian distribution $\mathcal{X} \sim \mathcal{N}(\mu, I)$, where $\mu \in \mathbb{R}^p$, and a linear model $f = \omega^T x + b$, where $\omega \in \mathbb{R}^p$, $\omega \neq \bar{0}$ and $b \in \mathbb{R}$. Let $r = \frac{\omega^T \mu + b}{\|\omega\|}$ be the signed distance from μ to the decision boundary of f. Then,*

$$P_{x \sim \mathcal{X}}(f(x) > 0) = \Phi(r),$$

where Φ is the CDF of the univariate normal distribution $\mathcal{N}(0, 1)$.

Proof. Let $O \in \mathbb{R}^{p \times p}$ be a matrix with the first row equal to $\frac{\omega}{\|\omega\|}$, and let the other rows be chosen so that the rows of O form an orthonormal basis of $\mathbb{R}^p$. Note that O is an orthogonal matrix, so O is bijective and $O^T O = O O^T = I$. Let $g(t) = f(O^{-1}(t + O\mu))$ and e_1 be a unit vector $e_1 = \{1, 0, ..., 0\}$, then:

$$\begin{aligned} g(t) &= f(O^{-1}(t + O\mu)) = f(O^{-1}t + \mu) = f(O^T t + \mu) \\ &= \omega^T O^T t + \omega^T \mu + b = \|\omega\| \, (e_1^T t) + \omega^T \mu + b \\ &= \|\omega\| \, t_1 + \omega^T \mu + b, \end{aligned}$$

where t_1 is the first element of t, and $\omega^T O^T = \|\omega\| \, e_1^T$ comes from the fact that ω is orthogonal to every row of O except for the first row. Note that $g(t) > 0$ when $\|\omega\| \, t_1 + \omega^T \mu + b > 0$, which leads to $t_1 > -\frac{\omega^T \mu + b}{\|\omega\|} = -r$. Correspondingly, $g(t) < 0$ when $t_1 < -r$.

Now, let $\mathcal{Z} = O(\mathcal{X} - \mu)$. From the properties of the normal distribution, $\mathcal{Z} \sim \mathcal{N}(\bar{0}, I)$ since:

$$\mathcal{Z} = O(\mathcal{X} - \mu) \sim \mathcal{N}(O(\mu - \mu), OIO^T) = \mathcal{N}(\bar{0}, I).$$

Moreover, since the standard multivariate normal distribution is the joint distribution of independent univariate normal distributions, $z_1 \sim \mathcal{N}(0, 1)$.

From Lemma 24 and definitions of O, g, $\mathcal{Z}$, we get that $P_{x \sim \mathcal{X}}(f(x) > 0) = P_{z \sim \mathcal{Z}}(g(z) > 0)$. Therefore:

$$\begin{aligned} P_{x \sim \mathcal{X}}(f(x) > 0) &= P_{z \sim \mathcal{Z}}(g(z) > 0) = P_{z \sim \mathcal{Z}}(z_1 > -r) \\ &= P_{z_1 \sim \mathcal{N}(0,1)}(z_1 > -r) = P_{z_1 \sim \mathcal{N}(0,1)}(z_1 \leq r) \\ &= \Phi(r), \end{aligned}$$

where the strict inequality becomes non-strict since for the Gaussian distribution, the probability $P_{z_1 \sim \mathcal{N}(0,1)}(z_i = r) = 0$. Thus, $P_{x \sim \mathcal{X}}(f(x) > 0) = \Phi(r)$ as desired. ■

Now, we show that the variance of losses increases under margin noise in the theorem below:

Theorem 26. *Consider data distribution $\mathcal{D} = \mathcal{X} \times \mathcal{Y}$, where, $\mathcal{X} \in \mathbb{R}^p$, $\mathcal{Y} \in \{-1, 1\}$, classes are balanced $P(Y = -1) = P(Y = 1)$ and generated by Gaussian distributions $P(X|Y = -1) = \mathcal{N}(\bar{0}, \Sigma)$, $P(X|Y = 1) = \mathcal{N}(\mu, \Sigma)$, where Σ is a diagonal matrix with non-zero elements. Let the hypothesis space $\mathcal{F}$ be the set of linear models, $f = \omega^T x + b$, where $\omega \in \mathbb{R}^p$. $\omega \neq 0$ and $b \in \mathbb{R}$. We add margin noise by moving the means of the Gaussians towards each other by a factor of k, where $0 < k < 1$, meaning that the mean of the positive class becomes $\mu_k = k \cdot \mu$. For a fixed $f \in \mathcal{F}$, if $L_\mu(f) < 0.5$, we get that the variance of losses increases with more noise,*

$$\sigma(f, \mu) < \sigma(f, \mu_k).$$

Proof. Without loss of generality, we will show that the variance of the losses increases for data generated from two Gaussian distributions $P(X|Y = -1) = \mathcal{N}(\bar{0}, I)$ and $P(X|Y = 1) = \mathcal{N}(\mu, I)$ (where I is the identity matrix) when we move them towards each other. More specifically, since normalization by variance $\left(\frac{1}{\Sigma_{i,i}}\right)$ is a bijective linear transformation, by Lemma 24 we can work with $P(X|Y = -1) = \mathcal{N}(\bar{0}, I)$ and $P(X|Y = 1) = \mathcal{N}(\mu, I)$ instead of $P(X|Y = -1) = \mathcal{N}(\bar{0}, \Sigma)$ and $P(X|Y = 1) = \mathcal{N}(\mu, \Sigma)$.

Let $r_1 = \frac{b}{\|\omega\|}$ and $r_2 = \frac{\omega^T \mu + b}{\|\omega\|}$ be the signed distances from the centers of the two Gaussians to the decision boundary. Then, from Lemma 25 (see illustration in Figure 4.1), the loss can be computed using the CDFs based on the signed distance:

$$\begin{aligned} L_\mu(f) &= P(f(x) > 0|Y = -1)P(Y = -1) + P(f(x) \leq 0|Y = 1)P(Y = 1) \\ &= \frac{1}{2} P(f(x) > 0|Y = -1) + \frac{1}{2} \left(1 - P(f(x) > 0|Y = 1)\right) \\ &= \frac{1}{2} \left[(\Phi(r_1) + (1 - \Phi(r_2))\right]. \end{aligned}$$

Next, we will show that $\omega^T \mu > 0$. If $L_\mu(f) < \frac{1}{2}$, then we get that $\frac{1}{2}[(\Phi(r_1) + (1 - \Phi(r_2))] < \frac{1}{2}$, which means that $\Phi(r_2) > \Phi(r_1)$. Since the CDF of the Gaussian distribution $\mathcal{N}(\bar{0}, I)$ is strictly increasing, we have that $r_2 > r_1$, which means that $\frac{\omega^T \mu + b}{\|\omega\|} > \frac{b}{\|\omega\|}$, and so $\omega^T \mu > 0$.

Recall that we induce noise by moving the Gaussians towards each other by decreasing

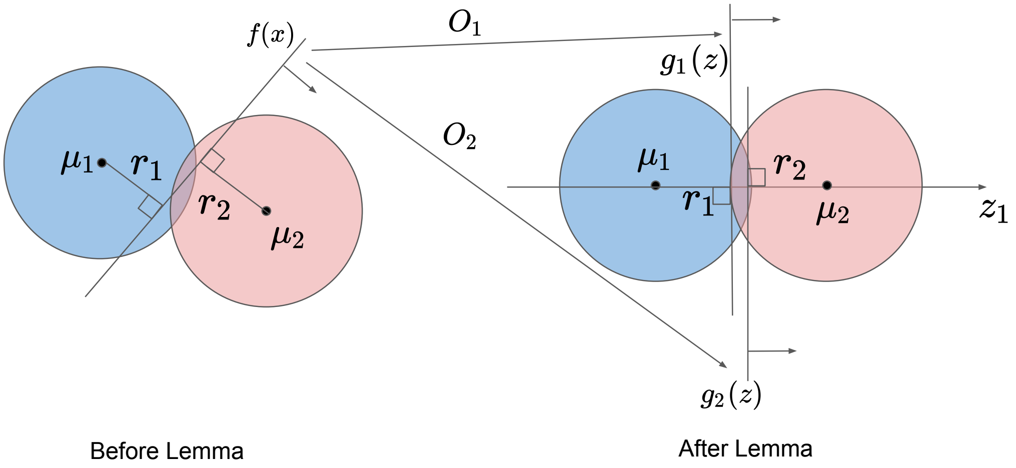


FIGURE 4.1: An illustration of how Lemma 25 rotates each of the Gaussians $\mathcal{N}(\mu_1, I)$, $\mathcal{N}(\mu_2, I)$ and the decision boundary $f(x)$ in order to compute loss as CDF of the signed distances (r_1, r_2) from means (μ_1, μ_2) to the rotated boundaries $(g_1(z), g_2(z))$. Note that we apply Lemma 25 separately to each Gaussian, thus there are two rotation operators O_1, and O_2.

k. Now we will show that loss is monotonically decreasing with respect to increasing values of k, or equivalently that $\frac{\partial}{\partial k} L_{\mu_k}(f) < 0$:

$$\begin{aligned}\frac{\partial}{\partial k} L_{\mu_k}(f) &= \frac{\partial}{\partial k}\left(\frac{1}{2}\left[(\Phi(r_1) + (1 - \Phi(r_2))\right]\right) \\ &= \frac{\partial}{\partial k}\left(\frac{1}{2}\left[(\Phi\left(\frac{b}{\|\omega\|}\right) + 1 - \Phi\left(\frac{k\omega^T\mu + b}{\|\omega\|}\right)\right]\right) \\ &= -\frac{1}{2}\left[\frac{\partial}{\partial k}\Phi\left(\frac{k\omega^T\mu + b}{\|\omega\|}\right)\right] = -\frac{1}{2}\frac{\omega^T\mu}{\|\omega\|}\phi\left(\frac{k\omega^T\mu + b}{\|\omega\|}\right) < 0,\end{aligned}$$

since as we showed above, $\omega^T\mu > 0$, and ϕ is the PDF of normal distribution $\mathcal{N}(\bar{0}, I)$ which is always positive. Therefore, $L_{\mu_k}(f)$ is monotonically decreasing with respect to k, and we have that $L_\mu(f) < L_{\mu_k}(f)$ for all $0 < k < 1$.

For the true distribution $\mathcal{D}$, since l is 0-1 loss, then for a given model f, $l(f, z)$ is Bernoulli distributed with mean $p_{Ber} = \mathbb{E}_{z\sim\mathcal{D}}\, l(f, z) = L_{\mathcal{D}}(f)$ and variance $\sigma_f^2 = p_{Ber}(1 - p_{Ber}) = L_{\mathcal{D}}(f)(1 - L_{\mathcal{D}}(f))$. Therefore, the expected variance for a given model $f \in R_{set}(\mathcal{F}, \gamma)$ on distributions $\mathcal{D}_\mu$ and $\mathcal{D}_{\mu_k}$ obeys:

$$\sigma^2(f, \mu) = L_\mu(f)(1 - L_\mu(f))$$

$$< L_{\mu_k}(f)(1 - L_{\mu_k}(f))$$
$$= \sigma^2(f, \mu_k),$$

where the inequality arises from the fact that the parabola $x(1-x)$ is monotonically increasing along the interval $x \in [0, 0.5]$, and $\mu_k = k\mu$ is closer to $\bar{0}$ than μ. ■

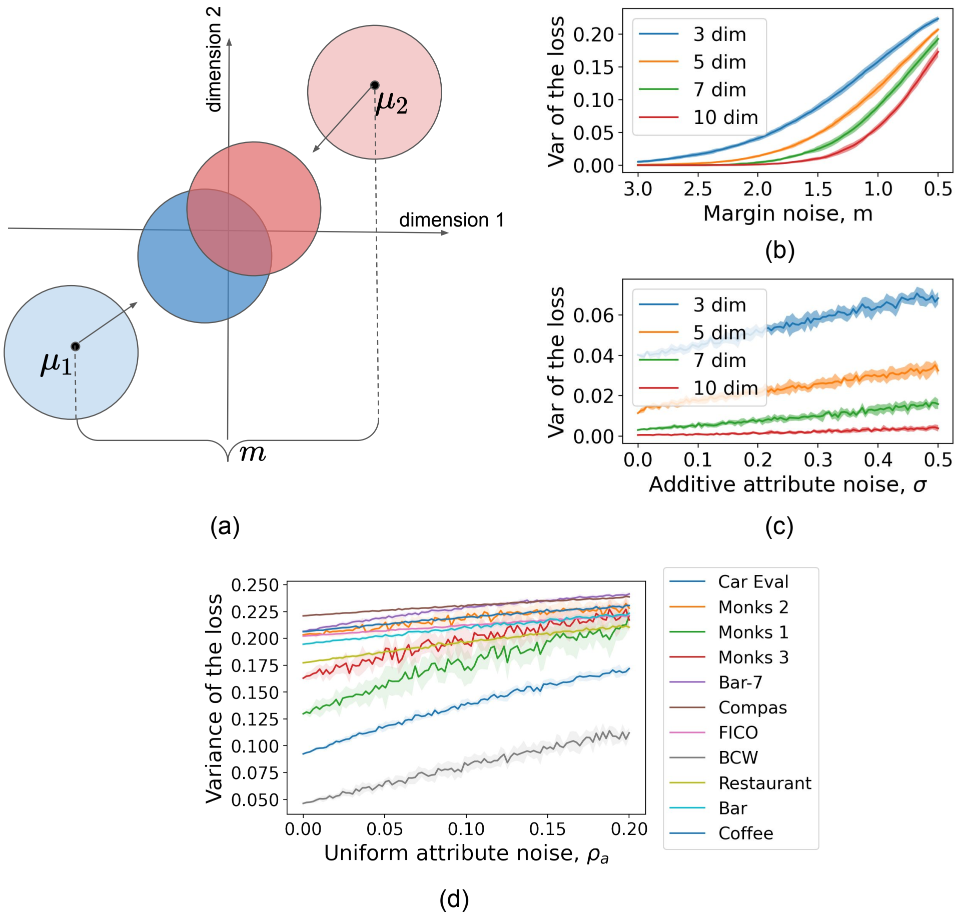
Note, that we can generalize Theorem 26 to the case when Σ is any positive-definite matrix that is not necessarily diagonal (covariance matrices are always positive semi-definite, and we now additionally assume that Σ does not have zero eigenvalues). Since Σ is real and symmetric, by the spectral theorem, there exists an orthogonal matrix $Q \in \mathbb{R}^{p \times p}$ such that $D = Q\Sigma Q^T$ where D is diagonal and contains eigenvalues of Σ. The diagonal elements of D must be real and positive since Σ is positive-definite. Then, consider the data distribution $(Q\mathcal{X}) \times \mathcal{Y}$. From the properties of the Gaussian distribution, $Q\mathcal{X}$ is Gaussian with mean $Q\mu$ and covariance matrix $Q\mathcal{X}Q^T = D$. Thus, we can generalize the results of Theorem 26 to apply to positive-definite non-diagonal matrices Σ.

For a fixed model, we additionally verify the results of Theorem 26 empirically, by generating Gaussian distributions and introducing margin noise by moving the Gaussians closer together (see Figure 4.2(b).) The variance of losses increases with additive and uniform random attribute noise as well, as we show empirically in Figure 4.2(c)-(d).

Label noise in datasets is common. In fact, real-world datasets reportedly have between 8.0% and 38.5% label noise (K.-H. Lee et al., 2018; W. Li et al., 2017; Song et al., 2019, 2022; Xiao et al., 2015). We hypothesize that a significant amount of label noise in real-world datasets is a combination of Gaussian (due to the central limit theorem) and random noise (for example, because of clerical errors causing label noise).

For the true Rashomon set $R_{set}(\mathcal{F}, \gamma)$, we consider the maximum variance for all models in the true set: $\sigma^2 = \sup_{f \in R_{set}(\mathcal{F},\gamma)} \mathrm{Var}_{z\sim\mathcal{D}}\, l(f, z)$. Then, from Theorem 22 we have that the maximum expected variance over the Rashomon set increases with noise.

Corollary 27 (Maximum variance increases with label noise)**.** *Under the same assumptions*



(a)

(b)

(c)

(d)

FIGURE 4.2: The variance of losses increases with margin (b) and additive attribute (c, d) noise. For (b) and (c) we generated data from Gaussians in 3, 5, 7, and 10 dimensions. For margin noise (b), as illustrated in (a), the negative class is generated from $\mathcal{N}(\bar{\mu_1}, I)$ and positive from $\mathcal{N}(\bar{\mu_2}, I)$, where I is the identity matrix, $\bar{\mu_1} = -m/2 \times \bar{1}$, $\bar{\mu_2} = m/2 \times \bar{1}$, and m controls the distance between Gaussians that determines the amount of margin noise. For additive noise, data is generated from $\mathcal{N}(\bar{0}, I)$ and $\mathcal{N}(\bar{2}, I)$. The noise model is $x' = x + \epsilon$, where $\epsilon \sim \mathcal{N}(\bar{0}, \sigma I)$ is the noise vector added to every sample and σ determines how much noise is added to the data. For evaluation, as a fixed model we consider a random linear model from the Rashomon set. For (d), we chose 3 features with the highest AUC value and introduced uniform noise by negating the attribute values with probability ρ_a. As a fixed model, we consider a tree generated by the CART algorithm that uses at least one of the features to which noise was applied (this is because if the model does not use these features, the variance of losses for that model will not change). All plots are based on 0-1 loss and are averaged over 10 iterations.

as in Theorem 22, we have that

$$\sup_{f \in R_{set_{\mathcal{D}_{\rho_1}}}(\mathcal{F},\gamma)} \sigma^2(f, \mathcal{D}_{\rho_1}) < \sup_{f \in R_{set_{\mathcal{D}_{\rho_2}}}(\mathcal{F},\gamma)} \sigma^2(f, \mathcal{D}_{\rho_2}).$$

Proof. Let f_1^{sup} and f_2^{sup} be maximizers of the variance of the loss in their respective Rashomon sets:

$$f_1^{\text{sup}} \in \arg \sup_{f \in R_{set_{\mathcal{D}_{\rho_1}}}(\mathcal{F},\gamma)} Var_{z\sim\mathcal{D}_{\rho_1}}\left[l(f,z)\right],$$

$$f_2^{\text{sup}} \in \arg \sup_{f \in R_{set_{\mathcal{D}_{\rho_2}}}(\mathcal{F},\gamma)} Var_{z\sim\mathcal{D}_{\rho_2}}\left[l(f,z)\right].$$

Given that for any $f \in R_{set_{\mathcal{D}_{\rho_2}}}(\mathcal{F},\gamma)$, $Var_{z\sim\mathcal{D}_{\rho_2}}\left[l(f,z)\right] \leq Var_{z\sim\mathcal{D}_{\rho_2}}\left[l(f_2^{\text{sup}},z)\right]$ and since $Var_{z\sim\mathcal{D}_{\rho}}\left[l(f,z)\right]$ is monotonically increasing in ρ, we have that:

$$\begin{aligned}\sup_{f \in R_{set_{\mathcal{D}_{\rho_1}}}(\mathcal{F},\gamma)} \sigma^2(f,\mathcal{D}_{\rho_1}) &= Var_{z\sim\mathcal{D}_{\rho_1}}\left[l(f_1^{\text{sup}},z)\right]\\ &< Var_{z\sim\mathcal{D}_{\rho_2}}\left[l(f_1^{\text{sup}},z)\right]\\ &\leq Var_{z\sim\mathcal{D}_{\rho_2}}\left[l(f_2^{\text{sup}},z)\right]\\ &= \sup_{f \in R_{set_{\mathcal{D}_{\rho_2}}}(\mathcal{F},\gamma)} \sigma^2(f,\mathcal{D}_{\rho_2}).\end{aligned}$$

■

While the results of Theorems 23, 26 are for a given and fixed model f, they hold for the f that achieves the maximum variance in the Rashomon set as well, meaning that Corollary 27 extends to Theorems 23, 26. The next step is to show that this increased maximum variance leads to worse generalization.

4.1.2 Step 2. Higher Variance Leads to Worse Generalization

Here we use an argument based on generalization bounds. Generalization bounds have been the key theoretical motivation for much of machine learning, including support vector machines (SVMs), because the margin that SVMs optimize appears in a bound. While bounds themselves are not directly used in practice, the terms in the bounds tend to be important quantities in practice. Our bound cannot be calculated in practice because it uses population information on the right side, but it still provides insight and motivation.

Unlike standard bounds, we will use the fact that the user is using empirical risk minimization, and cross-validation to assess overfitting. Thus, for $\hat{f}$, there are two possibilities: $\hat{f}$ is in the true Rashomon set, or it is not. If it is not, then for the empirical risk minimizer $\hat{f}$, the difference between the true and the empirical risk must be at least γ, which will be detected with high probability in cross-validation (Kearns, 1995; Mukherjee et al., 2006). In that case, the user will reduce their hypothesis space and we move to Step 3. If $\hat{f}$ is in the true Rashomon set, it obeys the following bound.

Theorem 28 (Variance-based "generalization bound")**.** *Consider dataset S, 0-1 loss l, and finite hypothesis space $\mathcal{F}$. With probability at least $1-\delta$, we have that for every $f \in R_{set}(\mathcal{F},\gamma)$:*

$$L(f) - \hat{L}(f) \leq \frac{2}{3n}\log\left(\frac{|R_{set}(\mathcal{F},\gamma)|}{\delta}\right) + \sqrt{\frac{2\sigma^2}{n}\log\left(\frac{|R_{set}(\mathcal{F},\gamma)|}{\delta}\right)}, \tag{4.2}$$

where $\sigma^2 = \sup_{f\in R_{set}(\mathcal{F},\gamma)} \mathrm{Var}_{z\sim\mathcal{D}}\, l(f,z)$, and n is number of samples in $S = \{z_i\}_{i=1}^n \sim \mathcal{D}$.

Proof. For each fixed model $f \in R_{set}(\mathcal{F},\gamma)$ in the true Rashomon set, from Bernstein's inequality, using that the maximum value for the 0-1 loss is 1, we have that

$$P(L(f) - \hat{L}(f) > \varepsilon) \leq e^{\frac{-n\varepsilon^2}{2\sigma_f^2 + 2\varepsilon/3}}.$$

According to the union bound:

$$\begin{aligned}
P\left(\exists f \in R_{set}(\mathcal{F},\gamma) : L(f) - \hat{L}(f) > \varepsilon\right) &\leq \sum_{f\in R_{set}(\mathcal{F},\gamma)} P\left(L(f) - \hat{L}(f) > \varepsilon\right) \\
&\leq \sum_{f\in R_{set}(\mathcal{F},\gamma)} e^{\frac{-n\varepsilon^2}{2\sigma_f^2 + 2\varepsilon/3}} \\
&\leq \sum_{f\in R_{set}(\mathcal{F},\gamma)} e^{\frac{-n\varepsilon^2}{2\sigma^2 + 2\varepsilon/3}} \\
&= |R_{set}(\mathcal{F},\gamma)| \cdot e^{\frac{-n\varepsilon^2}{2\sigma^2 + 2\varepsilon/3}},
\end{aligned}$$

where we used the fact that $e^{-\frac{1}{\sigma_f^2}} \leq e^{-\frac{1}{\sup_{f \in R_{set}(\mathcal{F},\gamma)} \sigma_f^2}} = e^{-\frac{1}{\sigma^2}}$, since the exponential function is monotonic.

Let $\delta = |R_{set}(\mathcal{F},\gamma)| \, e^{\frac{-n\varepsilon^2}{2\sigma^2 + 2\varepsilon/3}}$, then we have the following quadratic equation to find ε:

$$\varepsilon^2 - \frac{2}{3n} \log\left(\frac{|R_{set}(\mathcal{F},\gamma)|}{\delta}\right) \varepsilon - \frac{2\sigma^2}{n} \log\left(\frac{|R_{set}(\mathcal{F},\gamma)|}{\delta}\right) = 0.$$

Setting $a = \frac{2}{n} \log\left(\frac{|R_{set}(\mathcal{F},\gamma)|}{\delta}\right)$, we find that the roots of the quadratic equation with respect to ε are:

$$\varepsilon = \frac{a}{2 \cdot 3} \pm \frac{1}{2} \sqrt{\left(\frac{a}{3}\right)^2 + 4a\sigma^2}.$$

Since $4a\sigma^2 \geq 0$, we see that $\frac{a}{2 \cdot 3} - \frac{1}{2}\sqrt{\left(\frac{a}{3}\right)^2 + 4a\sigma^2} < 0$ which is not a valid root as $\varepsilon > 0$. Thus,

$$\begin{aligned}
\varepsilon &= \frac{1}{3n} \log\left(\frac{|R_{set}(\mathcal{F},\gamma)|}{\delta}\right) + \sqrt{\left(\frac{1}{3n} \log\left(\frac{|R_{set}(\mathcal{F},\gamma)|}{\delta}\right)\right)^2 + \frac{2\sigma^2}{n} \log\left(\frac{|R_{set}(\mathcal{F},\gamma)|}{\delta}\right)} \\
&\leq \frac{2}{3n} \log\left(\frac{|R_{set}(\mathcal{F},\gamma)|}{\delta}\right) + \sqrt{\frac{2\sigma^2}{n} \log\left(\frac{|R_{set}(\mathcal{F},\gamma)|}{\delta}\right)},
\end{aligned}$$

where the latter inequality arises from the inequality $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$. Therefore, we get that with probability at least $1 - \delta$:

$$\forall f \in R_{set}(\mathcal{F},\gamma) : L(f) - \hat{L}(f) \leq \varepsilon = \frac{2}{3n} \log\left(\frac{|R_{set}(\mathcal{F},\gamma)|}{\delta}\right) + \sqrt{\frac{2\sigma^2}{n} \log\left(\frac{|R_{set}(\mathcal{F},\gamma)|}{\delta}\right)}.$$

■

Note that generalization bounds are usually based on Hoeffding's inequality (see Lemma 38), which is a special case of Bernstein's inequality (see Lemma 37). In fact, we show in Appendix B.1 that Bernstein's inequality, which we used to prove Theorem 28, can be sharper than Hoeffding's when the variance is less than $\sigma_f^2 < \frac{1}{12}$ for a given f. Theorem 28 is easily generalized to continuous hypothesis spaces through a covering argument over the true Rashomon set (as an example, see Theorem 34), where the complexity is measured as the size of the cover over the true Rashomon set instead of the number of models in the true Rashomon set.

Let $c(\mathcal{F}, n) = \frac{2}{3n} \log \left(\frac{|R_{set}(\mathcal{F}, \gamma)|}{\delta} \right)$, which is the first term in the bound (4.2) in Theorem 28. According to the next theorem, under random label noise, the true Rashomon set does not decrease in size.

Theorem 29 (Size of the true Rashomon set cannot decrease under random label noise)**.** *Consider hypothesis space $\mathcal{F}$, data distribution $\mathcal{D} = \mathcal{X} \times \mathcal{Y}$, where $\mathcal{X} \in \mathbb{R}^p$, and $\mathcal{Y} \in \{-1, 1\}$. Let $\rho \in (0, \frac{1}{2})$ be a probability with which each label y_i is flipped independently, and $\mathcal{D}_\rho$ denotes the noisy version of D. For 0-1 loss function, the true Rashomon set over $\mathcal{D}$ is a subset of the true Rashomon set over $\mathcal{D}_\rho$, $R_{set_\mathcal{D}}(\mathcal{F}, \gamma) \subseteq R_{set_{\mathcal{D}_\rho}}(\mathcal{F}, \gamma)$.*

Proof. Recall that f^* is an optimal function, meaning that $f^* \in \text{argmin}_{f \in \mathcal{F}} L_\mathcal{D}(f)$. Given the noisy distribution $\mathcal{D}_\rho$, denote $f^*_\rho \in \text{argmin}_{f \in \mathcal{F}} L_{\mathcal{D}_\rho}(f)$. From the proof of Theorem 22, for any model $f \in \mathcal{F}$, we have $L_{\mathcal{D}_\rho}(f) = (1 - 2\rho) L_\mathcal{D}(f)_\rho$.

Since $L_\mathcal{D}(f^*) \leq L_\mathcal{D}(f^*_\rho)$ as f^* is an optimal model over $\mathcal{D}$, $0 < \rho < 0.5$ by assumption, and for any f in the true Rashomon set $R_{set_\mathcal{D}}(\mathcal{F}, \gamma)$, we have:

$$\begin{aligned} L_{\mathcal{D}_\rho}(f) - L_{\mathcal{D}_\rho}(f^*_\rho) &= L_\mathcal{D}(f)(1 - 2\rho) + \rho - L_\mathcal{D}(f^*_\rho)(1 - 2\rho) - \rho \\ &= (L_\mathcal{D}(f) - L_\mathcal{D}(f^*_\rho))(1 - 2\rho) \\ &\leq (L_\mathcal{D}(f) - L_\mathcal{D}(f^*))(1 - 2\rho) \\ &\leq \gamma(1 - 2\rho) \leq \gamma. \end{aligned} \tag{4.3}$$

Therefore, f is in the true Rashomon set $R_{set_{\mathcal{D}_\rho}}(\mathcal{F}, \gamma)$. As this calculation holds for every model f from the true Rashomon set $R_{set_\mathcal{D}}(\mathcal{F}, \gamma)$, then $R_{set_\mathcal{D}}(\mathcal{F}, \gamma) \subseteq R_{set_{\mathcal{D}_\rho}}(\mathcal{F}, \gamma)$.

■

Therefore, we have that $c(\mathcal{F}, n)$ at least does not decrease with more noise as it depends only on complexity. However, the second term $\sqrt{3\sigma^2 c(\mathcal{F}, n)}$ in Theorem 28 depends on the maximum loss variance, which increases with label noise, as motivated in the previous section.

This means with more noise in the labels, we would expect worse generalization, which would generally lead practitioners who are using a validation set to reduce the complexity of the hypothesis space.

As discussed earlier, we use cross-validation to assess whether $\hat{f}$ overfits. From Proposition 30 we infer that if the empirical risk minimizer (ERM) does not highly overfit, it has a high chance to be in the true Rashomon set and thus Theorem 28 applies.

Proposition 30 (ERM can be close to the true Rashomon set)**.** *Assume that through the cross-validation process, we can assess ξ such that $L(\hat{f}) - \hat{L}(\hat{f}) \leq \xi$ with high probability (at least $1 - \epsilon_\xi$) with respect to the random draw of data. Then, for any $\epsilon > 0$, with probability at least $1 - e^{-2n\epsilon^2} - \epsilon_\xi$ with respect to the random draw of training data, when $\xi + \epsilon \leq \gamma$, then $\hat{f} \in R_{set}(\mathcal{F}, \gamma)$.*

Proof. For a fixed $f \in \mathcal{F}$ for 0-1 loss by Hoeffding's inequality (B.2):

$$P\left[\hat{L}(f) - L(f) > \epsilon\right] \leq e^{-2n\epsilon^2}.$$

Therefore, with probability at least $1 - e^{-2n\epsilon^2}$ with respect to the random draw of data, $\hat{L}(f) - L(f) \leq \epsilon$. This is true for the optimal model as well, thus with high probability $\hat{L}(f^*) - L(f^*) \leq \epsilon$.

Since $\hat{f}$ is the empirical risk minimizer, and $\epsilon + \xi \leq \gamma$ by assumption, we have that $\hat{L}(\hat{f}) \leq \hat{L}(f^*)$. We use that for two events A and B, $P(\neg(A \cup B)) = 1 - P(A \cup B) \geq 1 - P(A) - P(B)$, where A is the event that cross-validation gives us an incorrect generalization bound, and B is the event that f^* does not generalize. Thus, $P(A) \leq e^{2n\epsilon^2}$ and $P(B) \leq \epsilon_\xi$. Thus, with probability at least $1 - e^{-2n\epsilon^2} - \epsilon_\xi$,

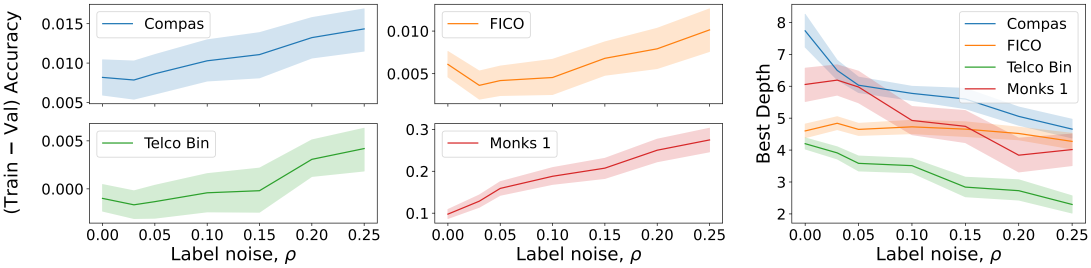
$$L(\hat{f}) \leq \hat{L}(\hat{f}) + \xi \leq \hat{L}(f^*) + \xi \leq L(f^*) + \epsilon + \xi \leq L(f^*) + \gamma.$$

Therefore $\hat{f} \in R_{set}(\mathcal{F}, \gamma)$.

■

4.1.3 Step 3. Practitioner Chooses a Simpler Hypothesis Space

We have shown earlier that noisier datasets lead to higher variance and worse generalization. The question we consider here is whether one can see the results of these bounds in practice and would actually reduce the hypothesis space. For example, consider four real-world datasets and the hypothesis space of decision trees of various depths. In Figure 4.3 (a) we show that as label noise increases, so does the gap between risks (1− accuracy) evaluated on the training set and on a holdout dataset for a fixed depth tree, and thus, as a result, during the validation process, the smaller depth would be chosen by a reasonable analyst. We simulated this by using cross-validation to select the optimal tree depth in CART, as shown in Figure 4.3 (b). As predicted, the optimal tree depth decreases as noise increases. We also show that a similar trend happens for the gradient-boosted trees in Figure 4.4. More specifically, with more noise, the best number of estimators, as chosen based on cross-validation, decreases. We describe the setup in detail in Appendix B.3.



(a)

(b)

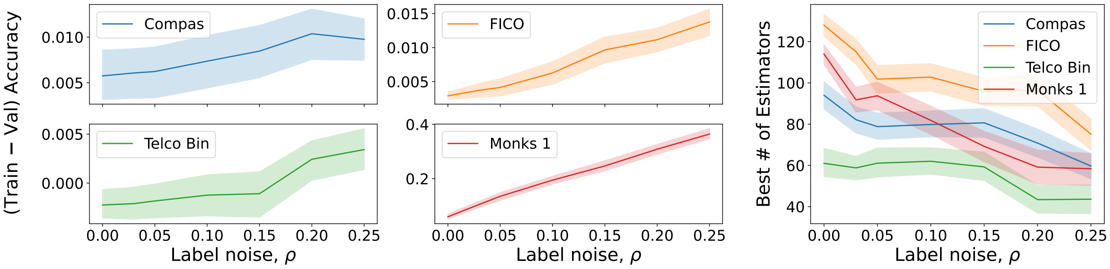
FIGURE 4.3: Practitioner's validation process in the presence of noise for CART. For a fixed tree depth, as we add noise, the gap between training and validation accuracy increases (Subfigure a). As we use cross-validation to select tree depth, the best tree depth decreases with noise (Subfigure b).

4.1.4 Step 4. Rashomon Ratio is Larger for Simpler Spaces

For simpler hypothesis spaces, it may not be immediately obvious that the Rashomon ratio is larger, i.e., a larger fraction of a simpler model class consists of "good" models. Intuitively, this is because the denominator of the ratio (the total number of models) increases faster than the numerator (the number of good models) as the complexity of the model class increases. As we will see, this is because the good models in the simpler model class tend to give rise to more bad models in the more complex class, and the bad models do not tend to give rise to good models as much. We explore two popular model classes: decision trees and linear models. The results thus extend to forests (collections of trees) and generalized additive models (which are linear models in enhanced feature space).

Consider data that live on a hypercube (e.g., the data have been binarized, which is a common pre-processing step (Angelino et al., 2018; Lin et al., 2020; Verwer & Zhang, 2017; Xin et al., 2022)) and a hypothesis space of fully grown trees (complete trees with the last level also filled) of a given depth d. Denote this hypothesis space as $\mathcal{F}_d$. For example, a depth 1 tree has 1 node and 2 leaves. Under natural assumptions on the quality of features of classifiers and the purity of the leaves of trees in the Rashomon set, we show that the Rashomon ratio is larger for hypothesis spaces of smaller-depth trees.

Proposition 31 (Rashomon ratio is larger for decision trees of smaller depth)**.** *For a dataset $S = X \times Y$ with binary feature matrix $X \in \{0,1\}^{n \times p}$, consider a hypothesis space $\mathcal{F}_d$ of fully grown trees of depth d. Let the number of dimensions $p < 2^{2^d}$. Assume: (Leaves are correct) all leaves in all trees in the Rashomon set have at least $\lceil \theta n \rceil$ more correctly classified points than incorrectly classified points; (Bad features) there is a set of $p_{bad} \geq d$ "bad" features*



(a) (b)

Figure 4.4: Practitioner's validation process in the presence of noise for gradient-boosted trees. For a fixed number of estimators, as we add noise, the gap between training and validation accuracy increases (Subfigure a). As we use cross-validation to select the number of estimators, the best number of estimators decreases with noise (Subfigure b).

such that if a model is not in the Rashomon set, it is using only the bad features. Then $\hat{R}_{ratio}(\mathcal{F}_{d+1}, \theta) < \hat{R}_{ratio}(\mathcal{F}_d, \theta)$.

Proof. As in the proof of Theorem 21, the hypothesis space of fully-grown trees of depth d contains

$$|\,\mathcal{F}_d\,| = 2^{2^d} \prod_{k=1}^{d} (p-k+1)^{2^{k-1}}$$

Now let's compute the size of the Rashomon set. First, since each leaf of every tree in the Rashomon set has correctly classified $\lceil \theta n \rceil$ points more than misclassified, flipping the label of this leaf will add more than θ to the loss and thus will push the tree out of the Rashomon set. Therefore, for every tree, every leaf label is determined by the data.

Second, the trees in the Rashomon set must have at least one good feature. The cardinality of the set of good features is $\bar{p} = p - |p_{bad}|$, then the cardinality of the Rashomon set is:

$$\hat{R}_{set}(\mathcal{F}_d, \theta) = \prod_{k=1}^{d} (p-k+1)^{2^{k-1}} - \prod_{k=1}^{d} (p-\bar{p}-k+1)^{2^{k-1}},$$

meaning that we do not consider trees that consist of bad features only (note that $\hat{R}_{set}(\mathcal{F}_d, \theta) < |\,\mathcal{F}_d\,|$ since $p_{bad} \geq d$). Then the Rashomon ratio is:

$$\begin{aligned}
\hat{R}_{ratio}(\mathcal{F}_d, \theta) &= \frac{|\hat{R}_{set}(\mathcal{F}_d, \theta)|}{|\,\mathcal{F}_d\,|} = \frac{\prod_{k=1}^{d}(p-k+1)^{2^{k-1}} - \prod_{k=1}^{d}(p-\bar{p}-k+1)^{2^{k-1}}}{2^{2^d}\prod_{k=1}^{d}(p-k+1)^{2^{k-1}}} \\
&= \frac{1}{2^{2^d}}\left(1 - \frac{\prod_{k=1}^{d}(p-\bar{p}-k+1)^{2^{k-1}}}{\prod_{k=1}^{d}(p-k+1)^{2^{k-1}}}\right) \\
&= \frac{1}{2^{2^d}}\left(1 - \prod_{k=1}^{d}\left(1 - \frac{\bar{p}}{p-k+1}\right)^{2^{k-1}}\right) \\
&= \frac{1-\alpha(d)}{2^{2^d}},
\end{aligned}$$

where $\alpha(d) = \prod_{k=1}^{d}\left(1 - \frac{\bar{p}}{p-k+1}\right)^{2^{k-1}}$. Since $d > 1, \bar{p} > 1$, and $\frac{\bar{p}}{p-k+1} > \frac{\bar{p}}{p}$ for $k > 2$, we get that

$$\alpha(d) = \prod_{k=1}^{d}\left(1 - \frac{\bar{p}}{p-k+1}\right)^{2^{k-1}} < \prod_{k=1}^{d}\left(1 - \frac{\bar{p}}{p}\right)^{2^{k-1}} < 1 - \frac{\bar{p}}{p}.$$

Note as well that $\alpha(d) < 1$ for any d. Recall that $p < 2^{2^d}$, then for the ratio of ratios:

$$\begin{aligned}
\frac{\hat{R}_{ratio}(\mathcal{F}_d, \theta)}{\hat{R}_{ratio}(\mathcal{F}_{d+1}, \theta)} &= \frac{|\hat{R}_{set}(\mathcal{F}_d, \theta)|}{|\,\mathcal{F}_d\,|}\frac{|\,\mathcal{F}_{d+1}\,|}{|\hat{R}_{set}(\mathcal{F}_{d+1}, \theta)|} \\
&= \frac{1-\alpha(d)}{2^{2^d}}\frac{2^{2^{d+1}}}{1-\alpha(d+1)} = 2^{2^d}\frac{1-\alpha(d)}{1-\alpha(d+1)} \\
&> 2^{2^d}(1-\alpha(d)) > 2^{2^d}\left(1 - \left(1 - \frac{\bar{p}}{p}\right)\right) \\
&= 2^{2^d}\frac{\bar{p}}{p} > 2^{2^d}\frac{1}{2^{2^d}} = 1.
\end{aligned}$$

Thus we showed that $\hat{R}_{ratio}(\mathcal{F}_d, \theta) > \hat{R}_{ratio}(\mathcal{F}_{d+1}, \theta)$, meaning that the Rashomon ratio grows as we consider less deep trees.

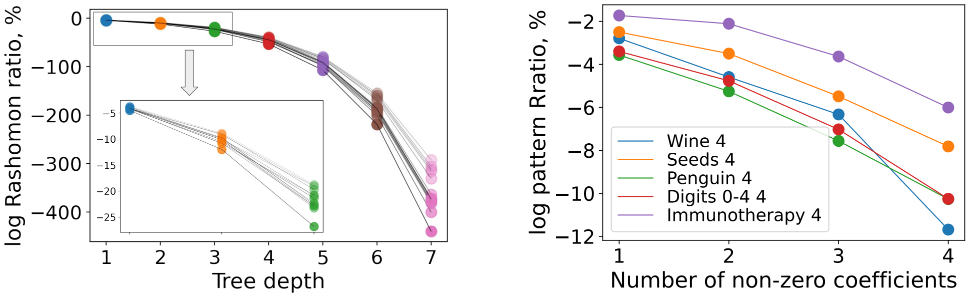
■

Both assumptions of Proposition 31 are typically satisfied in practice.

To demonstrate our point, we computed the Rashomon ratio and pattern Rashomon ratio for 19 different datasets for hypothesis spaces of decision trees and linear models of different complexity (see Figure 4.5). As the complexity of the hypothesis space increases, we see an obvious decrease in the Rashomon ratio and pattern Rashomon ratio.

For trees, to compute the numerator of the Rashomon ratio, we used TreeFARMS (Xin et al., 2022), which allows us to enumerate the whole Rashomon set for sparse trees as we discussed in Section 2.3.4. We set the Rashomon parameter to 5% (however, as we show in Figure 4.13(a) the choice of the Rashomon parameter does not change results). To compute the denominator of the Rashomon ratio, we used the recursive formula (2.5) in Section 2.3.4.

For linear models, we computed the pattern Rashomon ratio using the branch and bound approach as in Section 2.4.2. For the hierarchy of regularized linear models (Figure 4.5 (b)), we considered regularization for 1 non-zero coefficient, 2 non-zero coefficients, 3 non-zero coefficients, and 4 non-zero coefficients. We set the Rashomon parameter to 3%. To compute the denominator of the pattern Rashomon ratio, we used the formula as in (2.8) that gives the number of all possible patterns for the hypothesis space of linear models.



(a) (b)

FIGURE 4.5: Calculation showing that the Rashomon ratio (a) and pattern Rashomon ratio (b) decrease for the hypothesis space of decision trees of fixed depth from 1 to 7 for 14 different datasets (a) and for the hypothesis space of linear models of sparsity from 1 to 4 for 5 different datasets (b). Each line represents a different dataset, each dot represents the log of the Rashomon ratio or pattern Rashomon ratio. Both ratios decrease as we move to a more complex hypothesis space.

Completion of the Path. After reducing the complexity of the hypothesis space in Step 3, the practitioner has already arrived at a simpler hypothesis space, which is the goal. Continuing on the path, they can reduce complexity further. Specifically, they find a larger Rashomon ratio for the newly chosen lower complexity hypothesis space according to Step 4. As a reminder of the thesis of Chapter 3, with large Rashomon ratios, there are many good models, among which may exist even simpler models that perform well. Thus, the path, starting from noise, is a powerful way to explain what we see in practice, which is that simple models often perform well (Holte, 1993).

Note that to follow the path, the machine learning practitioner does not need to know the exact amount of noise. As long as they suspect *some* noise present in the dataset, the results in this Chapter apply, and the practitioner would expect a good performance from simpler models.

4.2 Rashomon Ratio for Ridge Regression Increases under Additive Attribute Noise

For linear regression, adding multiplicative or additive noise to the training data is known to be equivalent to regularizing the model parameters (Bishop, 1995). Moreover, the more noise is added, the stronger this regularization is. Thus, noise leads directly to Step 3 and a choice of a smaller (simpler) hypothesis space. Also, for ridge regression, the Rashomon volume (the numerator of the Rashomon ratio) can be computed directly (Section 2.3.2) and depends on the regularization parameter. Building upon these results, we prove that noise leads to an increase of the Rashomon ratio (as in Step 4 of the path).

Given dataset S, the ridge regression model is learned by minimizing the penalized sum of squared errors: $\hat{L}(\omega) = \hat{L}_{LS}(\omega) + C\omega^T\omega$, where $\hat{L}_{LS}(\omega) = \frac{1}{n}\sum_{i=1}^{n}\left(x_i^T\omega - y_i\right)^2$ is the least squares loss, C is a regularization parameter, and $\omega \in \mathbb{R}^p$ is a parameter vector for a linear model $f = \omega^T x$.

We will assume that there is a maximum loss value $\hat{L}_{\max}$, such that any linear model that has higher regularized loss than $\hat{L}_{\max}$ is not being considered within the hypothesis space. For instance, an upper bound for $\hat{L}_{\max}$ is the value of the loss at the model that is identically $\bar{0}$, namely $\hat{L}(\bar{0}) = \frac{1}{n}\sum_i y_i^2$. Thus, for every reasonable model $f = \omega^T x$, $f \in \mathcal{F} \equiv \hat{L}(\omega) \leq \hat{L}_{\max}$. On the other hand, the best possible value of the least squares loss is $\hat{L}_{LS} = 0$, and therefore we get that $Cw^T w \leq \hat{L}_{\max}$, or alternatively, $w^T w \leq \hat{L}_{\max}/C$. This defines the hypothesis space as an ℓ_2-norm ball in m-dimensional space, the volume of which we can compute.

To measure the numerator of the Rashomon ratio we will use the Rashomon volume $\mathcal{V}(\hat{R}_{set}(\mathcal{F}, \theta))$, as defined in Semenova et al. (2022). In the case of ridge regression, the Rashomon set is an ellipsoid in m-dimensions, thus the Rashomon volume can be computed directly. Therefore, we have the Rashomon ratio as the ratio of the Rashomon volume to the volume of the ℓ_2-norm ball that defines the hypothesis space.

Next, we show that under additive attribute noise, the Rashomon ratio increases:

Theorem 32 (Rashomon ratio increases with noise for ridge regression)**.** *Consider dataset $S = X \times Y$, X is a non-zero matrix, and a hypothesis space of linear models $\mathcal{F} = \{f = \omega^T x, \omega \in \mathbb{R}^p, \omega^T\omega \leq \hat{L}_{\max}/C\}$. Let ϵ_i, such that $\epsilon_i \sim \mathcal{N}(\bar{0}, \lambda I)$ ($\lambda > 0$, I is identity matrix), be i.i.d. noise vectors added to every sample: $x_i' = x_i + \epsilon_i$. Consider options $\lambda_1 > 0$ and $\lambda_2 > 0$ that control how much noise we add to the dataset. For ridge regression, if $\lambda_1 < \lambda_2$, then the Rashomon ratios obey $\hat{R}_{ratio_{\lambda_1}}(\mathcal{F}, \theta)) < \hat{R}_{ratio_{\lambda_2}}(\mathcal{F}, \theta))$.*

Proof. For simplicity denote $\mathbb{E}_{\epsilon_1,\ldots,\epsilon_n \sim \mathcal{N}(\bar{0},\lambda I)}$ as $\mathbb{E}_\epsilon$. To find the optimal solution, under added noise, we would like to minimize expected regularized least squares:

$$\begin{aligned}
\mathbb{E}_\epsilon \hat{L}(\omega) &= \mathbb{E}_\epsilon \left[\frac{1}{n}\sum_{i=1}^{n}((x_i + \epsilon_i)^T\omega - y_i)^2 + C\omega^T\omega\right] \\
&= \mathbb{E}_\epsilon \left[\frac{1}{n}\sum_{i=1}^{n}\left((x_i^T\omega - y_i)^2 + 2\epsilon^T\omega(x_i^T\omega - y_i) + \omega^T\epsilon_i\epsilon_i^T\omega\right)\right] + C\omega^T\omega \\
&= \frac{1}{n}\sum_{i=1}^{n}\Big((x_i^T\omega - y_i)^2 + 2\,\mathbb{E}_\epsilon\left[\epsilon_i\right]^T\omega(x_i^T\omega - y_i) + \omega^T\,\mathbb{E}_\epsilon\left[\epsilon_i\epsilon_i^T\right]\omega\Big) + C\omega^T\omega \\
&= \frac{1}{n}\sum_{i=1}^{n}\left((x_i^T\omega - y_i)^2 + \omega^T(\lambda I)\omega\right) + C\omega^T\omega \\
&= \frac{1}{n}\sum_{i=1}^{n}(x_i^T\omega - y_i)^2 + (C + \lambda)\omega^T\omega,
\end{aligned}$$

where $\mathbb{E}_\epsilon\left[\epsilon_i\epsilon_i^T\right] = \lambda I$, I is identity matrix, and $E_\epsilon\left[\epsilon_i\right] = \bar{0}$.

Therefore, adding attribute noise to the training data becomes equivalent to ℓ_2-regularization, and the new regularization parameter is $C + \lambda$. According to Theorem 8 in Section 2.3.2, the Rashomon volume can be computed as:

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

where σ_i are singular values of matrix X, and $\Gamma(\cdot)$ is the Gamma-function.

On the other hand, for the regularization parameter $C + \lambda$, the hypothesis space is defined as $(C + \lambda)w^Tw \leq \hat{L}_{\max}$, meaning that $w^Tw \leq \frac{\hat{L}_{\max}}{C+\lambda}$. The volume of the ball defined by the ℓ_2-norm in m-dimensional space with radius R, $\|x\|_2 = \left(\sum_{i=1}^m |x_i|^2\right)^{\frac{1}{2}} \leq R$, can be computed as:

$$\mathcal{V}_p^2(R) = \frac{\pi^{\frac{p}{2}}}{\Gamma(\frac{p}{2}+1)} R^p.$$

Since for $\|\omega\|_2^2 = w^T w \leq \frac{\hat{L}_{\max}}{C+\lambda}$, we have radius $R_\lambda = \sqrt{\frac{\hat{L}_{\max}}{C+\lambda}}$, we get that the Rashomon ratio obeys:

$$\begin{aligned}\hat{R}_{ratio_\lambda}(\mathcal{F}, \theta)) &= \frac{\mathcal{V}(\hat{R}_{set_\lambda}(\mathcal{F}, \theta))}{\mathcal{V}_p^2(R_\lambda)} \\ &= \frac{(\pi\theta)^{\frac{p}{2}}}{\Gamma(\frac{p}{2}+1)} \left[\prod_{i=1}^p \frac{1}{\sqrt{\sigma_i^2 + C + \lambda}}\right] \frac{\Gamma(\frac{p}{2}+1)}{\pi^{\frac{p}{2}}} \frac{(C+\lambda)^{\frac{p}{2}}}{(\hat{L}_{\max})^{\frac{p}{2}}} \\ &= \left(\frac{\theta}{\hat{L}_{\max}}\right)^{\frac{p}{2}} \prod_{i=1}^p \sqrt{\frac{C+\lambda}{\sigma_i^2 + C + \lambda}}.\end{aligned}$$

Since $0 < \lambda_1 < \lambda_2, C > 0$, without loss of generality, let $\lambda_C = \lambda_1 + C$, and $\lambda_C + \delta = \lambda_2$ + C, where $\delta = \lambda_2 - \lambda_1 > 0$. Consider function $\frac{x}{a+x}$, where $a > 0$. This function is monotonically increasing for all $x > 0$, since $\frac{\partial}{\partial x}\left(\frac{x}{a+x}\right) = \frac{a}{(a+x)^2} > 0$. Therefore, for all non-zero σ_i^2:

$$\frac{\lambda_C}{\sigma_i^2 + \lambda_C} < \frac{\lambda_C + \delta}{\sigma_i^2 + \lambda_C + \delta}.$$

Since X is a non-zero matrix, there is at least one non-zero singular value σ_i^2. Given the monotonicity of the square root function, we have that for the Rashomon ratios for noise levels λ_1 and λ_2:

$$\begin{aligned}\hat{R}_{ratio_{\lambda_1}}(\mathcal{F}, \theta)) = \hat{R}_{ratio_{\lambda_C}}(\mathcal{F}, \theta)) &= \left(\frac{\theta}{\hat{L}_{\max}}\right)^{\frac{p}{2}} \prod_{i=1}^p \sqrt{\frac{\lambda_C}{\sigma_i^2 + \lambda_C}} \\ &< \left(\frac{\theta}{\hat{L}_{\max}}\right)^{\frac{p}{2}} \prod_{i=1}^p \sqrt{\frac{\lambda_C + \delta}{\sigma_i^2 + \lambda_C + \delta}} \\ &= \hat{R}_{ratio_{\lambda_C+\delta}}(\mathcal{F}, \theta)) = \hat{R}_{ratio_{\lambda_2}}(\mathcal{F}, \theta)).\end{aligned}$$

Therefore we proved that with the additive attribute noise, the Rashomon ratio increases.

■

Compared to the Rashomon ratio, the relationship between the regularization parameter and the Rashomon volume is inverted: the stronger the regularization, the smaller the Rashomon volume. This means that adding more noise leads to stronger regularization and smaller Rashomon volume. In some ways, this is consistent with what we saw in Figure 4.3(b), where CART preferred shorter trees in the presence of noise.

Note that while in Theorem 32 we directly show that adding noise to the training data is equivalent to stronger regularization leading us directly to Step 3, Steps 1 and 2 of the path identified in the previous section are automatically satisfied. We next formally prove that additive noise still leads to an increase of the variance of losses for least squares loss (similar to Theorems 22, 23, and 26) in Theorem 33, and we show that an increase in the maximum variance of losses leads to worse generalization bound (similar to Theorem 28) for the squared loss in Theorem 34.

Theorem 33 (Variance of least squares loss increases with noise)**.** *Consider dataset* $S = X \times Y$*, where* $X \in \mathbb{R}^{n \times p}$*,* $Y \in \mathbb{R}^n$*,* $z_i = (x_i, y_i)$*. Let* $\epsilon_i = \{\epsilon_{ij}\}_{j=1}^p$*, such that* $\epsilon_{ij} \sim \mathcal{N}(0, \sigma_\mathcal{N}^2)$*, be i.i.d. noise vectors added to every sample:* $x_i' = x_i + \epsilon_i$*. Consider* $\sigma_{\mathcal{N}_1}^2 > 0$*,* $\sigma_{\mathcal{N}_2}^2 > 0$ *that control how much noise is added to the dataset. For the least squares loss* $l(z_i) = r_i^2 = (w^T x_i - y_i)^2$ *and a fixed model* $f(x) = \omega^T x$*, where* $\omega \in \mathbb{R}^p$*,* $\omega \neq \bar{0}$*, the variance of losses increases with more noise: if* $\sigma_{\mathcal{N}_1}^2 < \sigma_{\mathcal{N}_2}^2$*, then:* $\sigma^2(f, S_{\sigma_{\mathcal{N}_1}}) < \sigma^2(f, S_{\sigma_{\mathcal{N}_2}})$*.*

Proof. For simplicity, denote $\mathbb{E}_{\epsilon_{11},\dots,\epsilon_{1p},\dots,\epsilon_{n1},\dots,\epsilon_{np}}$ as $\mathbb{E}_{\bar{\epsilon}}$, and $\mathbb{E}_{x_i,y_i}$ as $\mathbb{E}_z$. The variance of losses for the least squares loss under the additive normal noise is: $\sigma^2(f, S_{\sigma_\mathcal{N}}) = Var_{z,\bar{\epsilon}}\left[\left((x_i + \epsilon_i)^T \omega - y_i\right)^2\right]$. Also, for simplicity, we will omit index i over samples (but keep index j over the dimensions). Recall that $r = x^T \omega - y$. From the definition of the variance we have that:

$$\begin{aligned} Var_{z,\bar{\epsilon}}\left[\left((x+\epsilon)^T\omega - y\right)^2\right] &= Var_{z,\bar{\epsilon}}\left[\left((x^T\omega - y) + \epsilon^T\omega\right)^2\right] = Var_{z,\bar{\epsilon}}\left[\left(r+\epsilon^T\omega\right)^2\right] \\ &= \mathbb{E}_{z,\bar{\epsilon}}\left[\left(r+\epsilon^T\omega\right)^4\right] - \left(\mathbb{E}_{z,\bar{\epsilon}}\left[\left(r+\epsilon^T\omega\right)^2\right]\right)^2. \end{aligned} \tag{4.4}$$

Since $\epsilon_j \sim \mathcal{N}(0,\sigma^2_{\mathcal{N}})$, we have that $\mathbb{E}_{\epsilon_j}\left[\epsilon_j\right] = 0$, $\mathbb{E}_{\epsilon_j}\left[(\epsilon_j)^2\right] = \sigma^2_{\mathcal{N}}$, $\mathbb{E}_{\epsilon_j}\left[(\epsilon_j)^3\right] = 0$ (this is a property of Gaussian random variables), and $\mathbb{E}_{\epsilon_j}\left[(\epsilon_j)^4\right] = 3\sigma^4_{\mathcal{N}}$. Also recall that the multinomial theorem states:

$$\left(\sum_{j=1}^{p} a_j\right)^t = \sum_{k_1+k_2+\ldots+k_p=t} \frac{t!}{k_1!\cdot k_2!\cdot\ldots\cdot k_p!}\cdot a_1^{k_1}\cdot a_2^{k_2}\cdot\ldots\cdot a_p^{k_p}.$$

The multinomial theorem helps us to compute coefficients of the first four moments for $\epsilon^T\omega$. More specifically, for the first and the second moments:

$$\mathbb{E}_{\bar{\epsilon}}\left[\epsilon^T\omega\right] = 0,$$

$$\begin{aligned} \mathbb{E}_{\bar{\epsilon}}\left[(\epsilon^T\omega)^2\right] &= \mathbb{E}_{\bar{\epsilon}}\left[\left(\sum_{j=1}^{p}\epsilon_j\omega_j\right)^2\right] = \mathbb{E}_{\bar{\epsilon}}\left[\sum_{j=1}^{p}(\epsilon_j\omega_j)^2 + \sum_{j=1..p,k=1..p,j\neq k}\epsilon_j\omega_j\epsilon_k\omega_k\right] \\ &= \sum_{j=1}^{p}\mathbb{E}_{\bar{\epsilon}}\left[\epsilon_j^2\right]\omega_j^2 = \sigma^2_{\mathcal{N}}\omega^T\omega. \end{aligned}$$

For the third moment, notice that from the multinomial theorem, $k_1+\cdots+k_p = 3$. Then there are three possible combinations of the values of k_j: some $k_a = 3$ and the rest are 0, some $k_a = 2$, $k_b = 1$, and the rest are 0, and finally some $k_a = 1$, $k_b = 1$, $k_c = 1$ and the rest are 0. All of these cases will lead to the presence of either ϵ_j^3 or ϵ_j in the product. Since $\mathbb{E}_{\epsilon_j}\left[\epsilon_j\right] = 0$, and $\mathbb{E}_{\epsilon_j}\left[\epsilon_j^3\right] = 0$, we have that

$$\mathbb{E}_{\bar{\epsilon}}\left[(\epsilon^T\omega)^3\right] = 0.$$

Similarly, for $\mathbb{E}_{\bar{\epsilon}}\left[\left(\sum_{j=1}^{p}\epsilon_j\omega_j\right)^4\right]$ we get non-zero terms for some of the combinations and the others are 0. In particular, non-zero terms arise when some $k_a = 4$ and the rest are 0, and some $k_a = 2$, $k_b = 2$, and the rest are 0s. This gives us:

$$\begin{aligned}
\mathbb{E}_{\bar{\epsilon}}\left[(\epsilon^T\omega)^4\right] &= \mathbb{E}_{\bar{\epsilon}}\left[\left(\sum_{j=1}^{p}\epsilon_j\omega_j\right)^4\right] \\
&= \sum_{j=1}^{p}\mathbb{E}_{\bar{\epsilon}}\left[\epsilon_j^4\right]\omega_j^4 + 6\sum_{j=1..p,k=1..p,j\neq k}\mathbb{E}_{\bar{\epsilon}}\left[\epsilon_j^2\right]\omega_j^2\mathbb{E}_{\bar{\epsilon}}\left[\epsilon_k^2\right]\omega_k^2 \\
&= 3\sigma_{\mathcal{N}}^4\sum_{j=1}^{p}\omega_j^4 + 6\sigma_{\mathcal{N}}^4\sum_{j=1..p,k=1..p,j\neq k}\omega_j^2\omega_k^2 \\
&= 3\sigma_{\mathcal{N}}^4(\omega^T\omega)^2.
\end{aligned}$$

Let's focus on the first term of the variance equation (4.4):

$$\begin{aligned}
\mathbb{E}_{z,\bar{\epsilon}}\left[\left(r+\epsilon^T\omega\right)^4\right] &= \mathbb{E}_{z,\bar{\epsilon}}\left[r^4 + 4r^3\epsilon^T\omega + 6r^2(\epsilon^T\omega)^2 + 4r(\epsilon^T\omega)^3 + (\epsilon^T\omega)^4\right] \\
&= \mathbb{E}_z\left[r^4\right] + 4\mathbb{E}_z\left[r^3\right]\mathbb{E}_{\bar{\epsilon}}\left[\epsilon^T\omega\right] + 6\mathbb{E}_z\left[r^2\right]\mathbb{E}_{\bar{\epsilon}}\left[(\epsilon^T\omega)^2\right] \\
&+ 4\mathbb{E}_z\left[r\right]\mathbb{E}_{\bar{\epsilon}}\left[(\epsilon^T\omega)^3\right] + \mathbb{E}_{\bar{\epsilon}}\left[(\epsilon^T\omega)^4\right] \\
&= \mathbb{E}_z\left[r^4\right] + 6\sigma_{\mathcal{N}}^2\omega^T\omega\mathbb{E}_z\left[r^2\right] + 3\sigma_{\mathcal{N}}^4(\omega^T\omega)^2.
\end{aligned}$$

Now, we focus on the second term of the variance equation (4.4):

$$\begin{aligned}
\left(\mathbb{E}_{z,\bar{\epsilon}}\left[\left(r+\epsilon^T\omega\right)^2\right]\right)^2 &= \left(\mathbb{E}_{z,\bar{\epsilon}}\left[r^2 + 2r\epsilon^T\omega + (\epsilon^T\omega)^2\right]\right)^2 \\
&= \left(\mathbb{E}_z\left[r^2\right] + \mathbb{E}_z\left[2r\right]\mathbb{E}_{\bar{\epsilon}}\left[\epsilon^T\omega\right] + \mathbb{E}_{\bar{\epsilon}}\left[(\epsilon^T\omega)^2\right]\right)^2 \\
&= \left(\mathbb{E}_z\left[r^2\right] + \sigma_{\mathcal{N}}^2\omega^T\omega\right)^2 \\
&= \left(\mathbb{E}_z\left[r^2\right]\right)^2 + 2\sigma_{\mathcal{N}}^2\omega^T\omega\mathbb{E}_z\left[r^2\right] + \sigma_{\mathcal{N}}^4(\omega^T\omega)^2.
\end{aligned}$$

Therefore, for the variance, we get that:

$$Var_{z,\bar{\epsilon}}\left[\left(r+\epsilon^T\omega\right)^2\right] = \mathbb{E}_z\left[r^4\right] + 6\sigma_{\mathcal{N}}^2\omega^T\omega\mathbb{E}_z\left[r^2\right] + 3\sigma_{\mathcal{N}}^4(\omega^T\omega)^2$$

$$
\begin{aligned}
&- \left(\mathbb{E}_z\left[r^2\right]\right)^2 - 2\sigma_{\mathcal{N}}^2\omega^T\omega\mathbb{E}_z\left[r^2\right] - \sigma_{\mathcal{N}}^4(\omega^T\omega)^2 \\
&= 2\sigma_{\mathcal{N}}^4(\omega^T\omega)^2 + 4\sigma_{\mathcal{N}}^2\omega^T\omega\mathbb{E}_z\left[r^2\right] + 2\left(\mathbb{E}_z\left[r^2\right]\right)^2 + \mathbb{E}_z\left[r^4\right] - 3\left(\mathbb{E}_z\left[r^2\right]\right)^2 \\
&= 2\left(\sigma_{\mathcal{N}}^2\omega^T\omega + \mathbb{E}_z\left[r^2\right]\right)^2 + \mathbb{E}_z\left[r^4\right] - 3\left(\mathbb{E}_z\left[r^2\right]\right)^2.
\end{aligned}
$$

Next, we will take the derivative of the variance with respect to $\sigma_{\mathcal{N}}^2$:

$$
\begin{aligned}
\frac{\partial}{\partial\sigma_{\mathcal{N}}^2}\left(Var_{z,\bar{\epsilon}}\left[\left(r+\epsilon^T\omega\right)^2\right]\right) &= \frac{\partial}{\partial\sigma_{\mathcal{N}}^2}\left(2\left(\sigma_{\mathcal{N}}^2\omega^T\omega + \mathbb{E}_z\left[r^2\right]\right)^2 + \mathbb{E}_z\left[r^4\right] - 3\left(\mathbb{E}_z\left[r^2\right]\right)^2\right) \\
&= 4\left(\sigma_{\mathcal{N}}^2\omega^T\omega + \mathbb{E}_z\left[r^2\right]\right)\omega^T\omega > 0,
\end{aligned}
$$

since $\sigma_{\mathcal{N}}^2 > 0$ by assumption, $\omega^T\omega > 0$ since $\omega \neq \bar{0}$, and the risk of the least squares loss $\mathbb{E}_z\left[r^2\right] \geq 0$. Therefore, the variance of losses for a fixed model $f = \omega^T x$ monotonically increases for $\sigma_{\mathcal{N}}^2 > 0$. Thus, for $\sigma_{\mathcal{N}_1}^2 < \sigma_{\mathcal{N}_2}^2$ we have that:

$$
\sigma^2(f, S_{\sigma_{\mathcal{N}_1}}) < \sigma^2(f, S_{\sigma_{\mathcal{N}_2}}).
$$

■

As before, Corollary 27 is easily extendable to the results of Theorem 33, meaning that the maximum variance of losses, $\sigma^2 = \sup_{f\in\mathbb{R}_{set}(\mathcal{F},\gamma)} Var_{z\sim\mathcal{D}}l(f,z)$, will also increase for the least squares loss under increasing additive attribute noise.

Next, we show that when the maximum variance σ^2 increases, the generalization bound becomes worse for the least squares loss and the continuous hypothesis space. Cucker and Smale (2002) proved the generalization bound based on Bernstein's inequality for the least squares loss (Theorem B). We state and provide proof of the theorem for the true Rashomon set. To derive the generalization bound, we use the covering number over the true Rashomon set. Recall that for the functional space $\mathcal{F}$ and any $\epsilon > 0$, the ℓ_∞ *covering number* $N(\mathcal{F},\epsilon)$ of $\mathcal{F}$ is the minimum number of balls of radius ϵ, such that they can cover $\mathcal{F}$, meaning that there exist $h_1, ..., h_{N(\mathcal{F},\epsilon)} \in \mathcal{F}$, such that for every $f \in \mathcal{F}$ there is $k \leq N(\mathcal{F},\epsilon)$ such that $\|f - h_k\|_\infty = \max_{x\in\mathcal{X}} |f(x) - h_k(x)| \leq \epsilon$. Now we focus on the theorem.

Theorem 34 (Variance-based "generalization bound" for least squares loss)**.** *Consider data distribution $\mathcal{D}$ over $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$, dataset $S = \{z_i\}_{i=1}^n \sim \mathcal{D}$, hypothesis space $\mathcal{F}$, and the least squares loss $l(f, z) = (f(x) - y)^2$. Let the loss be bounded by $C^2 > 0$ such that $l(f, z) \leq C^2$ for every $z \in \mathcal{Z}$. For any $\epsilon > 0$:*

$$P\left(\sup_{f \in R_{set}(\mathcal{F},\gamma)} L(f) - \hat{L}(f) > \varepsilon\right) \leq N\left(R_{set}(\mathcal{F},\gamma), \frac{\epsilon}{8C}\right) e^{\frac{-n(\varepsilon/2)^2}{2\sigma^2 + C^2\varepsilon/3}},$$

where $\sigma^2 = \sup_{R_{set}(\mathcal{F},\gamma)} Var_{z \in \mathcal{D}} l(f, z)$.

Proof. For each fixed model $f \in R_{set}(\mathcal{F}, \gamma)$ in the true Rashomon set, from Bernstein's inequality and given that loss is bounded by C^2, we have that

$$P(L(f) - \hat{L}(f) > \varepsilon) \leq e^{\frac{-n\varepsilon^2}{2\sigma_f^2 + 2C^2\varepsilon/3}}.$$

Let $B_1, \ldots B_{N\left(R_{set}(\mathcal{F},\gamma), \frac{\epsilon}{8C}\right)}$ be an ℓ_∞ cover of radius $\frac{\epsilon}{8C}$ of the true Rashomon set, meaning that $R_{set}(\mathcal{F},\gamma) \subseteq \bigcup_{k=1}^{N\left(R_{set}(\mathcal{F},\gamma), \frac{\epsilon}{8C}\right)} B_k$, where $N\left(R_{set}(\mathcal{F},\gamma), \frac{\epsilon}{8C}\right)$ is the covering number. Since the loss is bounded, $l(f, z) = (f(x) - y)^2 \leq C^2$, then $|f(x) - y| \leq C$. For every $f \in B_k$, we have that $\|f - h_k\|_\infty \leq \frac{\epsilon}{8C}$, where h_k is the center of the ball B_k. Therefore:

$$\begin{aligned}
(L(f) - \hat{L}(f)) - (L(h_k) - \hat{L}(h_k)) &= (L(f) - L(h_k)) + (\hat{L}(h_k) - \hat{L}(f)) \\
&= \mathbb{E}_{z\sim\mathcal{D}}\left[l(f, z) - l(h_k, z)\right] + \hat{\mathbb{E}}_{z_i\sim S}\left[l(f, z_i) - l(h_k, z_i)\right] \\
&= \mathbb{E}_{z\sim\mathcal{D}}\left[(f(x) - y)^2 - (h_k(x) - y)^2\right] + \hat{\mathbb{E}}_{z_i\sim S}\left[(f(x_i) - y_i)^2 - (h_k(x_i) - y_i)^2\right] \\
&= \mathbb{E}_{z\sim\mathcal{D}}\left[(f(x) - h_k(x))\left((f(x) - y) + (h_k(x) - y)\right)\right] \\
&+ \hat{\mathbb{E}}_{z_i\sim S}\left[(f(x_i) - h_k(x_i))\left((f(x_i) - y_i) + (h_k(x_i) - y_i)\right)\right] \\
&\leq \mathbb{E}_{z\sim\mathcal{D}}\left[\|f - h_k\|_\infty (C + C)\right] + \hat{\mathbb{E}}_{z_i\sim S}\left[\|f - h_k\|_\infty (C + C)\right] \\
&= 4C\|f - h_k\|_\infty \leq 4C\frac{\epsilon}{8C} = \frac{\epsilon}{2}.
\end{aligned}$$

Therefore, if $L(f) - \hat{L}(f) > \epsilon$, we have $L(h_k) - \hat{L}(h_k) \geq (L(f) - \hat{L}(f)) - \frac{\epsilon}{2} > \epsilon - \frac{\epsilon}{2} = \frac{\epsilon}{2}$.

This holds for every $f \in B_k$, and thus for $\arg\sup_{f \in B_k}$ as well:

$$P\left(\sup_{f \in B_k} L(f) - \hat{L}(f) > \varepsilon\right) \leq P\left(L(h_k) - \hat{L}(h_k) > \frac{\varepsilon}{2}\right). \tag{4.5}$$

Since the exponential function is monotonic, $e^{-\frac{1}{\left(\sigma^2_{h_k}\right)}} \leq e^{-\frac{1}{\sigma^2}}$. Based on the definition of the covering number, according to the union bound and (4.5) we have that:

$$\begin{aligned}
P\left(\sup_{f \in R_{set}(\mathcal{F},\gamma)} L(f) - \hat{L}(f) > \varepsilon\right) &= P\left(\exists f \in R_{set}(\mathcal{F},\gamma) : L(f) - \hat{L}(f) > \varepsilon\right) \\
&\leq P\left(\bigcup_{k=1}^{N\left(R_{set}(\mathcal{F},\gamma), \frac{\epsilon}{8C}\right)} \exists f \in B_k : L(f) - \hat{L}(f) > \varepsilon\right) \\
&\leq \sum_{k=1}^{N\left(R_{set}(\mathcal{F},\gamma), \frac{\epsilon}{8C}\right)} P\left(\exists f \in B_k : L(f) - \hat{L}(f) > \varepsilon\right) \\
&\leq \sum_{k=1}^{N\left(R_{set}(\mathcal{F},\gamma), \frac{\epsilon}{8C}\right)} P\left(L(h_k) - \hat{L}(h_k) > \frac{\varepsilon}{2}\right) \\
&\leq \sum_{k=1}^{N\left(R_{set}(\mathcal{F},\gamma), \frac{\epsilon}{8C}\right)} e^{\frac{-n(\varepsilon/2)^2}{2\sigma^2_{h_k} + C^2\varepsilon/3}} \\
&\leq \sum_{k=1}^{N\left(R_{set}(\mathcal{F},\gamma), \frac{\epsilon}{8C}\right)} e^{\frac{-n(\varepsilon/2)^2}{2\sigma^2 + C^2\varepsilon/3}} \\
&= N\left(R_{set}(\mathcal{F},\gamma), \frac{\epsilon}{8C}\right) e^{\frac{-n(\varepsilon/2)^2}{2\sigma^2 + C^2\varepsilon/3}}.
\end{aligned}$$

Therefore we obtained the desired bound. ■

Since $e^{-1/x}$ monotonically increases for $x > 0$, in Theorem 34, as the maximum variance of losses increases, the bound on the right side increases as well, and thus the generalization bound becomes worse.

Returning to the Path. For ridge regression, we have now built *a direct noise-to-Rashomon-ratio argument* showing that, in the presence of noise, the Rashomon ratios are larger. As before, for larger Rashomon ratios, there are multiple good models, including simpler ones that are easier to find.

4.3 Rashomon Set Characteristics in the Presence of Noise

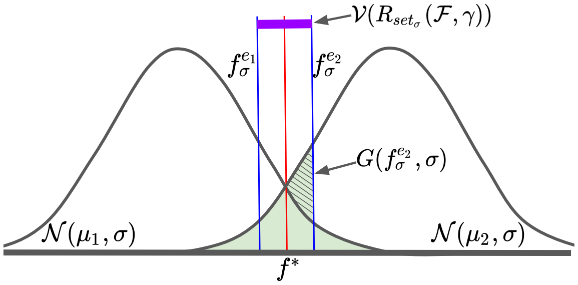
Now we discuss a different mechanism for obtaining larger Rashomon sets. Suppose the practitioner knows the data are noisy. They would then expect a large Rashomon set, which we speculate in this section is explained by noise in the data. Once the practitioner knows they have a large Rashomon set, they could hypothesize from the reasoning in Chapter 3 that a simple model might perform well for their dataset. In this section, will show theoretically and experimentally that characteristics of the Rashomon set are likely to increase under label noise.

4.3.1 Margin Noise is Likely to Increase Rashomon Set

Consider the setting of linear (Gaussian) discriminant analysis, where the data arise from two Gaussians, one with positive labels and one with negative labels. We will add margin noise to the dataset by either increasing the variances or changing the means of the Gaussians, making the two distributions overlap. In Section 4.1, we showed that margin noise leads to an increase in the variance of the loss, and thus the ML practitioner chooses a simpler hypothesis space, leading to larger Rashomon ratios. However, what happens if the hypothesis space remains fixed? In this case, will the Rashomon ratio increase in size? The answer to this question is addressed in Conjecture 35. More specifically, in Conjecture 35, we discuss that the Rashomon set (and correspondingly the Rashomon ratio, since the hypothesis space is fixed) increases with margin noise.

Conjecture 35 (The Rashomon set can increase with margin noise)**.** *Consider data distribution $\mathcal{D} = \mathcal{X} \times \mathcal{Y}$, where, $\mathcal{X} \in \mathbb{R}$, $\mathcal{Y} \in \{-1, 1\}$, and classes are balanced $P(Y = -1) = P(Y = 1)$ and generated by Gaussian distributions $P(X|Y = -1) = \mathcal{N}(\mu_1, \sigma^2)$, $P(X|Y = 1) = \mathcal{N}(\mu_2, \sigma^2)$, where $0 \leq \mu_1 < \mu_2$. For the hypothesis space $\mathcal{F} = \{f : f \in (\beta_1, \beta_2)\}$, where $(\mu_1, \mu_2) \subset (\beta_1, \beta_2)$, $\beta_1 \ll \mu_1$, and $\mu_2 \ll \beta_2$, and the Rashomon parameter $\gamma > 0$:*

(I) The volume of the Rashomon set is $\mathcal{V}(R_{set_\sigma}(\mathcal{F}, \gamma)) = |f^{e_1}_\sigma - f^{e_2}_\sigma|$, where $f^{e_1}_\sigma$ and $f^{e_2}_\sigma$



$\mathcal{V}(R_{set_\sigma}(\mathcal{F},\gamma))$
$f_\sigma^{e_1}$
$f_\sigma^{e_2}$
$G(f_\sigma^{e_2},\sigma)$
$\mathcal{N}(\mu_1,\sigma)$
$\mathcal{N}(\mu_2,\sigma)$
f^*

FIGURE 4.6: The setup for Conjecture 35. We show two Gaussians $\mathcal{N}(\mu_1, \sigma)$ and $\mathcal{N}(\mu 2, \sigma)$, with optimal model $f^* = \frac{\mu_1+\mu_2}{2}$ (shown in red). Models $f_\sigma^{e_1}$ and $f_\sigma^{e_2}$ (shown in blue) correspond to the left and right edges of the Rashomon set (shown in purple). The loss of model $f_\sigma^{e_2}$ is computed as the sum of two Gaussian tails and is equal to the area of the green region. The objective $G(f_\sigma^{e_2}, \sigma) = L(f_\sigma^{e_2}) - L(f^*)$ corresponds to the area of the shaded region to the left of the function $f_\sigma^{e_2}$.

are the two solutions to Eqn. (4.6), *where* Φ *is the CDF of the standard normal:*

$$2\Phi\left(\frac{\mu_2-\mu_1}{2\sigma}\right) - \Phi\left(\frac{\mu_2-f}{\sigma}\right) - \Phi\left(\frac{f-\mu_1}{\sigma}\right) = \gamma. \tag{4.6}$$

(II) We conjecture that[1] *for* $\mathcal{F} = \{f : f \in (\mu_1, \mu_2)\}$, *as we add feature noise to the data set by increasing the standard deviation* σ, *for all* σ *such that* $\sigma > \tilde{\sigma} = \frac{\mu_2-\mu_1}{2\sqrt{2}}$, *the volume of the Rashomon set increases as a function of* σ.

(III) Consider the setting where $\sigma = 1$ *for both Gaussians, and we add or remove noise by moving the means* μ_1 *and* μ_2 *of the Gaussians towards or away from each other. For any* $\gamma > 0$, *the volume of the Rashomon set is minimized when* $\mu_2 \approx \mu_1 + 2$. *Moving the Gaussians either away from or towards each other increases the volume of the Rashomon set.*

Proof. Without loss of generality, we will assume that $\mu_1 = 0$ and $\mu_2 = \mu > 0$. To get results for the original values μ_1 and μ_2, we can simply add μ_1 to μ and f.

Let us show the first point of the conjecture and show how to compute the volume of the Rashomon set.

[1] The hypothesis space for Part II conservatively includes all reasonable candidates for the empirical risk minimizer. In other words, we assume that the decision boundary can be anywhere between the means of the two distributions.

Evidence for Part (I) Denote ϕ_1 and ϕ_2 as probability density functions (PDF) and Φ_1 and Φ_2 as cumulative distribution functions (CDF) of classes $Y = -1$ and $Y = 1$ correspondingly. For a given model $f \in \mathcal{F}$, the loss can be computed as the sum of areas under the PDF corresponding to misclassification errors:

$$\begin{aligned} L(f) &= P(X > f|Y = -1) + P(X \leq f|Y = 1) \\ &= \int_f^\infty \phi_1(t)dt + \int_{-\infty}^f \phi_2(t)dt = 1 - \Phi_1(f) + \Phi_2(f) \\ &= 1 - \Phi\left(\frac{f-0}{\sigma}\right) + \Phi\left(\frac{f-\mu}{\sigma}\right) \\ &= 2 - \Phi\left(\frac{f}{\sigma}\right) - \Phi\left(\frac{\mu-f}{\sigma}\right), \end{aligned} \tag{4.7}$$

where Φ is the CDF of the normal distribution $\mathcal{N}(0,1)$.

The optimal model, f^*, can be obtained when $P(X|Y=-1) = P(X|Y=1)$, meaning that $\frac{1}{\sigma\sqrt{2\pi}} \exp\left(\frac{-(x-0)^2}{2\sigma^2}\right) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(\frac{-(x-\mu)^2}{2\sigma^2}\right)$, which leads to $f^* = \frac{\mu}{2}$. The loss of the optimal model is:

$$L(f^*) = 2 - 2\Phi\left(\frac{\mu}{2\sigma}\right).$$

Denote $G(f, \sigma)$ as the difference in loss between a model f from the Rashomon set and optimal model f^*,

$$G(f,\sigma) := L(f) - L(f^*) = 2\Phi\left(\frac{\mu}{2\sigma}\right) - \Phi\left(\frac{\mu-f}{\sigma}\right) - \Phi\left(\frac{f}{\sigma}\right).$$

To find a model on the edge of the true Rashomon set, we set $G(f,\sigma) = \gamma$ and obtain Equation (4.6). As the problem is symmetric with respect to f^* and $\gamma > 0$, solutions $f_\sigma^{e_1}$ and $f_\sigma^{e_2}$ to Equation (4.6) correspond to the left and right edges of the true Rashomon set. Thus, for a given γ and σ, we can estimate the volume of the Rashomon set as $\mathcal{V}(R_{set_\sigma}(\mathcal{F},\gamma)) = |f_\sigma^{e_1} - f_\sigma^{e_2}|$.

Now we will show evidence for the second point of the conjecture that the volume of the Rashomon set increases when σ increases.

Evidence for Part (II) Let $\tilde{\sigma} = \frac{\mu}{2\sqrt{2}}$. Let's focus on the right edge of the true Rashomon set where $f \in (f^*, \mu)$. We will show that:

(i) For any fixed σ, and for $f \in (f^*, \mu)$, $G(f, \sigma)$ is monotonically increasing in f.

(ii) For any fixed $f \in (f^*, \mu)$ and $\sigma > \tilde{\sigma}$, $G(f, \sigma)$ is monotonically decreasing in σ.

(iii) If (i) and (ii) hold, then for any σ_1 and σ_2 such that $\tilde{\sigma} < \sigma_1 < \sigma_2$, we have that $\mathcal{V}(R_{set_{\sigma_1}}(\mathcal{F}, \gamma)) < \mathcal{V}(R_{set_{\sigma_2}}(\mathcal{F}, \gamma))$.

We will prove (i) and (iii), and conjecture that (ii) is true, based on numerical experiments.

Let's focus on (iii) first. Consider σ_1 and σ_2 such that $\tilde{\sigma} < \sigma_1 < \sigma_2$. Let $f^e_{\sigma_1}, f^e_{\sigma_2} \in (f^*, \mu)$ be the right edge models of the corresponding true Rashomon sets $G(f^e_{\sigma_1}, \sigma_1) = \gamma$ and $G(f^e_{\sigma_2}, \sigma_2) = \gamma$. The volume of the Rashomon set is $\mathcal{V}(R_{set_\sigma}(\mathcal{F}, \gamma)) = |f^{e_1}_\sigma - f^{e_2}_\sigma| = 2|f^{e_2}_\sigma - f^*| = 2|f^e_\sigma - f^*|$. Using monotonicity of $G(f, \sigma)$ with respect to σ given fixed f (the result of part (ii)), we get that:

$$G(f^e_{\sigma_1}, \sigma_2) < G(f^e_{\sigma_1}, \sigma_1) = \gamma = G(f^e_{\sigma_2}, \sigma_2),$$

which means that $f^e_{\sigma_1} < f^e_{\sigma_2}$ due to monotonicity of $G(f, \sigma)$ with respect to f given fixed σ (the result of part (i)). Therefore, as we increase the standard deviation from σ_1 to σ_2, the Rashomon set increases as well:

$$\mathcal{V}(R_{set_{\sigma_1}}(\mathcal{F}, \gamma)) = 2|f^e_{\sigma_1} - f^*| < 2|f^e_{\sigma_2} - f^*| = \mathcal{V}(R_{set_{\sigma_2}}(\mathcal{F}, \gamma)).$$

Thus, we have proved part (iii).

Now, let us prove (i). Given fixed σ, the derivative of the objective $G(f, \sigma)$ with respect to f is:

$$\begin{aligned} G'_f(f, \sigma) &= \frac{1}{\sigma}\Phi'_f\left(\frac{\mu - f}{\sigma}\right) - \frac{1}{\sigma}\Phi'_f\left(\frac{f}{\sigma}\right) \\ &= \frac{1}{\sqrt{2\pi}\sigma}\left(e^{-\frac{(\mu-f)^2}{2\sigma^2}} - e^{-\frac{(f)^2}{2\sigma^2}}\right) > 0, \end{aligned}$$

since $\mu - f < f$ since $f > f^* = \frac{\mu}{2}$. Since the derivative $G'_f(f, \sigma) > 0$, the objective $G(f, \sigma)$ monotonically increases in f.

Finally, let's show (ii). Given fixed $f \in (f^*, \mu)$ and $\sigma > \tilde{\sigma}$, the derivative of the objective with respect to σ is:

$$
\begin{aligned}
G'_\sigma(f,\sigma) &= \frac{-2(\mu)}{2\sqrt{2\pi\sigma^2}}\Phi'_\sigma\left(\frac{\mu}{2\sigma}\right) + \frac{\mu - f}{\sqrt{2\pi\sigma^2}}\Phi'_\sigma\left(\frac{\mu - f}{\sigma}\right) + \frac{f}{\sqrt{2\pi\sigma^2}}\Phi'_\sigma\left(\frac{f}{\sigma}\right) \\
&= \frac{1}{\sqrt{2\pi\sigma^2}}\left[-\mu e^{-\frac{\mu^2}{8\sigma^2}} + (\mu - f)e^{-\frac{(\mu-f)^2}{2\sigma^2}} + f e^{-\frac{f^2}{2\sigma^2}}\right] \\
&= \frac{f}{\sqrt{2\pi\sigma^2}}\left[-\frac{\mu}{f}\left(e^{\frac{f^2}{2\sigma^2}}\right)^{\frac{-(\mu/f)^2}{4}} + \left(\frac{\mu}{f} - 1\right)\left(e^{\frac{f^2}{2\sigma^2}}\right)^{-((\mu/f)-1)^2}\right. \\
&\left.+\left(e^{\frac{f^2}{2\sigma^2}}\right)^{-1}\right].
\end{aligned}
$$

Denote $u = \frac{\mu}{f}$. Since, $\frac{\mu}{2} < f < \mu$, then $u \in (1,2)$. Denote $a = e^{\frac{f^2}{2\sigma^2}}$, then $\sigma^2 = \frac{f^2}{2\log(a)}$. Note that $a > 1$, and since $\sigma > \tilde{\sigma} = \frac{\mu}{2\sqrt{2}}$, $a < e^{\frac{4f^2}{\mu^2}} = e^{\frac{4}{u^2}} = s(u)$. Then $G'_\sigma(f,\sigma)$ can be expressed in terms of parameter u and variable a:

$$
\begin{aligned}
G'_\sigma(u,a) &= \frac{2\log(a)}{\sqrt{2\pi}f}\left[-ua^{\frac{-u^2}{4}} + (u-1)\,a^{-(u-1)^2} + a^{-1}\right] \\
&= \frac{2\log(a)}{\sqrt{2\pi}f}D(u,a).
\end{aligned}
$$

As $a > 1$ and $f > \frac{\mu}{2} > 0$, $\frac{2\log(a)}{\sqrt{2\pi}f} > 0$. To show that $D(u,a) < 0$ for any $u \in (1,2)$ and any $a \in (1, s(u))$, we perform exhaustive numerical calculations spanning the possible values of $u \in (1,2)$ and $a \in (1, s(u))$ in Figure B.1 in Appendix B.4. Indeed, $D(u,a) = 0$ when $u = 2$ or $a = 1$, and for all other values of u and a, $D(u,a) < 0$. Therefore, $G'_\sigma(f,\sigma) = G'_\sigma(u,a) < 0$. Since the derivative is negative, when $\sigma > \tilde{\sigma}$ the objective $G(f,\sigma)$ monotonically decreases in σ, which concludes our evidence for (ii), and thus part (II), of the conjecture.

Evidence for Part (III) Now we add or remove noise from the data set by moving the two means closer together or further apart (see Figure 4.7), for a fixed $\sigma = 1$. Recall that without a loss of generality, we take $\mu_1 = 0$ and denote $\mu = \mu_2$. The optimal model

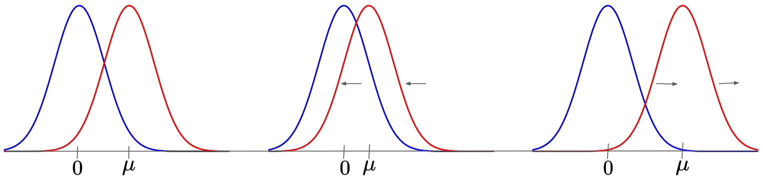


FIGURE 4.7: In part (III), we add or remove noise from the data by moving the mean of the right Gaussian. Both Gaussians have a standard deviation of 1.

f^* is $f^* = \frac{\mu}{2}$. Given $\sigma = 1$ and $\gamma > 0$, we are interested in finding $\mu > 0$ and edge model $f \in (0, \mu/2)$ such that the volume of the Rashomon set is minimal. Therefore we have the following optimization problem:

$$\begin{aligned} &\min_{\mu} \frac{\mu}{2} - f_\mu \text{ s.t. } f_\mu \text{ is defined by} \\ &2\Phi\left(\frac{\mu}{2}\right) - \Phi(\mu - f_\mu) - \Phi(f_\mu) = \gamma. \end{aligned} \tag{4.8}$$

We cannot solve optimization problem (4.8) directly for the best μ for each γ, but we provide numerical solutions in Figure 4.8 for a range of values of γ. We observe that, regardless of the value of $\gamma > 0$, as we minimize the volume of the Rashomon set to find μ, the μ corresponding to the optimal solution is always approximately equal to 2. The edge model $f^e \in (0, 1)$ is then the one that satisfies the constraint: $2\Phi(1) - \Phi(2 - f^e) - \Phi(f^e) = \gamma$. (The edge model must vary with γ since the optimal μ does not.)

The value $\mu = 2$ might seem surprising as a solution for *any* $\gamma > 0$. However, when $\mu_1 = 0$ and $\mu_2 = 2$, $f^* = 1$, which is one standard deviation away from each mean and therefore corresponds to a value where the inflection points of the two Gaussians coincide (normal distribution $\mathcal{N}(\mu, \sigma)$ has two inflection points $\mu \pm \sigma$). Given fixed γ, as we move either of the means of the Gaussians so that the inflection points no longer coincide, f^e moves outward to compensate so that the constraint in optimization problem (4.8) is satisfied. Therefore, the volume of the Rashomon set grows as we increase or decrease feature noise by moving μ away from 1 (in either direction) as shown in Figure 4.9.

■

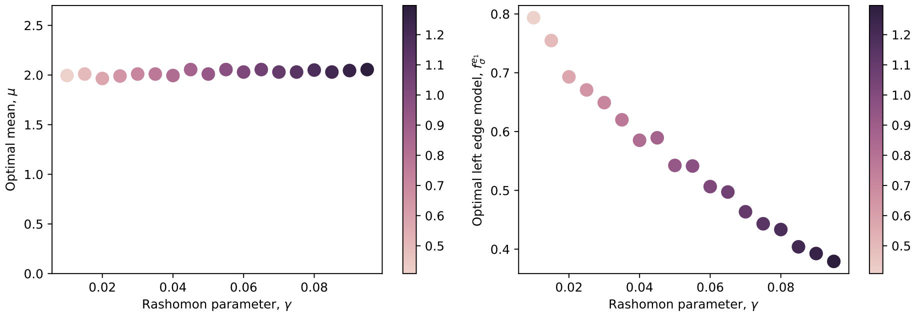


FIGURE 4.8: Numerical solution to the optimization problem (4.8). For each fixed γ, we plot $f_\sigma^{e_1} = f^e \in (0, \mu/2)$, such that the volume of the Rashomon set is minimized. The color of the scatter plot points corresponds to the value of the volume of the Rashomon set, where more intense color means higher value.

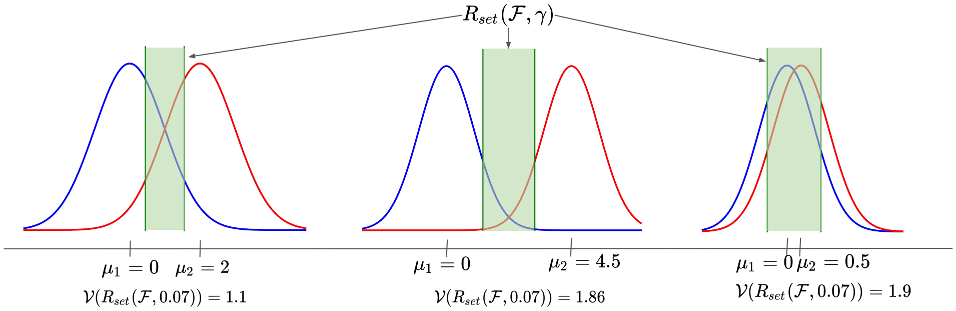

$R_{set}(\mathcal{F},\gamma)$
$\mu_1 = 0$
$\mu_2 = 2$
$\mathcal{V}(R_{set}(\mathcal{F}, 0.07)) = 1.1$
$\mu_1 = 0$
$\mu_2 = 4.5$
$\mathcal{V}(R_{set}(\mathcal{F}, 0.07)) = 1.86$
$\mu_1 = 0$
$\mu_2 = 0.5$
$\mathcal{V}(R_{set}(\mathcal{F}, 0.07)) = 1.9$

FIGURE 4.9: An example that shows that in part (III) as we move the right Gaussian away from a mean of 2 in either direction, the volume of the Rashomon set increases.

This conjecture is not a theorem because there is no analytical solution to the minimizer of the volume of the Rashomon set; the calculations are quite complex, involving differences of the CDF values of different Gaussians. However, *all parts of the conjecture have been fully checked numerically.* In part (II), we use an analytical derivation and exhaustive numerical computations to show that the derivatives of the left side of Eqn. (4.6) are either positive or negative sign. For part (III), we transpose the left Gaussian to $N(0, 1)$ to form a canonical problem in which all possible solutions can be computed numerically. We exhaustively search over the range of γ, finding the optimal μ_2 and volume of the Rashomon set for each γ. We find that μ_2 is very close to 2 for all γ.

This conjecture suggests that ***data that are approximately distributed according to two normal distributions, where the positive and negative normal distributions***

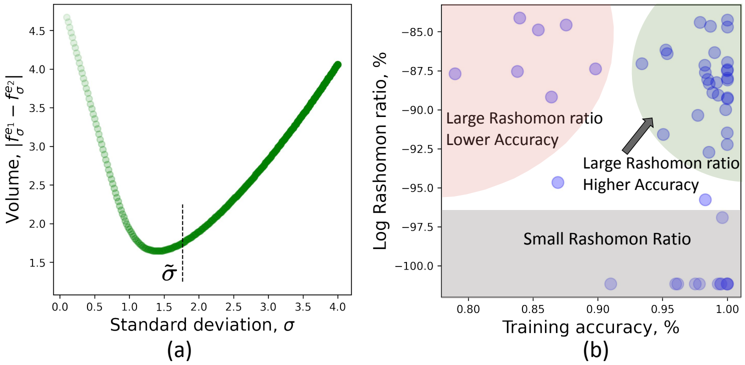


FIGURE 4.10: (a) Dependence of the Rashomon ratio on noise σ for the two Gaussians example in Conjecture 35. When $\sigma > \tilde{\sigma}$, as we add more noise, the size of the Rashomon set increases. (2). (b) Scatter plot of the log of Rashomon ratio versus the maximum accuracy across five different algorithms for 38 data sets in Section 3.2. We observe larger Rashomon ratios in both noisy and non-noisy data sets.

substantially overlap, will have a large Rashomon set. Figure 4.10(a) shows the dependence of the Rashomon set on the noise level σ for $\mu_1 = 1, \mu_2 = 6$ and $\sigma \in [0.2, 4]$. Figure 4.10(b) plots maximum accuracy versus Rashomon ratio for 38 data sets considered in Section 3.2. These figures indicate that large Rashomon sets occur both in noisy and non-noisy data.

4.3.2 Label Noise is Likely to Increase Pattern Diversity

Let S_ρ be a version of S with uniformly random label noise, creating randomly perturbed labels $\tilde{y}$ with probability $0 < \rho < \frac{1}{2}$: $P(\tilde{y}_i \neq y_i) = \rho$. Let $\Omega(S_\rho)$ be the uniform distribution over all S_ρ. From Theorem 16 denote the upper bound on pattern diversity as $U_{div}(\hat{R}_{set}(\mathcal{F}, \theta)) = 2(\hat{L}(\hat{f}) + \theta)(1 - (\hat{L}(\hat{f}) + \theta)) + 2\theta$. We show that it increases with uniform label noise.

Theorem 36 (Upper bound on pattern diversity increases with label noise)**.** *Consider a hypothesis space $\mathcal{F}$, 0-1 loss, and a dataset S. Let $\rho \in (0, \frac{1}{2})$ be the probability with which each label y_i is flipped independently, and let $S_\rho \sim \Omega(S_\rho)$ denote a noisy version of S. For the Rashomon parameter $\theta \geq 0$, if $\inf_{f \in \mathcal{F}} \mathbb{E}_{S_\rho} \hat{L}_{S_\rho}(f) < \frac{1}{2} - \theta$ and $\hat{L}_S(\hat{f}_S) < \frac{1}{2}$, then adding noise to the dataset increases the upper bound on pattern diversity of the expected Rashomon* *set:*

$$U_{div}(\hat{R}_{set_S}(\mathcal{F},\theta)) < U_{div}(\hat{R}_{set_{\mathbb{E}_{S_\rho \sim \Omega(S_\rho)}S_\rho}}(\mathcal{F},\theta)).$$

Proof. Given the noise model, hypothesis f is in the expected Rashomon set if $\mathbb{E}_{S_\rho}\hat{L}_{S_\rho}(f) \leq \inf_{f\in\mathcal{F}}\mathbb{E}_{S_\rho}\hat{L}_{S_\rho}(f)+\theta$. Let $\bar{f}\in\mathcal{F}$ be such that $\bar{f}\in\arg\inf_{f\in F}\mathbb{E}_{S_\rho}\hat{L}_{S_\rho}(f)$. Since $\rho\in(0,\frac{1}{2})$, and $\hat{L}_S(\hat{f}_S) < \frac{1}{2}$ by assumption, from (4.1) and the definition of the empirical risk minimizer, we have that:

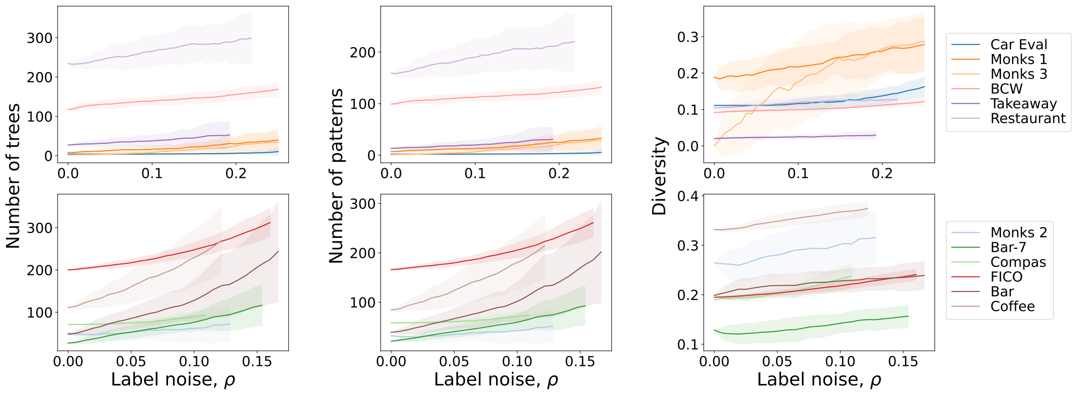
$$\begin{aligned}\inf_{f\in\mathcal{F}}\mathbb{E}_{S_\rho}\hat{L}_{S_\rho}(f) &= \mathbb{E}_{S_\rho}\hat{L}_{S_\rho}(\bar{f})\\ &= (1-2\rho)\hat{L}_S(\bar{f})+\rho\\ &\geq (1-2\rho)\hat{L}_S(\hat{f}_S)+\rho\\ &> \hat{L}_S(\hat{f}_S).\end{aligned}$$

Consider $g(x) = 2(x+\theta)(1-x-\theta)+2\theta$. For $x\in[0,\frac{1}{2}-\theta)$, $g(x)$ is monotonically increasing, as $g'(x) = 2(1-x-\theta)-2(x+\theta) = 4\left(\frac{1}{2}-x-\theta\right) > 0$. Given monotonicity, assumption of the theorem $\inf_{f\in\mathcal{F}}\mathbb{E}_{S_\rho}\hat{L}_{S_\rho}(f) < \frac{1}{2}-\theta$, and since $\hat{L}_S(\hat{f}_S) < \inf_{f\in\mathcal{F}}\mathbb{E}_{S_\rho}\hat{L}_{S_\rho}(f)$, we have that

$$\begin{aligned}U_{div}(\hat{R}_{set_S}(\mathcal{F},\theta)) &= 2\left(\hat{L}_S(\hat{f}_S)+\theta\right)\left(1-\hat{L}_S(\hat{f}_S)-\theta\right)+2\theta\\ &< 2\left(\inf_{f\in\mathcal{F}}\mathbb{E}_{S_\rho}\hat{L}_{S_\rho}(f)+\theta\right)\left(1-\inf_{f\in\mathcal{F}}\mathbb{E}_{S_\rho}\hat{L}_{S_\rho}(f)-\theta\right)+2\theta\\ &= U_{div}(\hat{R}_{set_{\mathbb{E}_{S_\rho}S_\rho}}(\mathcal{F},\theta)).\end{aligned}$$

■

In the general case, it is challenging to find a closed-form formula for pattern diversity or design a lower bound without strong assumptions about the data distribution or the hypothesis space. Therefore, we empirically examine the behavior of pattern diversity alongside other characteristics of the Rashomon set for different datasets and show these characteristics tend to increase with more noise.



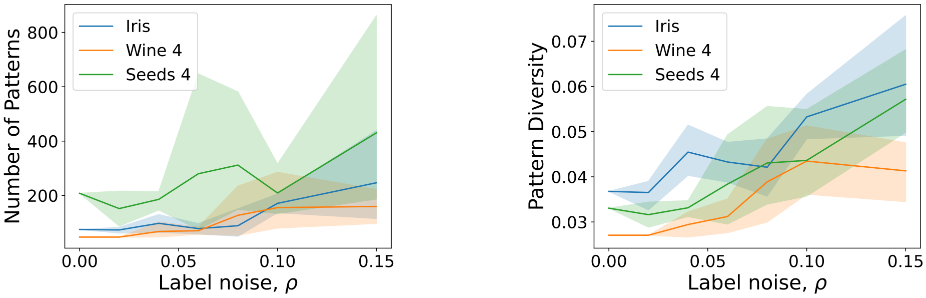
(a) (b) (c)

FIGURE 4.11: Rashomon set characteristics such as the number of trees in the Rashomon set (Subfigure a), the number of patterns in the Rashomon set (Subfigure b), and pattern diversity (Subfigure c) tend to increase with uniform label noise for hypothesis spaces of sparse decision trees. For readability, the top row of the figure shows datasets with lower empirical risk and the bottom row shows datasets with higher empirical risk.

4.3.3 Experiments for Rashomon Set Characteristics and Label Noise

We expect many different datasets to have larger Rashomon set measurements as data become more noisy. As before, we consider uniform label noise, where each label is flipped independently with probability ρ. For different noise levels, we computed diversity, and the number of patterns in the Rashomon set for 12 different datasets for the hypothesis space of sparse decision trees of depth 3 (note that we underfitted for some of the datasets); see Figure 4.11 (a)-(b). For every dataset, we introduced up to 25% label noise ($\rho \in [0, 0.25]$), or Accuracy($\hat{f}$) $-$ 50%, whichever is smaller. This means that data with lower accuracy (before adding noise) will have shorter plots, since noise is along the horizontal axis of our plots in Figure 4.11. For each noise level ρ, we performed 50 draws of S_ρ, and for each draw S_ρ we recomputed the Rashomon set. For decision trees, we used TreeFARMS (Xin et al., 2022), which allows us to compute the number of trees in the Rashomon set. We set the Rashomon parameter to 5%, however, the results hold for other values of the Rashomon parameter (Figure 4.13(b)-(c)).

Some key observations from Figure 4.11: First, the number of trees and patterns in the Rashomon set on average increases with noise. This means that the Rashomon ratio and pattern Rashomon ratio increase with noise as well since the hypothesis space stays the same. Second, datasets, that initially had higher empirical risk (e.g., COMPAS, Coffee House, FICO) tend to have more models in the Rashomon set (and thus higher Rashomon ratios) as compared to datasets with lower empirical risk (Car Evaluation, Monks1, Monks3). Finally, pattern diversity, on average, tends to increase with noise for the majority of the datasets.



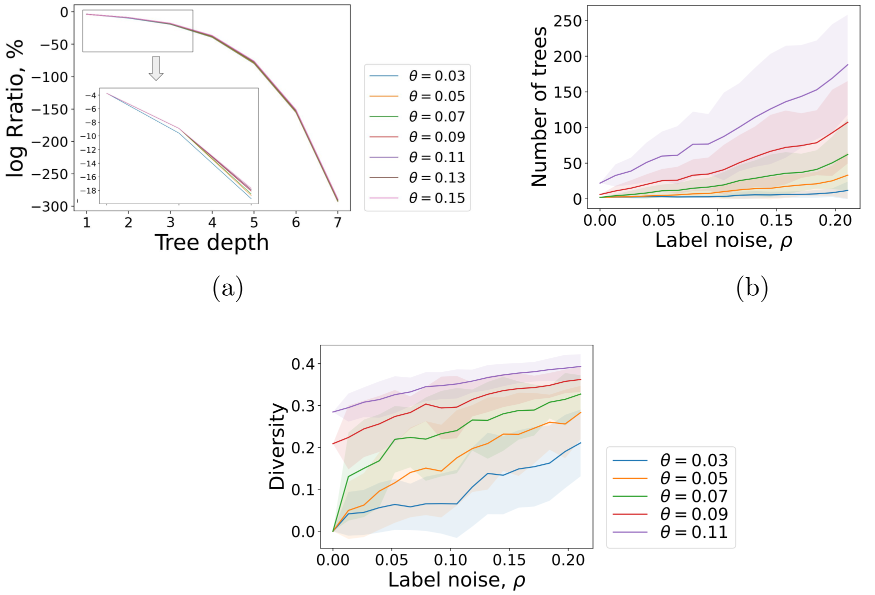
(a)

(b)

FIGURE 4.12: Rashomon set characteristics such as the number of patterns in the Rashomon set (Subfigure a) and pattern diversity (Subfigure b) tend to increase with uniform label noise for hypothesis spaces of linear models.

Note that the results in Figure 4.11 hold for unregularized hypothesis spaces. Since TreeFARMS has regularization, for more complex hypothesis spaces (such as decision trees of depth 5) and for larger amounts of noise, TreeFARMS's regularization on the number of leaves might prevent trees from fitting to this noise, resulting in trees of smaller depth. Thus, after reaching a certain point of adding noise, we might observe a decrease in metrics for some datasets. In this case, regularization acts along the path we described in Section 4.1. The hypothesis space of decision trees of depth 3 that we used in Figure 4.11 is simple enough for us to not see a decrease in depth due to regularization.

For the hypothesis space of linear classifiers, we show the pattern diversity and the number of patterns in the Rashomon set for different datasets in the presence of noise in Figure 4.12. We considered uniform label noise, where each label is flipped independently with probability ρ. We set noise level ρ to values in $\{0, 0.02, 0.04, 0.06, 0.08, 0.10, 0.15\}$ and performed five draws of S_ρ for every noise level. We then computed the pattern Rashomon set for each draw using the method described in Appendix 2.4 and finally computed the pattern diversity. Both the number of patterns and pattern diversity tend to increase with label noise.



(a)

(b)

(c)

FIGURE 4.13: (a) Curve of the hypothesis space complexity vs. Rashomon ratio (as in Figure 4.5) stays the same shape for different Rashomon parameters for the Monks 3 dataset for the hypothesis space of sparse decision trees. (b, c) Rashomon characteristics tend to increase with uniform label noise for the hypothesis space of decision trees (as in Figure 4.11) for different Rashomon parameters for the Monks 3 dataset. For (b) and (c), we averaged over 25 iterations.

Returning to the Path. If the practitioner observes more noise in the data, it could be already the case that the Rashomon set is large. Then, there are many good models, among which simpler or interpretable models are likely to exist.

5. The Role of Simplicity and the Rashomon Effect for Informed Decision Making

In the previous chapters, we established a theoretical foundation that explains when and why simple-yet-accurate models exist. This foundation is built upon two key observations: larger Rashomon sets and noise in data generation processes. Here, using a complex biology dataset as an example, we illustrate how the framework explains and influences the decisions of data analysts. Specifically, we examine a dataset of people with HIV who have undergone antiretroviral therapy. Intending to discover patterns in the data, we additionally demonstrate how the choices of models and algorithms are supported by the theoretical framework described in Chapters 3 and 4. We then further illustrate the power of sparse-yet-accurate methods by presenting tree-based density estimation approaches in Section 5.2 that can be used for different categorical datasets. We begin by describing the data of patients with HIV.

5.1 Interpretable Machine Learning Approaches to Better Understand Viral Reservoir for People with HIV

The development of antiviral therapy (ART) has vastly improved morbidity and mortality associated with HIV infection. However, a long-lived proviral reservoir, which is a group of cells where HIV can persist in a latent state, precludes a cure of infection (Chun et al., 1997, 1998; Finzi et al., 1997; Wong et al., 1997). Furthermore, despite the antiviral therapy, people with HIV (PWH) are still at increased risk of morbidity and mortality related to sequelae of persistent immune activation. There could be multiple factors that contribute to this persistent immune activation, including residual immune pertubation from untreated HIV infection, persistent proviral expression during ART, and specific clinical risk factors for immune activation that are enriched amongst people with HIV. Therefore, understanding the interplay between the HIV reservoir and persistent immune activation during long-term ART is important to inform strategies for a cure.

In this section, we use data science methods to assess the relationship between the immune system and the HIV reservoir across a cohort of 115 people with HIV. Our results highlight machine learning approaches to identify otherwise imperceptible global patterns in high-parameter studies of the HIV reservoir. Additionally, our results corroborate recent findings in the HIV persistence field regarding selective total proviral decay and immune dynamics and highlight the complex immunologic signatures of HIV latency, see Chomont et al. (2009), Falcinelli et al. (2021), Gandhi et al. (2021), and Peluso et al. (2020).

5.1.1 Cohort Description and Feature Analysis

We considered a cohort of 115 people with HIV (PWH) infection for at least one year that were recruited from two clinical sites: 66 PWH were recruited at Duke University Medical Center and 49 PWH at the University of North Carolina at Chapel Hill. Patients had been receiving antiretroviral therapy (ART) for at least 0.9 years (median 9 years) (Table 5.1). The study cohort was 77 percent male and the median participant age was 45. The total Proviral DNA assay and flow cytometry analysis were performed resulting in the determination of a total of 142 parameters for the cohort (Falcinelli et al., 2023; Semenova, Wang, et al., 2023). These parameters include 133 immunophenotypic cell frequencies, the HIV reservoir parameter (frequency of total HIV DNA), and 8 clinical-demographic parameters. Clinical-demographic variables (Table 5.1) included age, biological sex, years of ART, estimated years of HIV infection before ART (here, we use categorical variables: years before ART=NA, years before ART<1, years before ART≥ 1), CD4 nadir, most recent CD4 T cell count, and race (Caucasian, African-American, other).

The immunophenotypic cell frequencies and clinical-demographic variables formed the set of features for our study. For the label, we binarized the reservoir metric (total reservoir size) into high (above median) or low (below median).

First, we decided to identify a set of the most valuable variables from which ML models could be built. To achieve this, we generated receiver operator (ROC) curves for all the clinical-demographic and immunophenotype parameters (Figure 5.1A). The area under the curve (AUC) of the ROC curve indicates the ability of the feature (immune parameter or clinical/demographic information) to correctly identify a participant as having qualitatively low or high total HIV DNA. A model that randomly guesses a high or low reservoir has an AUC of 0.5. The most effective individual immune markers for classification of high versus low total HIV DNA included %NKG2A+ CD4 T (AUC = 0.70), %PD-1+ Tn CD4 T (AUC = 0.68), and %CD38+/HLA-DR- CD8 T (AUC = 0.68). Overall, this approach allowed us to derive a ranked list of the most predictive immune parameters for each aspect of the HIV reservoir, and these highly ranked features were thus used for subsequent ML modeling.

Table 5.1: Participant demographic and clinical characteristics. For demographics and clinical information, we report percentages for categorical variables, medians, and [Q1, Q3] for real-value variables. ART is antiretroviral therapy. CD4 counts reported in cells/mm^3. Years of HIV has 1 missing value, Years of ART has 7, and CD4 Nadir has 3; consequently, these missing values are not included in median and quantiles computations. Years before ART means years of HIV infection before ART initiation.

	Percentage (count)	Median	[Q1, Q3]	[Min, Max]
Age		45	[37, 53]	[23,65]
Sex (% male)	76.52% (88)			
Race				
Black	60% (69)			
White	37.39% (43)			
Other	2.61% (3)			
Years of HIV		11	[7, 19.85]	[1, 33.6]
Years before ART$<$ 1	55.65% (64)			
Years before ART$\geq$ 1	38.26% (44)			
Years before ART=NA	6.09% (7)			
Years of ART		9	[5.23, 16.63]	[0.9,33.5]
Recent CD4 count		799	[624.5, 962]	[319, 1970]
CD4 Nadir		313.5	[163.25, 463.25]	[2, 1080]

5.1.2 Data Visualization with Sparse Decision Trees

Since the interaction of the immune system and the HIV reservoir is multifactorial, we hypothesized that models that consider multiple parameters simultaneously could more accurately describe the overall dataset, and provide insights regarding the biology of the HIV reservoir and the host immune system. To accomplish this, we employed a decision tree approach to visualize combinations of variables that classify participants as having high or low reservoir size. Decision tree visualization is an interpretable supervised approach and does not require post-hoc analysis.

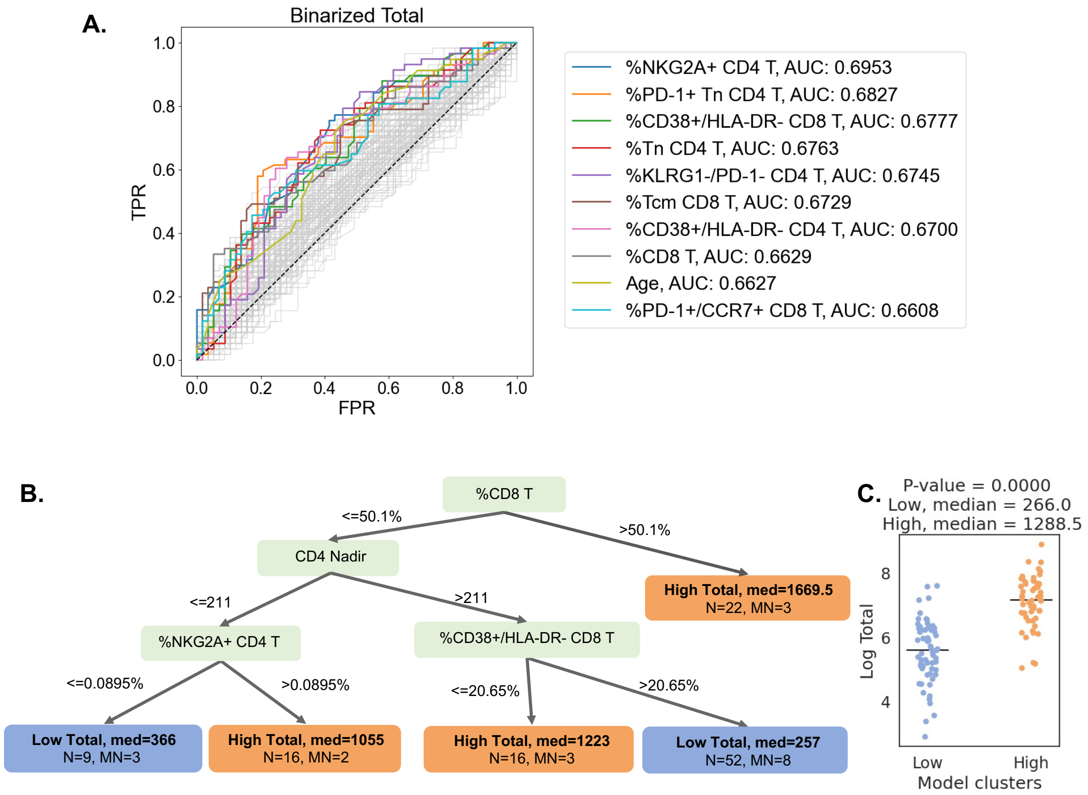


FIGURE 5.1: A. Receiver operating characteristic (ROC) curves identify PWH parameters that can classify reservoir characteristics. B. Decision tree visualization of the association of immune cell subsets with the reservoir characteristic. In each leaf, “med” denotes the median HIV characteristic of PWH, N is the number of PWH in the leaf, and MN is the number of mislabeled PWH. C. PWH in model leaves associated with high (orange) or low (blue) reservoir size characteristics were aggregated and a Mann-Whitney U test was performed to determine statistical significance between the actual total reservoir size of the “high” and “low” groups.

We first selected 35 variables with the highest ROC AUC values for total HIV DNA to be considered for model generation (Figure 5.1)A. Using these parameters, we fitted Generalized and Scalable Optimal Sparse Decision Trees (GOSDT) Lin et al., 2020 to the data. We required the trees to achieve at least 80% accuracy for classifying PWH in the cohort, as well as to have at least five PWH in each leaf. Since these trees are based on the entire dataset, these models are thus descriptive rather than predictive.

For the total HIV DNA decision tree, only four immune variables were required to accurately describe high versus low HIV DNA status (Figure 5.1B): CD8 T cell frequency, CD4 nadir, %CD38+HLA-DR- CD8 T cells, and %NKG2A+ CD4 T cells. The tree divided the cohort into five subgroups (leaves), among which three have a high total reservoir size and two have a low total reservoir size. Comparing the labels provided by the GOSDT model with the actual data, the tree achieved 83.5% accuracy (i.e. misclassifying 19 PWH among the overall cohort of 115 PWH). Notably, when we combined all samples from "high total reservoir" leaves and all samples from "low total reservoir" leaves, we observed a significant difference in the actual median total reservoir size for these two groups (266/M for low total and 1288.5/M for high total, Mann Whitney U test p-value is 3.56e-13, Figure 5.1C). This tree highlights combinations of immune parameters that can accurately describe qualitatively high versus low total HIV DNA in a cohort of n = 115 PWH. This visualization serves as a basis for mechanistic hypotheses about the interactions of the immune system and HIV reservoir during long-term ART.

Notably, *the Rashomon set for the hypothesis space of sparse decision trees of depth four and our dataset consists of 189673 models and 49519 patterns. This explains why we were able to accurately describe data with such a small tree (recall that we have 133 features in total), as per our theory in Chapter 3.* Recall that different models can realize the same pattern. We measured the Rashomon set using TreeFARMS (Xin et al., 2022), as we described in Section 2.3.4.

5.1.3 Machine Learning Models That Predict High versus Low Reservoir

Although clustering and decision tree analysis permit visualization and understanding of global structures within a dataset, we were curious if combinations of immune and clinical-demographic parameters could actually accurately predict (rather than only visualize) whether a given participant had high or low values of total HIV DNA. We considered five machine learning algorithms including Logistic Regression with L2 regularization (LR), CART, Support Vector Machines with RBF kernel (SVM), Random Forest (RF), and Gradient Boosted Trees (GBT). For reservoir size characteristics, we measured the accuracy of the models over 10 random splits of the data into training and test sets (Figure 5.2A).

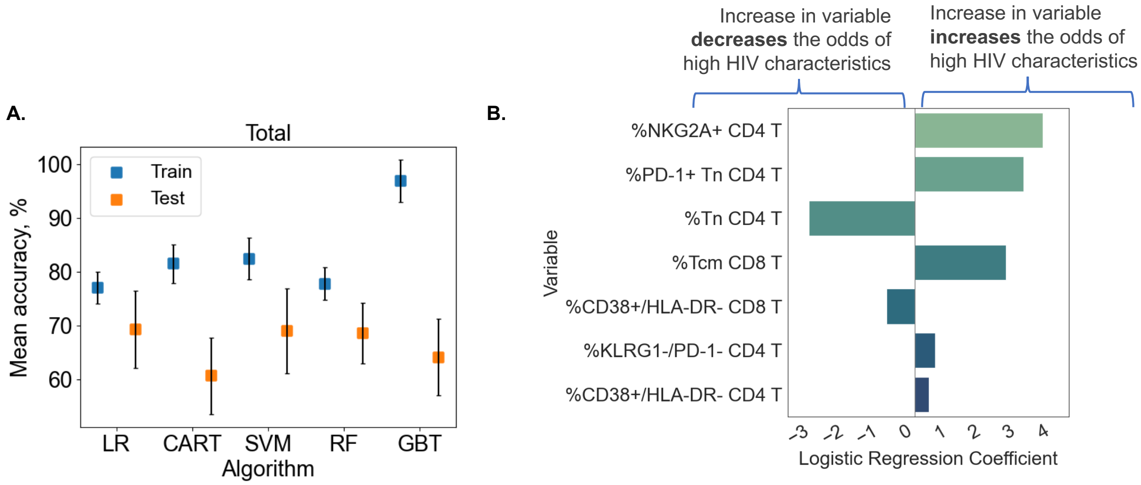


Figure 5.2: Predicting HIV reservoir characteristic with machine learning. A. Average training and test accuracies over ten training and test data split for Random Forest (RF), Gradient Boosted Trees (GBT), Support Vector Machines with RBF kernel (SVM), Logistic Regression (LR), and CART models for total reservoir size. B. For one split of training and test sets, the logistic regression model is visualized. On the y-axis, we show variables used by the model, while the x-axis displays coefficient values for individual variables used by the model.

Overall, we found that for total reservoir size, Logistic Regression achieved the highest mean classification accuracy (69.31%) in test data. *This is consistent with our findings in Chapter 4, where we provided mathematical proof that simpler models should be the default for data that rise from noisy data generation processes. Our data is cross-sectional and has a substantial amount of noise.*

To examine the contribution of individual immune features to model performance, we examined logistic regression coefficients for each immune cell variable in the model. Since Logistic Regression coefficients are associated with the expected change in log odds (based on $\log_e$), we can think about the coefficient β for variable X in the following way: increasing variable X by one unit multiplies the odds of high reservoir size (probability that the reservoir size is high divided by the probability that the reservoir size is low) by e^{β}.

In Figure 5.2B we visualize the logistic regression model for one data split among ten we considered for total reservoir size. For this split, we observe that higher values of %NKG2A+ CD4 T, %PD-1+ Tn CD4 T, and %Tcm CD8 T are associated with an increased probability of total reservoir size being high. On the other hand, an increase in %Tn CD4 T decreases the odds of high reservoir size. The model visualized in Figure 5.2B achieved 75.86% training and 75% test accuracy.

Our analysis demonstrates that we can use machine learning tools to construct models that can predict with approximately 70% accuracy whether a given PWH has qualitatively low or high total HIV DNA. The coefficients identified the model, including NKG2A expression on CD4 T cells, identify specific immunologic nodes that serve as a basis for the generation of hypotheses about the interplay between total versus total HIV DNA and the host of the immune system.

Inspired by the success of data visualization in Section 5.1.2, we further explore how simpler models can assist in decision-making by trying to understand the data. More specifically, we propose sparse tree-based methods for density estimation.

5.2 Sparse Density Trees and Lists for Categorical Datasets

5.2.1 Background and Related Work

In this section, we present sparse-density trees and rule lists. Our methods aim to globally optimize a Bayesian posterior possessing a sparsity prior, which acts as a regularization term. Bayesian tree models are commonly used for tasks other than density estimation (i.e., classification and regression). Some examples include Bayesian CART (Y. Wu et al., 2007), Bayesian Additive Regression Trees (BART) (Chipman et al., 2010), and Bayesian Rule Lists (Letham et al., 2015; H. Yang et al., 2017). Bayesian CART and BART use priors that specify the probability that a node is terminal and a uniform probability distribution over the choices for a split. Our priors function differently. We have priors over *global* properties of the trees such as the number of total leaves (our Method I and Method III). Also, we have prior parameters governing the number of branches at a node (our Method II), which is different from Bayesian CART or BART where there are only two branches at every node. Our rule list density approach (Method III) has a prior on the number of conditions used in each rule, which is similar to Bayesian Rule Lists, but not similar to Bayesian CART or BART, which have only one condition defining each split.

Unlike past work on density trees D. Li et al., 2016; Ram and Gray, 2011; K. Yang and Wong, 2014a, 2014b, including density estimation trees (DET) (Ram & Gray, 2011), our three methods are not constructed using greedy tree induction, they are optimized instead. DETs are built top-down in a greedy manner, which can result in lower-quality trees. DET was used by K. Wu et al. (2018), leveraging ideas from Lu et al. (2013) with random forests to perform density estimation. DET has also been applied to high energy physics (Anderlini, 2015). Techniques to avoid overfitting in tree-based density estimation models have been discussed by Anderlini (2016). In our methods, we optimize the splits instead of using a greedy approach and place prior directly on the shape of the trees and lists.

Some works (e.g., Q. Liu et al., 2021; Sasaki & Hyvärinen, 2018) use neural networks to perform density estimation, which do not aim to be interpretable. In Luo et al., 2019, a smoothing spline is used to perform density estimation. In Rehn et al., 2018, a non-parametric density estimator called "FRONT" segments a data stream through a periodically updated linear transformation. Other techniques have been proposed, including: nonparametric techniques (mainly variants of kernel density estimation) (Akaike, 1954; Cacoullos, 1966; Cattaneo et al., 2019; Devroye, 1991; Mahapatruni & Gray, 2011; Nadaraya, 1970; Parzen, 1962; Rejtö & Révész, 1973; Rosenblatt et al., 1956; Silverman, 1986; Varet et al., 2023; Wasserman, 2006), mixtures of Gaussians (Chen et al., 2006; J. Q. Li & Barron, 2000; Ormoneit & Tresp, 1996, 1998; Seidl et al., 2009; Zhuang et al., 1996), forest density estimation (H. Liu et al., 2011), RODEO (H. Liu et al., 2007), and nonparametric Bayesian methods (Müller & Quintana, 2004). However, all these methods either lack the interpretable logical structure found in histograms or do not focus on sparsity.

By placing the prior over the tree structure, we aim to improve issues with standard histograms in high dimensions. First, the density estimation models, produced by our methods, are easier to understand by following the paths in the tree of the list. Second, the prior promotes a sparse model, enhancing generalization. Finally, by encouraging sparsity, the prior aligns the model with a user-defined concept of interpretability Goh et al., 2024.

Our density estimation models can be useful in multiple domains to detect new patterns or errors in the data. For example, Figure 5.3 shows the sparse density tree for the COCO-Stuff (Caesar et al., 2018) labels. The data set contains 118k training samples over 91 stuff categories. While the labels are sparse, our method finds interesting combinations, such as *mirror* and *blanket*, or *railing*, *skyscraper*, and *ground* that are shown in Figure 5.4. In the next section, we provide an overview of priors and optimization methods of the posterior for our methods.

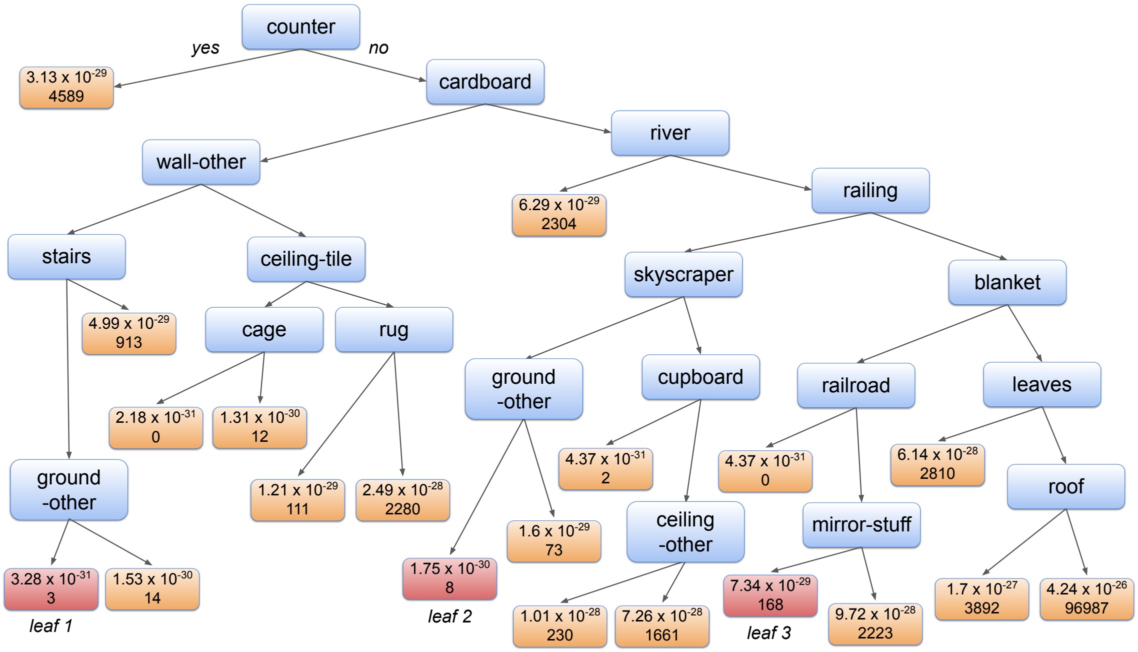
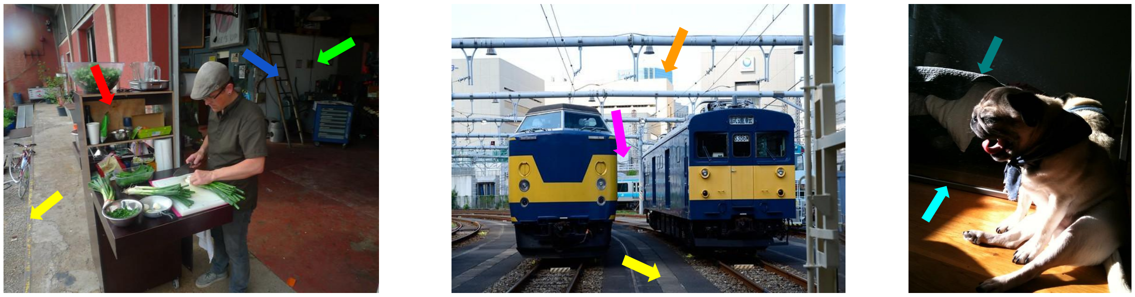


FIGURE 5.3: A sparse density tree to represent the COCO-stuff labels. Each leaf (orange or red color) shows the density and number of training points that belong to that leaf (since densities are small for large datasets such as COCO-stuff labels, the number of points might be easier to understand). Leaves 1, 2, and 3 are visualized in Figure 5.4. This tree contains 20 leaves. It took approximately 25 seconds to create the tree and around 6.4 minutes to run the validation process that tunes parameters and optimizes the tree for each parameter setting. Each run takes approximately 25 seconds; there are 5 repeats per parameter setting and 3 parameter settings.



(a) red – *cardboard*
green – *wall-other*
blue – *stairs*
yellow – *ground-other*

(b) magenta – *railing*
orange – *skyscraper*
yellow – *ground-other*

(c) teal – *blanket*
aqua – *mirror-stuff*

FIGURE 5.4: Examples of images that contain labels from leaf 1 (a), 2 (b) and 3 (c) in Figure 5.3.

5.2.2 Methods Description and Computational Optimization

We use a Bayesian approach to achieve sparsity, by introducing priors on the shape of the trees. In particular, in Method I, we define a prior on the number of leaves in the tree, then calculate the likelihood of the data having been generated by a particular tree, and multiply the prior and the likelihood to create a posterior to be optimized over all trees. In Method II, we instead choose a prior over the number of branches for each split in the tree, preferring a small number of branches. In Method III, we switch to rule lists, where the prior prefers models with a smaller number of rules and a smaller number of conjunctions per rule.

Before the introduction of the three methods, we first present notation. We will focus on problem of estimating the unknown distribution f with tree-structured approximations given a set of n data points $X = \{x_i, ..., x_n\}$ drawn i.i.d. from f on $\mathcal{X} \subset \mathbb{R}^p$. We will assume that our features are categorical. For the tree-structured estimations, we will describe each leaf l with a set of conditions (denote $\sigma_j(l)$ as a set of conditions on feature j) that occur alongside the path in the tree from the root to this leaf. For example, if we first split on feature $x_{\cdot 1}$ taking values "March, April, May", then $\sigma_1(l) = \{\text{March, April, May}\}$. Say we then split on feature $x_{\cdot 2}$ taking values "red, orange", then $\sigma_2(l) = \{\text{red, orange}\}$, and the leaf will be: $x \in \{x_{\cdot 1} \in \{\text{March, April, May}\}, x_{\cdot 2} \in \{\text{red, orange}\}\}$. Since we have p features in total, this means that there is no restriction on features $x_{\cdot 2}, .., x_{\cdot p}$ for this specific leaf l. $\sigma_j(l)$ includes all possible values for the feature j with no restrictions alongside the path. We compute the volume in a leaf l (which is required for computing the density) as a product of all cardinalities of $\sigma_j(l)$, meaning that $\mathbb{V}_l = \prod_{j=1}^{p} |\sigma_j(l)|$(Goh et al., 2024).

5.2.2.1 Overview of the Bayesian Approach for Density Tree Sparsity

Method I overview. For the leaf-sparse density tree method, the prior will define the user's desired number of leaves λ in the tree prior to seeing data. We denote the number of leaves in tree T as K_T. We chose Poisson prior centered at this user-defined parameter λ, $\text{Poisson}(K_T, \lambda)$. We assign a uniform prior over the probabilities associated with a data point landing in each of the leaves by using Dirichlet distribution with equal parameters $\boldsymbol{\alpha} = \{\alpha \dots \alpha\}$, where $\alpha \in \mathbb{Z}^+$ and $\alpha > 1$. The parameter α is a pseudocount that is typically chosen to be a small number (1 or 2) to avoid a 0 value for the estimated densities. We draw multinomial parameters $\boldsymbol{\theta} = [\theta_1, \dots, \theta_{K_T}]$ from $\text{Dir}(\boldsymbol{\alpha}_{K_T})$, which govern the prior on popularity of each leaf. The likelihood of the data to have arisen from a particular tree is $P(X|\boldsymbol{\theta}, T) = \prod_{l=1}^{K_T} \left(\frac{\theta_l}{\mathbb{V}_l}\right)^{n_l}$, where $\mathbb{V}_l$ is the volume of leaf l, and $\frac{\theta_l}{\mathbb{V}_l}$ is the probability to land at any specific value within leaf l. Please see Goh et al. (2024) for more details on prior, likelihood, and posterior (the likelihood times the prior). We choose the tree that maximizes the log-posterior.

As compared to full Bayesian approaches, by maximizing the posterior, we leverage relatively faster computation time and optimize for a single model, which can be important for interpretability. However, in turn, we miss out on uncertainty information, which we could get by modeling the posterior density.

Method II overview. For the branch-sparse density tree method, a Poisson distribution with parameter λ is used at each node to determine the number of branches, where the parameter λ is the desired number of branches set up by the user before seeing data. We denote the number of branches in tree T as $B_T = \{b_t\}$, where each b_t is the number of branches for every node t in the tree T. At each internal node t, we draw a sample from a Dirichlet distribution with parameter $[\alpha, \ldots, \alpha]$ (of size equal to the number of branches b_t) to determine the proportion of data that should go along the branch into each child node from its parent node. As before, α is a pseudo-count to avoid 0-valued estimated densities. The likelihood for this method is the same as for the leaf-sparse density tree method and the posterior is proportional to the prior times the likelihood terms. Please see Goh et al. (2024) for more details on prior, likelihood, and posterior for this method.

Method III overview. The sparse density rule list is a one-sided sparse tree. Each tree can be expressed as a rule list by creating a rule for each leaf, where the conditions defining the leaf also define the rule. An example of a sparse density rule list is as follows: **if** x obeys a_1 **then** density$(x) = f_1$ **else if** x obeys a_2 **then** density$(x) = f_2$ $\ldots$ **else if** x obeys a_m **then** density$(x) = f_m$ **else** density$(x) = f_0$. Here, as with the trees, the density is the probability mass function, which is constant for the entire portion of the feature space that falls into the leaf. The antecedents a_1,...,$a_m \in A$, where A is a set of pre-mined antecedents, are Boolean assertions, that are either true or false for each data point x_i. Let $a_{<j}$ be the antecedents before j in the rule list if there are any, and let c_j be their cardinality. We denote the rule list as d. Following the exposition of Letham et al., 2015, we use a prior over rule lists to encourage sparsity. The generative process is described in Algorithm 2 (Goh et al., 2024). It depends on the input prior parameters λ, η, and $\boldsymbol{\alpha}$. Prior parameter λ is the user's preference for the length of the sparse density list before seeing the data, and η is the user's preference for the number of conjunctions in each a_j. As before, usually all elements in $\boldsymbol{\alpha}$ are the same and indicate pseudo-counts. The distribution of m is the truncated at $|A|$ Poisson distribution; $P(c_j|c_{<j}, A, \eta)$ is a Poisson truncated to remove values for which no rules are available with that cardinality; distribution over antecedents in A of size c_j (excluding those in a_j) is uniform; the prior distribution over the leaves $\boldsymbol{\theta}$ is drawn from $Dir(\boldsymbol{\alpha})$. We use the same likelihood as for Method I, and the posterior is the product of likelihood and prior, see Goh et al. (2024) for more details on posterior and prior.

Algorithm 2 Density rule lists generation procedure

Input: Prior parameter λ and η, pseudo-count $\boldsymbol{\alpha}$, observations $X = \{x_i\}$, antecedents $A = \{a_j\}$

Output: Sparse density rule list

```
Sample a decision list length $m \sim P(m|A, \lambda)$
for decision list rule $j = 1, \dots, m$ do
    Sample the cardinality of antecedent $a_j$ in $d$ as $c_j \sim P(c_j|c_{<j}, A, \eta)$.
    Sample $a_j$ of cardinality $c_j$ from $P(a_j|a_{<j}, c_j, A)$.
end for
for observation $i = 1, \dots, n$ do
    Find the antecedent $a_j$ in $d$ that is the first that applies to $x_i$.
    If no antecedents in $d$ applies, set $j = 0$.
end for
Sample parameter $\boldsymbol{\theta} \sim$ Dirichlet $(\boldsymbol{\alpha})$ for the probability to be in each of the leaves
for each leaf $l$ in the rule list do
    Compute volume $V_l$
    Compute density $f_l = \frac{\theta_l}{\mathbb{V}_l}$
end for
```

5.2.2.2 Numerical Methods to Optimize the Objective Function

In the previous section, we have described the prior, likelihood, and posterior functions for three generative modeling methods (as before, see Goh et al. (2024) for more details). Since the search space of our problems is large, we use simulated annealing, a metaheuristic optimization algorithm, which allows us to approximate global solutions. More specifically, in this section, we describe a simulated annealing scheme that we implemented in order to find the optimal tree that maximizes the posterior for Method I and Method II as well as discuss Method III's optimization details.

Simulated annealing for tree-based methods. A successful simulated annealing scheme requires us to create a useful definition of a neighborhood. We define our neighborhood such that each move explores a neighboring tree where we are able to extend or shrink the tree.

At each iteration, we need to determine which neighboring tree to move to. To decide which neighbor to move to, we fix a parameter $\epsilon > 0$ beforehand, where ϵ is small, approximately 0.01 in our experiments. ϵ is the probability that we will perform a structural change to jump out of a possible local minimum. All other actions below are taken with equal probability. Thus, at each time, we generate a number from the uniform distribution on $(0, 1)$, then either:

1. (Shrink at a leaf) If the number is smaller than $\frac{1-\epsilon}{4}$, we select uniformly at random a "parent" node that has leaves as its children and remove its children. This is always possible unless the tree is the root node itself, in which case we cannot remove it and this step is skipped.

2. (Expand) If the random number is between $\frac{1-\epsilon}{4}$ and $\frac{1-\epsilon}{2}$, we pick a leaf randomly and a feature randomly. If it is possible to split on that feature, then we create children for that leaf. (If the feature has been used up by the leaf's ancestors, we cannot split, and we then skip this round.)

3. (Regroup) If the random number is between $\frac{1-\epsilon}{2}$ and $\frac{3(1-\epsilon)}{4}$, we pick a node randomly, delete its descendants, and split the node, creating two child nodes where the splitting is done on subsets of the node's feature values. Sometimes this is not possible, for example, if we pick a node where all the features have been used up by the node's ancestors, or if the node has only one category. In that case, we skip this step.

4. (Merge sibling nodes) If the random number is between $\frac{3(1-\epsilon)}{4}$ and $(1-\epsilon)$, we choose two nodes that share a common parent, delete all their descendants, and merge the two nodes (e.g., black, white, red, green becomes black-or-white, red, green).

5. (Structural change) If the random number is more than $1-\epsilon$, we perform a structural change operation where we remove all the children of a randomly chosen node of the tree.

Please see Algorithm 3 for the pseudo-code of the simulated annealing procedure. The last three actions avoid problems with local minima. The algorithms can be warm-started using solutions from other algorithms, e.g., DET trees. We found it useful to occasionally reset to the best tree encountered so far or the trivial root node tree.

Sparse Density Rule List Optimization. To search for optimal density rule lists that fit the data, we use local moves (adding rules, removing rules, and swapping rules) and use the Gelman-Rubin convergence diagnostic applied to the log posterior function.

A technical challenge that we need to address in our problem is the computation of the

Algorithm 3 Simulated annealing for tree-based methods

Input: Prior parameters, θ, maximum number of iterations, N, ϵ, cooling schedule Cool(iteration) for the simulated annealing

Output: Optimal density tree

```
Initialize the initial tree $T$ to be a single node, compute the posterior using the objective
  function and the prior parameters. Set the iteration number to be 0.
while iteration number $< N$ do
    Draw a random number $r$, from $Uni(0, 1)$.
    if $r < \frac{1-\epsilon}{4}$ then
        Perform shrink at leaf operation.
    else if $r < \frac{2(1-\epsilon)}{4}$ then
        Perform expand operation
    else if $r < \frac{3(1-\epsilon)}{4}$ then
        Perform regroup operation.
    else if $r < 1 - \epsilon$ then
        Perform merge sibling nodes operation.
    else
        Perform structural change.
    end if
    Compute the objective value for the modified tree.
    if a better objective value is obtained than the current best then
        Update $T$ to be the current tree.
    end if
    if the current tree is worse than the current best tree $T$ then
        With probability defined by the cooling schedule Cool(iteration), update $T$ to be
  the current tree. This will always be a small probability.
    end if
end while
Return $T$
```

volume of a leaf. Volume computation is not needed in the construction of a decision list classifier like that of Letham et al., 2015 but it is needed in the computation of a density list. There are multiple ways to compute the volume of a leaf of a density rule list.

Approach 1: Analytical Computation. Use the inclusion-exclusion principle to directly compute the volume of each leaf. Consider computing the volume of the i-th leaf in a density rule list. Let V_{a_i} denote the volume induced by the rule a_i, that is the number of points in the domain that satisfy a_i. To belong to that leaf, a data point has to satisfy a_i and not earlier rules $a_{<i}$. Hence the volume of the i-th leaf is equal to the volume obeying a_i alone, minus the volume that has been used by earlier rules. Using notation a_k^c to denote

the complement of condition a_k, we have the following:

$$\begin{aligned}
\mathbb{V}_i &= V_{a_i \wedge \bigwedge_{k=1}^{i-1} a_k^c} \quad (i \text{ is in } a_i \text{ and in the complement of all previous rules}) \\
&= V_{a_i} - V_{a_i \wedge (\bigvee_{k=1}^{i-1} a_k)} \\
&= V_{a_i} - V_{(\bigvee_{k=1}^{i-1} a_i \wedge a_k)} \\
&= V_{a_i} - \sum_{k=1}^{i-1} (-1)^{k+1} \sum_{1 \leq j_1 \leq \ldots j_k \leq n} V_{a_i \wedge a_{j_1} \ldots \wedge a_{j_k}}, \qquad (5.1)
\end{aligned}$$

where the last expression is due to the inclusion-exclusion principle and it only involves the volume resulting from conjunctions. The volume resulting from conjunctions can be easily computed from data. Without loss of generality, suppose we want to compute the volume of $V_{a_1 \wedge \ldots \wedge a_k}$. For each feature that appears, we examine if there is any contradiction between the rules; for example, if feature 1 is present in both a_1 and a_2, where rule a_1 specifies feature 1 to be 0 whereas a_2 specifies feature 1 to be 1, then we have found a contradiction and the volume of the intersection of a_1 and a_2 should be 0. If there is no contradiction, then the volume is equal to the product of the number of distinct categories of all the features that are not used. If all features are used, then the volume is 1. By using the inclusion-exclusion principle, we reduce the problem to just computing a volume of conjunctions as in (5.1). Note that for this approach, we still need to iterate over all conjunctions for each volume computation, which can be computationally expensive. *Approach 2: Uniform Sampling.* Create uniform data over the whole domain, and count the number of points that satisfy the antecedents. This approach would be expensive when the domain is huge but easy to implement for smaller problems.

Approach 3: MCMC. Use an MCMC sampling approach to sample the whole domain space. This approach is again not practical when the domain size is huge as the number of samples required will increase exponentially due to the curse of dimensionality.

We use the analytical computation approach 1 in our implementation.

5.2.3 Experiments and Empirical Performance Analysis

We considered two baselines to evaluate the effectiveness of our methods, including standard histograms and density estimation trees (DET) (K. Wu et al., 2018).

For the standard high-dimensional histogram baseline, we treated each possible set of feature values (e.g., $x_{.1} = 1, x_{.2} = 0, ..., x_{.10} = 1$) as a separate bin. (We call a set of feature values a *configuration*; it is a point in our feature space.) Histograms have the disadvantage that they create a large number of bins and thus may not generalize well to the test set; they are also not interpretable, since they cannot be visualized in a tree or list.

Algorithm 4 Conversion of categorical feature to real

```
Input: categorical feature $f_c$, $c_1, ..., c_J$ - categories of the feature $f_c$
Output: real-valued feature $f_r$
for every category $c_j$ do
    compute its frequency $p_j$ in the dataset
end for
sort the categories from the most frequent to the least frequent
split interval $[0, 1]$ in $J$ frequency intervals $I_j = [a_j, b_j]$ based on the cumulative probability, meaning that the length of the interval with index $j$ is $p_j$
for every sample $i$ do
    find its category $c_j^i$
    find corresponding frequency interval $I_j = [a_j, b_j]$
    chose a value $v \in I_j$ by sampling from a truncated Gaussian distribution with $\mu = a_j + p_j/2$ and $\sigma = (b_j - a_j)/6$
    assign $v$ to $f_r^i$
end for
```

The implementation of DET is meant for continuous variables, but we use it anyway for comparison. To compare our methods to DET, we pre-processed datasets using one-hot encoding and the Synthetic Data Vault algorithm from Patki et al., 2016 (see Algorithm 4). We also used our computations for the density in each leaf following the volume computations we described in the previous section. For DET hyperparameters, we tried $\{1, 3, 5\}$ for the minimum size of a leaf and $\left\{5, 10, 50, 100, \lfloor \frac{n}{10} \rfloor, \lfloor \frac{n}{5} \rfloor\right\}$ for the maximum size of a leaf, where n is the number of training data points. We used nested cross-validation over 5 folds. For each fold, we optimized parameters for validation log-likelihood and reported out-of-distribution log-likelihood. DET has the disadvantage of being a greedy algorithm and the available implementation of DET is not designed for categorical data (see Appendix B in Goh et al. (2024)), thus DET may not produce trees that are as useful or sparse as those from Methods I, II, or III.

For our Methods I, II, and II, we performed nested cross-validation over 5 folds and evaluated out-of-sample log-likelihood and sparsity of each method for every fold. For the leaf-sparse density tree model (Method I), the parameter to control the number of leaves was chosen from the set $\{5, 8, 10\}$, and α was set to 2, which corresponds to a pseudocount of 2 data points in each bin (to prevent bins with 0 data points). For the branch-sparse density tree model (Method II), the parameter to control the number of branches was chosen from the set $\{2, 3\}$, and α was set to 2. For the sparse density rule list (Method III), the parameters λ, η, and α were chosen among $[3, 5, 7], 1$, and 1 respectively. We provide a summary of parameters and their suggested values in Appendix C.2.

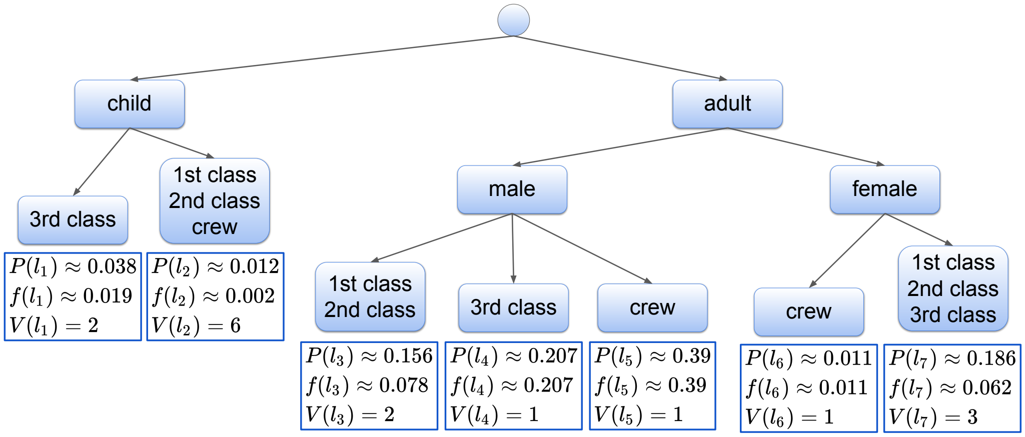
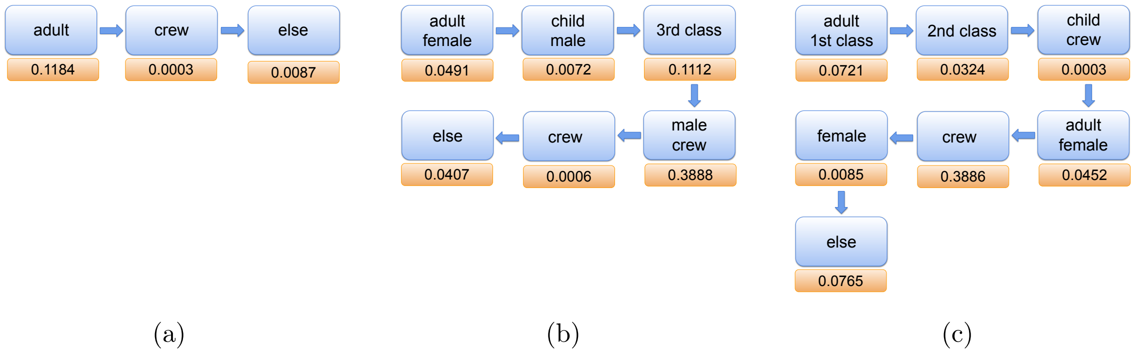


FIGURE 5.5: A sparse density tree to represent the Titanic dataset. The probability of belonging to the leaf (P), the densities (f), and the volume (V) are specified for each leaf of the sparse tree. The density is estimated to be constant within each leaf. Here, we can see that the volume times the density equals the probability of being in the leaf.

5.2.3.1 Titanic Dataset

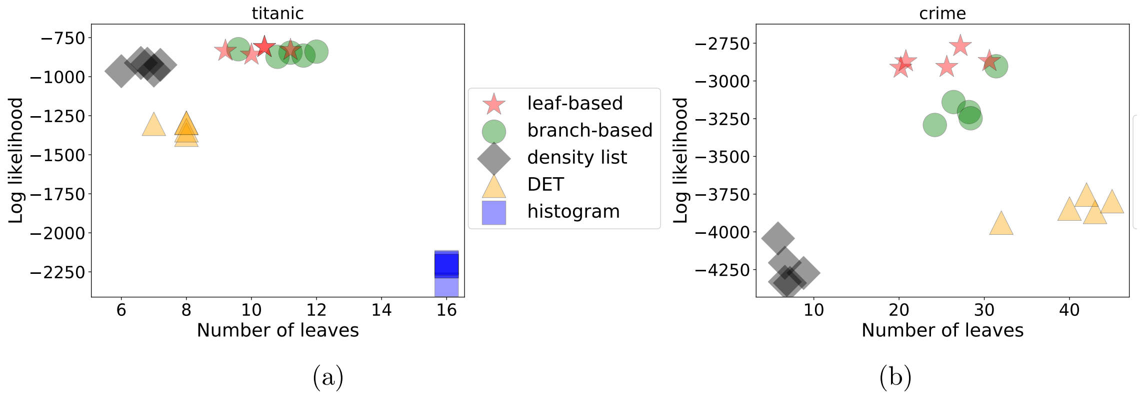
We will use a sparse density tree or list to help us understand the distribution of people on board the Titanic. The Titanic dataset has information for each of the 2201 people who



(a) (b) (c)

FIGURE 5.6: Density rule lists vary with different parameters λ, which indicate preferred list lengths. λ is set to 2 for (a), 4 for (b), and 7 for (c). Parameters η and α were chosen as 2 and 1 respectively. These density lists were chosen based on the maximum log-likelihood over 5 repeats. Each arrow represents an “else if" statement. E.g., for (b) if the passenger is an adult and female, then density is constant with respect to other variables at 0.0491, else if the passenger is a child and male, density is constant at 0.0072, etc.

were on the Titanic. It includes details like gender, whether someone is an adult, and the passenger’s class (first, second, or third class, or crew member).



(a) (b)

FIGURE 5.7: Performance comparison between our methods and baselines for Titanic (a), Crime (b) datasets.

For each method and each test fold, Figure 5.7 shows the out-of-sample log-likelihood (on the y-axis) and sparsity (on the x-axis). The histogram method had the most bins and thus tended to overfit. DET performance in terms of log-likelihood was lower than those of our methods. The sparse density trees and rule lists performed well in terms of sparsity and log-likelihood. Among them, the sparse density rule list method had a slightly better likelihood-sparsity tradeoff.

Figure 5.5 shows one of the density models generated by the leaf-sparse density tree method. The reason for the split is clear: there were fewer children than adults, the distributions of the males and females were different (mainly because the crew was mostly male), and the volume of crew members was very different than the volume of first, second, and third class passengers.

Figure 5.6 shows density rule lists for the Titanic dataset. By choosing the preferred list length λ from set $\{2, 4, 7\}$, we change the density rule lists as well as their length. Lower values of the parameter lead to shorter density rule lists, while larger preferred lengths correspond to longer density rule lists.

5.2.3.2 Crime Dataset

The data used in this experiment are from housebreaking incidents reported by the Cambridge Police Department in Massachusetts. The goal is to understand the common methods used in housebreaks, which is crucial for crime analysis. The dataset includes 3739 separate housebreaks in Cambridge from 1997 to 2012. The six categorical features in the Crime dataset include the location of entry (categorized as “window”, “door”, “wall”, and “basement”), means of entry (“forceful”, “open area”, “picked lock”, “unlocked”, and “other”), whether the resident is inside, whether the premises are judged to be ransacked by the reporting officer, whether the entry occurred on a “weekday” or “weekend”, and the type of premise, further categorized into “residence”, non-medical, non-religious “workplace”, “medical”, “parking”, “social” (social clubs), “storage”, “construction site”, “street” and “church”.

We provide a comparison of our methods and DET in Figure 5.7(b). The standard histogram’s results were not reported since they involve too many bins (1440) to fit on the figure, and are thus not competitive. The sparse density trees and rule lists dominate DET for the Crime dataset and are sparser than DET trees.

One of the trees obtained from the leaf-sparse density tree method is in Figure 5.8. It states that most burglaries happen at residences – the non-residential density has values less

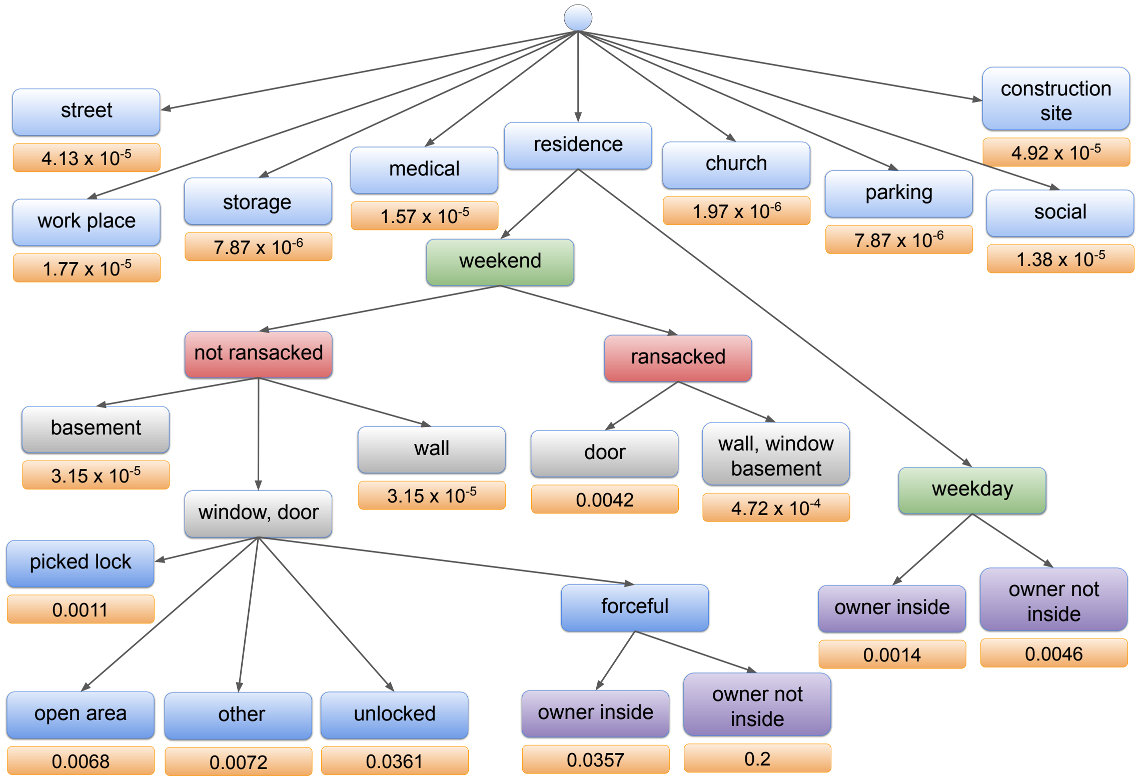


FIGURE 5.8: Leaf-sparse density tree representing the Crime dataset. Density is constant in each leaf. Different node colors represent different features in the dataset. This tree contains 20 leaves. It took around 1.4 seconds to create the tree and around 4 seconds to run the validation process.

than 2×10^{-4}. Given that a crime scene is a residence, most crimes happen on weekends. For residential crimes, burglary is more likely to happen when the owner is not inside (density 0.0046 if weekday and 0.2 if weekend, the premise is judged to be not ransacked and there is forceful entry through the door or window). When the premise is judged to be ransacked, the crime is more likely to happen with the door as the location of entry (density 0.0042) compared to wall, window, and basement (density 4.72×10^{-4}). On weekends, for residential and not-ransacked premises, doors and windows are more common locations of entry. If the entry is not forceful, unlocked windows and doors are the most common means of entry (density is 0.0361). If the means of entry is picked lock, the density is 0.0011 and if the area is open, the density is 0.0068.

Important aspects of the modus operandi are within the leaves of the tree, for instance,

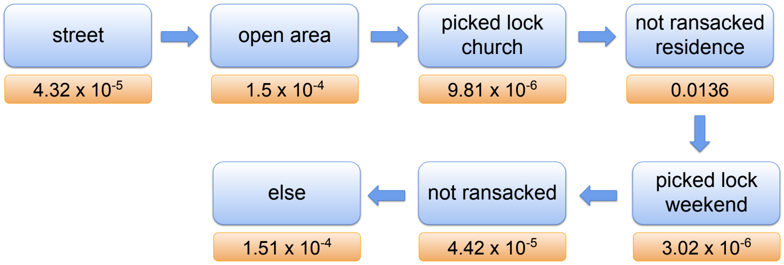


FIGURE 5.9: List representing the Crime dataset. Each arrow represents an “else if" statement.

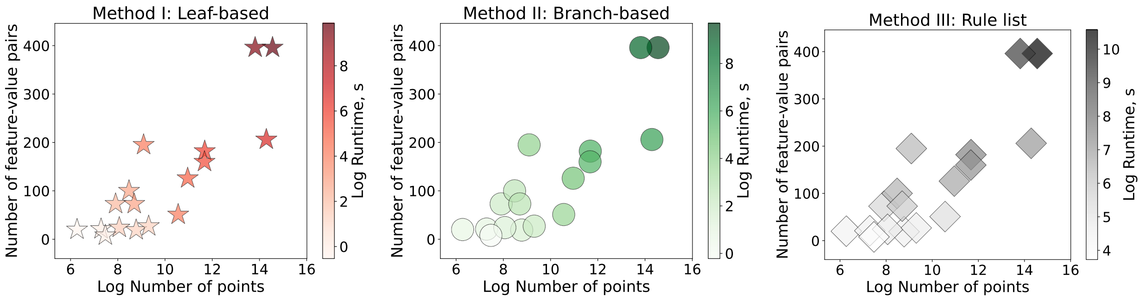
that the owner of a residence is not inside, and the house was not ransacked and the entry was forceful through the door or window. If this approach had been performed using a regular histogram, it would require 1440 different markers (discrete states), whereas the crime tree groups the crimes into just 20 bins.

These types of results can be useful for crime analysts to assess whether a particular modus operandi is unusual. For instance, according to the tree, it is clearly more unusual for the owner to be inside during the break-in, as shown by the smaller density values in the leaves when the owner is inside. Also, according to the densities in the leaves, it is more common for the means of entry to be forceful, and for the location of entry to be windows and doors. A density list for these data is presented in Figure 5.9. The preferred length of the list was chosen from the set $\{3, 5, 7\}$.

5.2.3.3 Run Time Analysis

The experiments below are designed to provide insight into how the methods operate.

We studied the performance of the density estimation methods on datasets of different complexity. We chose 17 datasets, including financial datasets (Bank-Full, Telco Customer Churn, HELOC), recidivism risk score data (COMPAS), UCI repository datasets (e.g., Car, Mushroom, US Census data), and detection labels of COCO stuff+thing data. The complexity of the dataset, in this case, is defined based on the number of samples and/or the number of feature-value pairs (the sum of categories for each feature). For the considered datasets, the number of samples ranged from 625 to around 2.5 million. The number of



(a)

(b)

(c)

FIGURE 5.10: Algorithm run time for all datasets as a function of dataset complexity (the log of the number of samples and the number of feature-value pairs) for (a) leaf-sparse density tree, (b) branch-sparse density tree, and (c) sparse density rule list estimation methods. Please refer to Table C.1 for more details.

feature-value pairs ranged from 9 to around 400 and there were from 3 to 68 features. Please see Table C.2 in Appendix C.1 for dataset statistics and pre-processing steps taken.

For datasets with less than 100k samples and less than 200 feature-value pairs, the tree-based methods (Method I and II) performed in under a minute and the rule list method (Method III) in under 7 minutes. The most complicated dataset, in terms of both the number of samples and feature-value pairs, is the US Census dataset (∼2.09M training samples, ∼400 feature-value pairs) which took around 1.75 hours for Method I, 2.5 hours for Method II, and $8\frac{2}{3}$ hours for Method III (note that majority of the time for Method III went into data loading and pre-processing). In our implementation, we represent data through bit vectors, thus an increase in the number of samples causes a relatively smaller increase in the run time compared to the run time increase caused by a larger number of feature categories. We also found that the leaf-sparse density tree method (Method I) performed fastest on average for all datasets considered. In Figure 5.10 we provide a visualization of the time taken to estimate the density of each dataset given its complexity for all three methods. Table C.1 in Appendix C.1 shows more details on the timing. All time measurements are averaged over 5 repeats.

5.2.4 Summary for Sparse Density Trees and Lists

In this section, we presented sparse density trees and rule lists – a Bayesian method for estimating density using sparse piece-wise constant estimators. Our methods are designed for categorical or binary data. The user-defined prior for each method encourages sparsity and thus enables interpretability. For tree-based methods, the prior is the user's desired number of leaves or branches in each node in the tree, while for the density rule list, the prior regularizes the length of the list. We designed a simulated annealing scheme, which alongside the inclusion-exclusion principle, and efficient data representation via bit vectors, allows us to find an optimal solution relatively fast.

Further, we illustrated the effectiveness of the methods for high-stakes decision domain datasets, such as Cambridge Police data. Our density functions can be easily visualized which helps in better understanding the data distribution.

6. Conclusions

In this dissertation, we have established a foundation for the existence of simpler-yet-accurate models. We have proposed Rashomon sets and ratios as another perspective on the relationship between hypothesis spaces and datasets, 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. More specifically, the main conclusions include: (1) Large Rashomon sets can embed models from simpler hypothesis spaces; (2) Similar performance across different machine learning algorithms may correlate with large Rashomon sets; (3) Large Rashomon sets correlate with existence of models that have good generalization performance; (4) The Rashomon ratio is a measure of a learning problem's complexity; (5) Noise is theoretical and practical motivator for the existence of simple, accurate models; (6) The Rashomon set characteristics tend to increase with noise; (7) Data that approximately arise from overlapping Gaussian distributions tend to have large Rashomon sets; (8) The interpretability of density trees and rule lists allows easier visualization of the estimated density values. These results validate our main thesis that for a lot of high-stakes decision domains, sparse models perform as well as black-box models. We illustrated this point for supervised learning and density estimation scenarios.

6.1 Practical Guidance for a Machine Learning Researcher

How can a machine learning practitioner benefit from these insights? Consider a researcher conducting a standard set of machine learning experiments in which the performance of several different algorithms is compared, and generalization is assessed. In the possible scenario where *all algorithms perform similarly*, and when their models tend to generalize well on validation data, the learning problem is likely to have a large Rashomon set. Based on the result in Chapter 3, simpler models are likely to exist in a large Rashomon set. If the researcher is interested in simpler models, they can search the simpler function class to locate simpler models within it. While optimizing for simplicity or interpretability constraints is usually much more computationally expensive than running standard machine learning algorithms, our thesis is that this search would be likely to succeed in the presence of a large Rashomon set. In the converse case, if the researcher's algorithms perform *differently from each other*, the researcher might then select a more complex model class that achieves better performance yet does not overfit. If the researcher suspects that the dataset comes from noisy generation processes, the results in Chapter 4 suggest that the Rashomon ratio might be large, therefore the researcher could assume that it is worthwhile to search for a simple model that performs well.

Recall from Chapter 4 that for the data that are likely to have arisen from overlapping Gaussian distributions, the Rashomon set tends to be large. There can be many datasets with such characteristics (that come from overlapping Gaussian distributions). For example, let us consider criminal recidivism data, whose Rashomon sets have been studied (Dong & Rudin, 2020; Fisher et al., 2019) and that admit simple-yet-accurate models (Rudin et al., 2020; Zeng et al., 2017). Each data point is generated based on a set of random events happening in the world; whether someone enters a job training program, whether someone associates with criminal associates after release, and whether someone commits a crime each day are all random variables whose random effects are cumulative over time and thus could be modeled by Gaussians by the central limit theorem. By this logic, we would expect many criminal recidivism prediction problems to admit large Rashomon sets. Other high-stakes predictions such as loan defaults may have similar characteristics.

In a sense, this full analysis paints a much clearer picture as to why such problems admit simple yet similarly accurate models: their distributions are approximately Gaussian with significant overlap, such overlap leads to large Rashomon sets, and large Rashomon sets lead to the existence of simple yet similarly accurate models.

Further, the machine learning practitioner can assess the data or its distribution using the sparse density trees and rule lists methods that we presented in Chapter 5. Models produced by these methods are easier to visualize, aiding in the understanding of the data distribution, detection of outliers and errors, and model selection, and could assist with decision-making.

6.2 Policy Implications

The results presented in this dissertation have profound policy implications, as they underscore the critical need to prioritize interpretable models in high-stakes decision-making. In a world where black box models are often used for high-stakes decisions, and yet the data generation processes are known to be noisy, our work sheds light on the false premise of this dangerous practice – that black box models are likely to be more accurate. Our findings have particular relevance for critical domains such as criminal justice, healthcare, and loan decisions, where individuals are subjected to the outputs of these models. The use of interpretable models in these areas can safeguard the rights and well-being of these individuals and ensure that decision-making processes are transparent, fair, and accountable.

6.3 Future Directions

In this dissertation, we study only one, but very important, side of the Rashomon Effect – its connection to the existence of accurate simpler models. However, when the Rashomon Effect is present, it changes *everything* we know about the machine learning problem as there is no single truth anymore. Instead, there is a diversity of models that can be contradictory but coexist and explain the same data from different perspectives. This acknowledgment prompts a reevaluation of traditional methodologies across various areas of machine learning, including but not limited to model selection, model evaluation, feature importance, data visualization, ethical considerations, bias mitigation, and the incorporation of human-in-the-loop feedback. The Rashomon Effect especially emerges in datasets from the high-stakes decision domains, requiring machine learning practitioners to identify its presence in the datasets they use and try to use the Effect to their advantage. Therefore, new methods and insights are needed to be developed for these datasets. Here we list a few possible directions for future research on the Rashomon set and Effect:

Causes for the Rashomon Effect. A better understanding of the data generation process and dataset properties is needed to figure out what causes the Rashomon Effect. One reason for larger Rashomon ratios is the label or attribute noise (as we discussed in

Chapter 4), while another contributing factor is a large geometric margin (as in Chapter 2). However, there has to be more to how data are collected and generated, especially for high-stakes decision domains, that cause larger Rashomon sets. The implications of these discoveries are vast: they will allow ML practitioners to have an understanding of whether they can expect the Rashomon Effect in the dataset without even attempting to measure it (which can be very computationally expensive). As a result, practitioners could make more informed decisions while working with a specific dataset. Further, such understanding of causes for the Rashomon Effect might help to detect when underspecification (D'Amour et al., 2022) and predictive multiplicity (when explanations are contradictory) (Marx et al., 2020) occur.

Navigating the Rashomon set. When the Rashomon set is large, many models can live in it (Qinyu Zhu et al., 2023; Xin et al., 2022). For example, there are 87500 trees in the Rashomon set for the recidivism prediction COMPAS dataset (Xin et al., 2022). While being able to obtain all the models in the Rashomon set provides incredible flexibility for ML practitioners and domain experts, a large number of available models may limit users' ability to interpret the data and overwhelm them during the model selection process. In this case, I will provide a summary of the models in the Rashomon set. This summary is essentially a cover over the Rashomon set, where the models are grouped based on their properties. The challenge is to define the distance between the models, taking into account loss, contradictory predictions, sparsity, feature importance, and other relevant properties for the user, so that the cover can be learned effectively. The representative models (centers) of the cover will highlight the different aspects of the well-performing models, and consequently help ML practitioners and domain experts to understand the data better.

Learning robust models in the Rashomon set. If the Rashomon set is large, it contains models with different properties, including those that are robust to various distributional changes or data permutations. Therefore, I would like to identify such robust models. One hypothesis is that a robust model might be an ensemble of various models from the Rashomon set, especially if these models exhibit different predictions based on the

Hamming distance. Alternatively, the ensemble can rely on a combination of the centers from the Rashomon set summary cover. A possible drawback is that an ensemble of models will likely reside in a more complex hypothesis space. Another hypothesis is to optimize for the model that is most resilient to leaving the Rashomon set. This approach is similar to distributionally robust optimization, but instead of considering the worst-case scenario, it optimizes over the Rashomon set. If the Rashomon set is large and we can find a robust model within it, then the cost of robustness is the existence of the Rashomon set, which happens often in practice.

A. Performance of Different Machine Learning Algorithms and Rashomon Ratio

A.1 Description of Datasets Used in Chapter 3

We provide a description of the datasets used in our experiments in Table A.1. All of them were downloaded from the UCI Machine Learning Repository (Dua & Graff, 2019). We show the number of features in each dataset, the sizes of the dataset, and any preprocessing steps that we used mainly to convert data to binary classification. For each dataset, we performed cross-validation over ten folds for datasets with more than 200 points and over five folds for datasets 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 datasets 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 no natural way to convert them to categorical features.

A.2 Rashomon Ratio Computation and Figures for Chapter 3

Figure A.1 and Figure A.2 show a performance comparison of different machine learning algorithms with regularization for the categorical and real-valued datasets. Datasets shown in Figures A.1 and A.2 are shown in decreasing order of the Rashomon ratio, from the highest in Figure A.1 to the Rashomon ratios that were so small that we were not able to measure them.

Influence of regularization on the Rashomon ratio. 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 size of the hypothesis space decreases, which increases 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 size of the Rashomon set would be larger, and, therefore, the Rashomon ratio would be larger too.

Figure A.3 and Figure A.4 show a comparison of the performance of different machine learning algorithms without regularization for the categorical and real-valued datasets.

Importance Sampling. As we mentioned before, we estimate the Rashomon ratio with importance sampling. As we discussed in Section 2.3.3, the probability of sampling a given tree from the target distribution is $p_t = p_f \times \prod_{i=1}^{2^D} \frac{1}{2}$. Thus, for one tree of depth seven, the 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. Note that if there are more than two classes in the dataset, the importance weight will be even smaller, as the probability of sampling a random tree from a target distribution decreases. Thus, trees built on a dataset with three classes will have a lower probability than trees built on a dataset with two classes. That is why we considered binary classification only and modified data as described in Table A.1, as it is essential for us to compare ratios over different datasets.

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 datasets and feature spaces, we use a consistent method of leaf-based importance sampling across all datasets and sample a manageable number of trees (250,000 in our case).

Large Rashomon ratios may appear artificially small. Even when the Rashomon ratio is a good driver of generalization performance, it may appear artificially small because of a poor representation of data or poor choice of hypothesis space. For instance, if the features are highly correlated, this artificially deflates the size of the Rashomon ratio as discussed in Appendix 3.3. Moreover, if 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 3.3(b), which includes some datasets with high-performing algorithms, yet (by the way we measured it) a small Rashomon ratio. In any case, small Rashomon ratios are not our main interest; here we are interested in what we would observe under large Rashomon ratios.

Table A.1: Description of the datasets used in Chapter 3 and processing notes.

Dataset 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

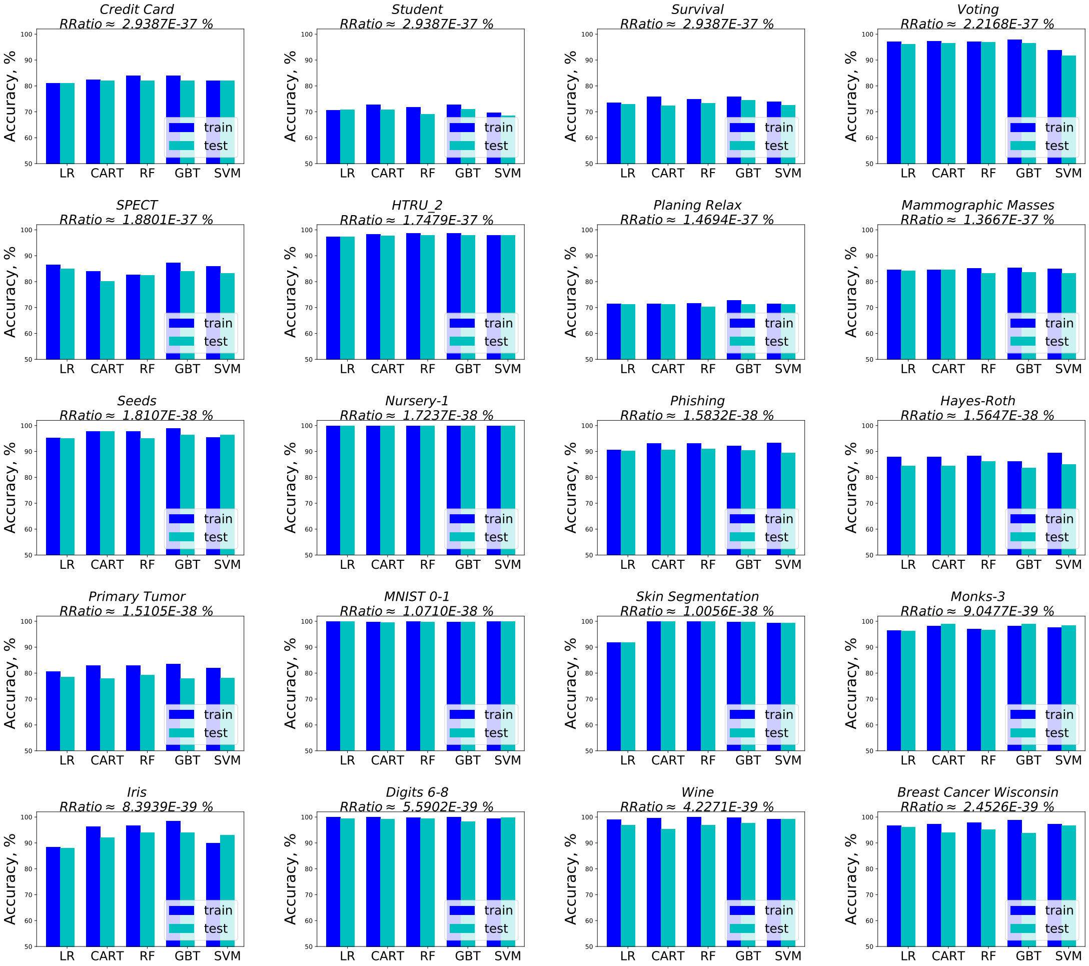


Figure A.1: Performance of five machine learning algorithms with regularization for the UCI classification datasets. Datasets are listed in decreasing order of Rashomon ratio. Rashomon ratios, train, and test accuracies are averaged over ten folds for datasets with more than 200 points and over five folds for datasets with less than 200 points. These plots continue in Figure A.2. 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 A.2.

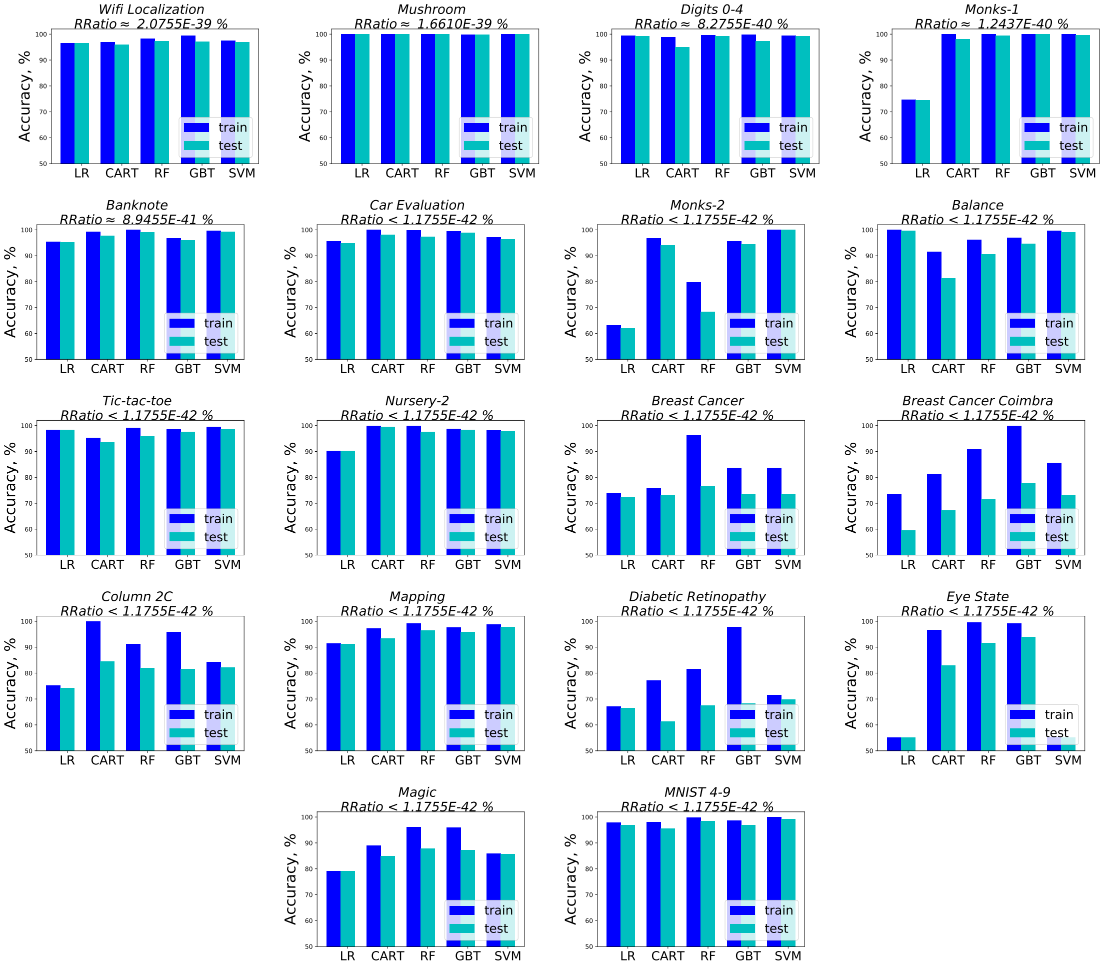


FIGURE A.2: Performance of five machine learning algorithms with regularization for the UCI classification datasets. Datasets are listed in decreasing order of the Rashomon ratio, continued from Figure A.1. Rashomon ratios, training accuracies, and test accuracies are averaged over ten folds for datasets with more than 200 points and over five folds for datasets with less than 200 points. Test performance sometimes varies across algorithms.

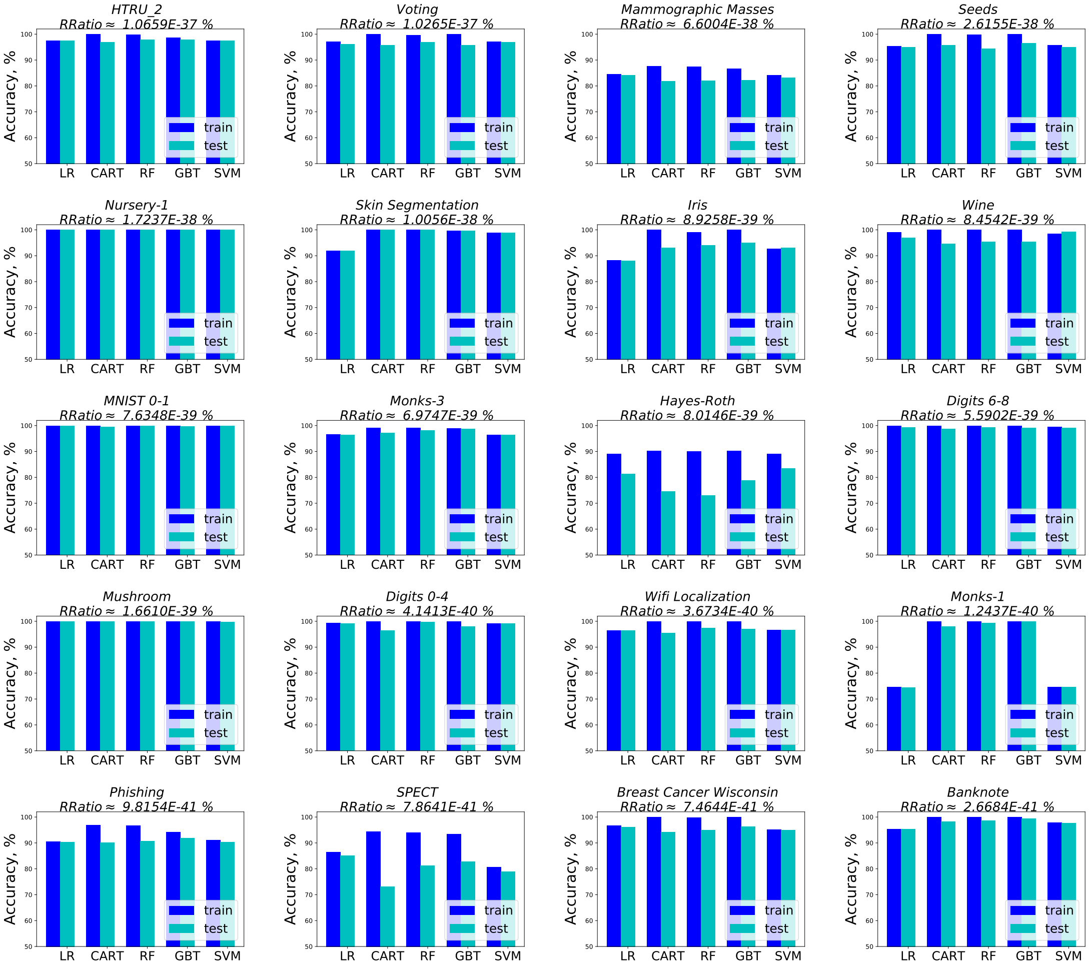


Figure A.3: Performance of five machine learning algorithms without regularization for the UCI classification datasets. Datasets are listed in decreasing order of Rashomon ratio. Rashomon ratios, train, and test accuracies are averaged over ten folds for datasets with more than 200 points and over five folds for datasets with less than 200 points. These plots continue in Figure A.4. The datasets with larger Rashomon ratios correlate with similar performance of machine learning algorithms and good generalization.

.

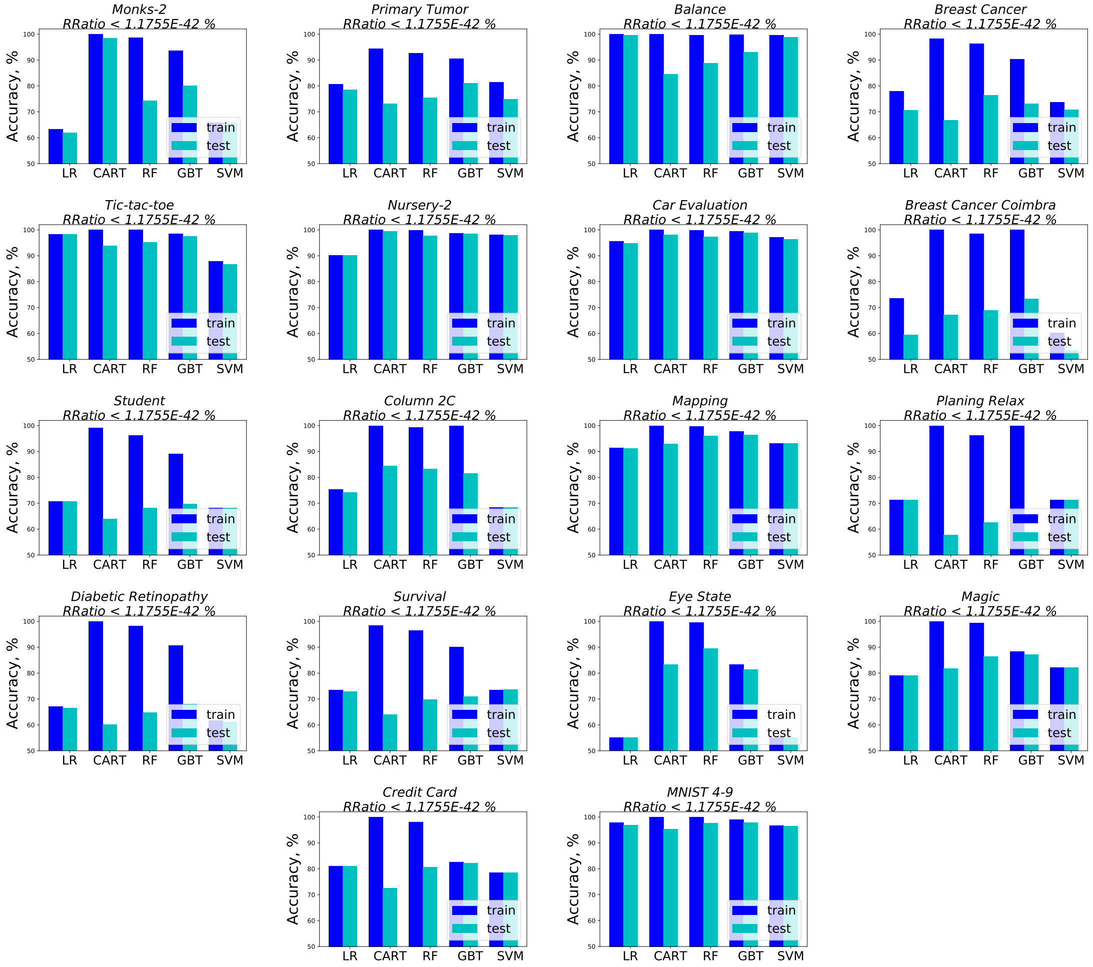


Figure A.4: Performance of five machine learning algorithms without regularization for the UCI classification datasets. Datasets are listed in decreasing order of the Rashomon ratio continuing from Figure A.3. Rashomon ratios, train, and test accuracies are averaged over ten folds for datasets with more than 200 points and over five folds for datasets with less than 200 points

.

B. Rashomon Sets in the Presence of Noise

B.1 Bernstein's and Hoeffding's Inequalities

We provide Bernstein's inequality in Lemma 37 and Hoeffding's inequality in Lemma 38.

Lemma 37 (Bernstein's inequality for loss class)**.** *Consider a hypothesis space $\mathcal{F}$. For a fixed $f \in \mathcal{F}$, let loss l be bounded by $C > 0$ such that $|l(f,z)| \leq C$ for every $z \in \mathcal{Z}$. For any $\varepsilon > 0$,*

$$P\left(L(f) - \hat{L}(f) > \varepsilon\right) \leq e^{\frac{-n\varepsilon^2}{2\sigma_f^2 + 2C\varepsilon/3}}, \tag{B.1}$$

where $\sigma_f^2 = \mathrm{Var}_{z\sim\mathcal{D}}\, l(f,z)$, and n is number of samples in $S = \{z_i\}_{i=1}^n \sim \mathcal{D}$.

Lemma 38 (Hoeffding's inequality for loss class)**.** *Consider a hypothesis space $\mathcal{F}$. For a fixed $f \in \mathcal{F}$, let loss l be bounded by $a, b \geq 0$ such that $a \leq l(f,z) \leq b$ for every $z \in \mathcal{Z}$. For any $\varepsilon > 0$,*

$$P\left(L(f) - \hat{L}(f) > \varepsilon\right) \leq e^{\frac{-2n\varepsilon^2}{(b-a)^2}}, \tag{B.2}$$

where n is the number of samples in $S = \{z_i\}_{i=1}^n \sim \mathcal{D}$.

Note that for 0-1 loss in the lemmas above, $a = 0$, $b = 1$, and $C = 1$. Now we show that Bernstein's inequality is stronger than Hoeffding's if the variance is lower than $\frac{(b-a)^2}{12}$.

Theorem 39 (Bernstein's inequality is stronger than Hoeffding's for lower variance)**.** *For a fixed $f \in \mathcal{F}$, let loss $l \in [a,b]$ so that $a \leq l(f,z) \leq b$ for every $z \in \mathcal{Z}$. Then, Bernstein's inequality is stronger than Hoeffding's inequality for all $\varepsilon \in (0, b-a)$ if $\sigma_f^2 \leq \frac{(b-a)^2}{12}$ or if $\left|L(f) - \frac{a+b}{2}\right| > \frac{b-a}{\sqrt{6}}$ where $\sigma_f^2 = \mathrm{Var}_{z\sim\mathcal{D}}\, l(f,z)$.*

Note that since the true risk and empirical risk can only differ by at most $b-a$, ϵ is not meaningful if $\epsilon \geq b-a$.

Proof. According to Hoeffding's inequality (B.2), we have that

$$P\left(\left|L(f) - \hat{L}(f)\right| > \varepsilon\right) \leq 2e^{\frac{-2n\varepsilon^2}{(b-a)^2}}.$$

Recall that Bernstein's inequality (B.1) states

$$P\left(\left|L(f)-\hat{L}(f)\right|>\varepsilon\right)\leq 2e^{\frac{-n\varepsilon^2}{2\sigma_f^2+2C\varepsilon/3}}$$

where $C=\frac{b-a}{2}$. Without loss of generality, let $l'(f,z)=l(f,z)-\frac{a+b}{2}$ so that $l'\in[-C,C]$. Then, we get that $L'(f)=L(f)-\frac{a+b}{2}$, $\mathrm{Var}_{z\sim\mathcal{D}}\, l'(f,z)=\mathrm{Var}_{z\sim\mathcal{D}}\, l(f,z)$, and $\hat{L}'(f)=\hat{L}(f)-\frac{a+b}{2}$. Therefore, we can rewrite Bernstein's inequality as

$$P\left(\left|L(f)-\hat{L}(f)\right|>\varepsilon\right)=P\left(\left|L'(f)-\hat{L}'(f)\right|>\varepsilon\right)\leq 2e^{\frac{-2n\varepsilon^2}{4\sigma_f^2+2(b-a)\varepsilon/3}}.$$

Consider $\sigma_f^2\leq\frac{(b-a)^2}{12}$. Then, we can upper-bound the right side of Bernstein's inequality by

$$2e^{-\frac{2n\varepsilon^2}{4\sigma_f^2+2(b-a)\varepsilon/3}}<2e^{-\frac{2n\varepsilon^2}{(b-a)^2/3+2(b-a)^2/3}}=2e^{\frac{-2n\varepsilon^2}{(b-a)^2}},$$

where $2e^{\frac{-2n\varepsilon^2}{(b-a)^2}}$ is the bound given by Hoeffding's inequality. Therefore, we showed that, if $\sigma_f^2\leq\frac{(b-a)^2}{12}$, then Bernstein's inequality is stronger than Hoeffding's inequality for all $\varepsilon\in(0,b-a)$.

We now consider $\left|L(f)-\frac{a+b}{2}\right|>\frac{b-a}{\sqrt{6}}$. Recall that $L'(f)=L(f)-\frac{a+b}{2}$, so we can rewrite this as $|L'(f)|>\frac{b-a}{\sqrt{6}}$. Since $-C\leq l'(f,z)\leq C$, we know that

$$\begin{aligned}\mathrm{Var}_{z\sim\mathcal{D}}(l'(f,z))&=E_{z\sim\mathcal{D}}((l'(f,z))^2)-(E_{z\sim\mathcal{D}}(l'(f,z)))^2\\&\leq C^2-(L'(f))^2\\&\leq\frac{(b-a)^2}{4}-\frac{(b-a)^2}{6}\\&=\frac{(b-a)^2}{12}.\end{aligned}$$

Then, we can follow the same argument as in the previous case to conclude that Bernstein's inequality is stronger than Hoeffding's inequality for all $\varepsilon\in(0,b-a)$. ■

B.2 Description of Datasets Used in Chapter 4

Please see Table B.2 for the description of the datasets used in the paper and all the processing steps. We normalize all real-valued features.

Table B.1: Description of the datasets used in Chapter 4 and processing notes

Dataset	Number of Samples	Number of Features	Notes
Car Evaluation	1728	16	We use one-hot encoding for features
Breast Cancer Wisconsin	699	11	We use one-hot encoding for features
Monks 1	124	12	We use one-hot encoding for features
Monks 2	169	12	We use one-hot encoding for features
Monks 3	122	12	We use one-hot encoding for features
SPECT	267	23	We use one-hot encoding for features
COMPAS	6907	13	Processed in (Xin et al., 2022)
FICO	10459	18	Processed in (Xin et al., 2022)
Bar 7 (Coupon)	1913	15	Processed in (Xin et al., 2022)
Expensive Restaurant	1417	16	Processed in (Xin et al., 2022)
Carryout Takeaway	2280	16	Processed in (Xin et al., 2022)
Cheap Restaurant	2653	16	Processed in (Xin et al., 2022)
Coffee House	3816	16	Processed in (Xin et al., 2022)
Bar	1913	16	Processed in (Xin et al., 2022)
Telco Bin	7043	6	We use only the binary features
Iris	100	4	We consider classes Versicolour and Setosa
Wine	130	13	
Wine 4	130	4	We use PCA to create 4 features
Seeds 4	140	4	We consider classes 1 and 2 and use PCA to create 4 features
Immunotherapy 4	90	4	(Khozeimeh, Alizadehsani, et al., 2017; Khozeimeh, Jabbari Azad, et al., 2017). We use one-hot encoding for feature "type". We use PCA to create 4 features
Penguin 4	265	4	We use one-hot encoding for feature "island." We consider classes "adelie" and "gentoo" only and use PCA to create 4 features
Digits 0-4 4	359	4	We consider digit 0 and digit 4. We use PCA to create 4 features

B.3 Description of Cross-Validation Process in Step 3 of the Path

We considered uniform label noise where each label is flipped independently with probability ρ. For each dataset, we performed five random splits into a train set and a validation set, where the validation set size is 20% of the number of samples. For the tree depth of CART, we considered the values $d \in \{1, \ldots, m\}$, where m is the number of features for a given dataset.

For Figure 4.3(a), we tuned the parameters and then added noise to see what happens, which is that performance degrades. For every train/validation split, we performed 5-fold cross-validation on the training set and computed the best depth. We fixed this depth (and thus hypothesis space). Then, we start adding noise to the dataset. We considered six different noise levels, $\rho \in \{0, 0.03, 0.05, 0.10, 0.15, 0.20.0.25\}$. For every level, we performed 25 draws of S_ρ. For every noise level, noise draw, and train/validation split, we evaluated train and validation performance and reported the average.

For Figure 4.3(b), we tuned the parameters for each noise level. We will see that noisier datasets lead us to use more regularization. We started adding noise to the dataset and then chose the best parameter based on cross-validation. More specifically, we considered six different noise levels, $\rho \in \{0, 0.03, 0.05, 0.10, 0.15, 0.20.0.25\}$. For every level, we performed 25 draws of S_ρ. Then we performed 5-fold cross-validation on the training data to choose the best depth for CART. For every noise level, noise draw, and train/validation split, we report mean depth based on cross-validation results.

For Figure 4.4, we varied the number of tree estimators. We used the same level of noise and cross-validation procedure as discussed above. For the number of estimators, we considered values $d \in \{5, 10, 20, \ldots, 150\}$.

B.4 Numerical Proof that Derivative is Negative for Conjuncture 35

See Figure B.1 for the numerical computation of the derivative for Conjuncture 35.

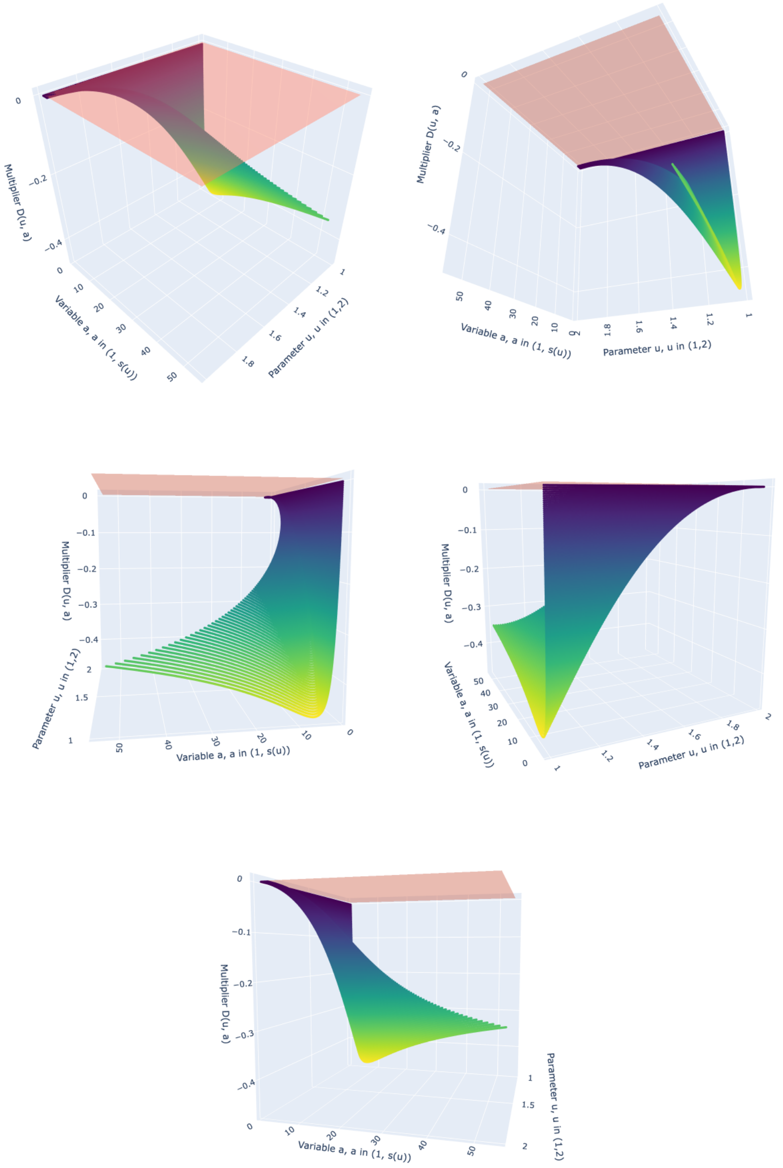


FIGURE B.1: Numerically showing that $D(u, a) < 0$ for any $u \in (1, 2)$, and $a \in (1, e^{\frac{4}{u^2}})$. The red plane corresponds to the value 0.

C. Additional Analysis of Sparse Density Trees and Rule Lists

C.1 Discussion on Run time Performance

Table C.1: Run time analysis of three methods for different complexity datasets. All time measurements are averaged over 5 repeats. "Leaf-sparse, best," "Branch-sparse, best," and "Rule list, best" show the average run time in seconds of the methods with best parameters λ and η. "Leaf-sparse, multiple," "Branch-sparse, multiple," and "Rule list, multiple" show the average run time in seconds of the methods, including the time needed to try multiple parameters.

Dataset	Number of train points	Number of feature-value pairs	Leaf-sparse, best, (sec)	Branch-sparse, best, (sec)	Rule list, best, (sec)	Leaf-sparse, multiple, (sec)	Branch-sparse, multiple, (sec)	Rule list, multiple, (sec)
Balance	531	20	**0.207**	1.036	8.226	0.587	2.166	82.402
Bank Full	38429	51	**19.652**	21.115	31.340	61.567	42.971	189.415
Car	1468	21	**0.276**	1.324	7.123	0.826	2.610	88.097
Chess (King-Rook vs. King-Pawn)	2716	73	**2.808**	3.631	58.269	8.683	7.147	246.194
COCO stuff+thing labels hierarchy	1607458	206	**315.418**	339.948	788.144	886.045	632.347	1175.841
COCO staff labels	118280	182	**119.142**	169.533	1056.583	382.756	316.800	1999.329
COCO things labels	117266	160	**104.491**	127.561	545.789	278.445	247.029	1476.896
COMPAS	6489	19	**1.225**	2.171	14.503	3.606	4.345	95.327
Connect 4	57423	126	**53.692**	58.286	404.347	143.833	111.995	882.699
Crime	3178	24	**1.393**	1.998	24.709	3.990	3.933	120.641
HELOC	8890	195	**19.402**	23.349	111.346	62.415	48.889	607.118
Mushroom	4797	100	**5.317**	6.362	31.890	16.360	12.685	699.592
Nursery	11016	27	**1.196**	4.418	10.667	3.587	8.571	106.414
Telco Customer Churn	5977	73	**6.405**	7.402	138.462	17.196	14.860	370.096
Titanic	1761	8	**0.324**	0.425	7.323	0.947	0.804	41.436
US Census 1990 (1m)	1000000	396	3580.249	**2813.563**	8382.199	10602.726	6299.193	10145.797
US Census 1990	2089542	396	**6216.533**	8965.318	31199.718	19469.654	16118.220	39081.668

We measured the performance of density estimation methods on 17 categorical datasets that are described in Table C.2. Continuous features in HELOC and Telco Customer Churn datasets were divided into 10 bins uniformly to create categorical features. We considered labels from the COCO stuff+thing dataset, and this resulted in three datasets: (1) COCO stuff that consists of binary features, where each indicates whether or not the stuff category is present in the image; (2) COCO thing that is built the same way except features are thing categories; (3) COCO stuff+thing, a four feature dataset that utilizes the hierarchical

Table C.2: Datasets statistics, including details on the dataset size, number of features and categories, and pre-processing notes if any.

Dataset	Total number of points	Number of train points	Number of validation points	Train validation split	Number of features	Number of feature-value pairs	Processing notes
Balance	625	531	94	15%	4	20	
Bank Full	45211	38429	6782	15%	12	51	
Car	1728	1468	260	15%	6	21	
Chess (King-Rook vs. King-Pawn)	3196	2716	480	15%	36	73	
COCO stuff+thing labels hierarchy	1677309	1607458	69582	4%	4	206	Features are formed from COCO detection labels hierarchy: stuff or thing, indoor or outdoor, super-category, and category
COCO stuff labels	123280	118280	5000	4%	91	182	Features are detection label categories
COCO thing labels	122218	117266	4952	4%	80	160	Features are detection label categories
COMPAS	7210	6489	721	10%	7	19	
Connect 4	67557	57423	10134	15%	42	126	
Crime	3739	3178	561	15%	6	24	
HELOC	10459	8890	1569	15%	23	195	Cut all features in 10 bins
Mushroom	5644	4797	847	15%	23	100	Dropped entries with missing values
Nursery	12960	11016	1944	15%	8	27	
Telco Customer Churn	7032	5977	1055	15%	19	73	Cut 3 continuous features in 10 bins; Dropped entries with missing values
Titanic	2201	1761	440	20%	3	8	
US Census 1990 (1m)	1150000	1000000	150000	15%	68	396	Considered 1 million samples
US Census 1990	2458285	2089542	368743	13%	68	396	

structure of COCO detection labels, where each feature is a hierarchy level (such as “animal,” “dog,” “things,” or “outdoor” where a “dog” is an “animal” and is “outdoor” and is in the “thing” dataset). We removed data samples with missing values. The train and validation data split for the run time experiments is fixed and shown in Table C.2.

For each setting of the parameters for each algorithm, we ran the algorithm five times to account for randomness in the optimization. We chose the best parameter values, and reported the average run time (over the 5 repeats) for these best parameters. We also reported the average (over the 5 repeats) total run time, including the time needed to choose parameter values. For the leaf-sparse density tree model, parameter λ (number of leaves) was chosen from the set $5, 8, 10$. For the branch-sparse density tree, λ (number of branches) was chosen from $2, 3$; for the sparse density rule list λ (length of the list), was chosen from the set $3, 5, 7$; and η (number of conjunctions in a rule) was chosen from $1, 2$. α was fixed to be 2 for the tree-based methods and 1 for the density rule list. Run time results for all 17 datasets are shown in Table C.1. For tree-based methods, the run time for evaluating multiple parameters is approximately three times (for leaf-sparse) and two times (for branch-sparse) larger than the run time for the methods when we knew the best parameters, simply because each run of the algorithm took approximately the same amount of time. For the density rule list, a significant portion of the run time is spent on data and volume pre-processing computations that are executed only once at the beginning. Thus, running the algorithm with multiple (i.e., 6) parameters is not 6 times the run time for running once when knowing the best parameters.

Run times in Table C.1 are computed by running our methods on Duke University's Computer Science Department cluster. On a single CPU machine (Intel(R) Core(TM) i7-6700 CPU @ 3.40GHz), for the Titanic dataset the average run time (over 5 iterations) for the leaf-based method is 0.276 sec, branch-based – 0.416 sec, density list – 6.928 sec; for Crime dataset: leaf-based – 0.1 sec, branch-based – 1.4 sec, density list – 12.28 sec; for Balance dataset: leaf-based – 0.202 sec, branch-based – 0.99 sec, density list – 7.544 sec; Bank Full dataset: leaf-based – 20.218 sec, branch-based – 23.077 sec, density list – 23.722 sec. The run times are reported for one average run of the algorithms.

C.2 Recommendations on the Choice of Algorithm and Parameters

From the experiments we conducted, we found that the leaf-based density trees were the most useful and intuitive. They also run the fastest (Table C.1) for the vast majority of the datasets we considered.

Leaf-based vs. branch-based. Leaf-based and branch-based methods are similar in structure but differ in how the shape of the model is controlled by the prior, specifically, the user controls either the number of leaves or the number of branches. This matters when there are many categories per feature: the leaf-based approach may try to put these

Table C.3: Description of the parameters for sparse density trees and lists. Additionally, initial guidance regarding the values of parameters is provided.

P	Meaning	Recommendations
		Leaf-Sparse Density tree
λ	Desired number of leaves in the tree.	8, 10. Experimentally we found that setting the prior for the number of leaves to 8 achieves good cross-validation log-likelihood.
α	Pseudocount, used to avoid zero values for the estimated densities.	1, 2 A small value.
		Branch-Sparse Density tree
λ	Desired number of branches at each internal node of the tree.	2, 3 Depends on the number of categories for each feature, and how much the user would like to aggregate them. If sparsity is preferred, then 2 or 3 is a good choice. Otherwise, may be set to 4 or 5.
α	Pseudocount, used to avoid zero values for the estimated densities.	1, 2 A small value.
		Sparse Density Rule List
λ	Desired length of the rule list.	7 or higher (for larger datasets).
η	Desired number of conjunctions in a rule.	1, 2, or 3 (for a smaller number of feature/value pairs). For larger datasets, mining rules and storing the data can be memory-expensive, so smaller values of this parameter may be preferred.
α	Pseudocount, used to avoid zero values for the estimated densities.	1, 2 A small value.

categories in one node, while the branch-based method might keep them in separate nodes. However, it takes longer to run the branch-based method (Table C.1), and its models are typically more complex than the leaf-based method's models on the same dataset (Table C.4).

Trees vs Lists. When comparing density trees and density rule lists, rule lists are one-sided trees, but they have multiple conditions defining each rule. Density rule lists can be more helpful if the user prefers a very sparse density model or has a smaller dataset. Concerning run time, the rule lists were the slowest method per run on average. However, one of the major bottlenecks for rule lists was memory space and time needed to process the data and mine the rules.

Parameters. Our sparse density lists and trees have priors on the model structure, such

Table C.4: Description of priors and the model complexities for models that maximized log-likelihood during the tuning procedure described in Appendix C.1.

Dataset			Leaf-based		Branch-based		Rule list		
	Number of train points	Number of feature-value pairs	Prior, λ	Number of leaves in the optimal model	Prior, λ	Number of leaves in the optimal model	Prior, λ	Prior, η	Length of the optimal model
Balance	531	20	5	2	3	40	3	1	3
Bank Full	38429	51	8	100	2	52	5	2	8
Car	1468	21	8	2	3	31	3	1	4
Chess (King-Rook vs. King-Pawn)	2716	73	5	26	3	47	7	2	8
COCO stuff+thing labels hierarchy	1607459	206	5	202	2	208	3	2	6
COCO stuff labels	118280	182	8	32	2	41	7	2	9
COCO thing labels	117266	160	5	47	3	29	7	1	7
COMPAS	6489	19	8	22	2	32	3	1	5
Connect 4	57423	126	10	47	3	42	7	2	7
Crime	3178	24	5	20	2	40	7	1	7
HELOC	8890	195	8	84	3	93	7	1	8
Mushroom	4797	100	10	52	2	73	3	1	6
Nursery	11016	27	8	2	3	48	3	2	2
Telco Customer Churn	5977	73	8	41	3	73	7	1	10
Titanic	1761	8	5	11	2	13	5	1	7
US Census 1990 (1m)	1000000	396	5	78	3	82	7	1	7
US Census 1990	2089542	396	5	100	2	105	7	1	5

as the number of leaves (Method I), branches (Method II), or length of the list (Method III). In Table C.3, we summarize all parameters that one needs to define in order to run our methods. Pseudocounts are typically set to small values in order to avoid zero densities. For all methods, λ regularizes the complexity of the model and reflects the prior belief on how sparse the user expects/would like the model to be. However, the resulting model complexity also depends on the data distribution. To give specific examples of priors and optimal model complexities, we analyzed trees and lists that we computed while evaluating run time in Appendix C.1. For every dataset, we reported the prior value that led to the maximum log-likelihood model and the complexity of this model (see Table C.4). For example, for the

COMPAS dataset with ~ 6500 training samples and 19 feature-value pairs, a prior of 8 on the number of leaves led to a tree with 22 leaves; a prior of 2 on the number of branches led to a tree with 32 leaves; a prior of 3 on the length of the rule list led to a model of length 5. While Table C.4 is a posthoc analysis of experiments conducted in Appendix C.1, it can still serve as a reference point for the prior value and optimal model complexity for different datasets. We also encourage users to perform cross-validation to choose parameters similar to the experiments we conducted in Chapter 5 and Appendix C.1.

Bibliography

Ahanor, I., Medal, H., & Trapp, A. C. (2023). DiversiTree: A new method to efficiently compute diverse sets of near-optimal solutions to mixed-integer optimization problems. *INFORMS Journal on Computing*.

Aïvodji, U., Arai, H., Gambs, S., & Hara, S. (2021). Characterizing the risk of fairwashing. *Advances in Neural Information Processing Systems*, *34*, 14822–14834.

Akaike, H. (1954). An approximation to the density function. *Annals of the Institute of Statistical Mathematics*, *6*(2), 127–132.

Anderlini, L. (2015). Density estimation trees in high energy physics. *arXiv preprint arXiv:1502.00932*.

Anderlini, L. (2016). Density estimation trees as fast non-parametric modelling tools. *Journal of Physics: Conference Series*, *762*(1), 012042.

Angelino, E., Larus-Stone, N., Alabi, D., Seltzer, M., & Rudin, C. (2018). Certifiably optimal rule lists for categorical data. *Journal of Machine Learning Research*, *18*, 1–78.

Bartlett, P., Freund, Y., Lee, W. S., & Schapire, R. E. (1998). Boosting the margin: A new explanation for the effectiveness of voting methods. *The Annals of Statistics*, *26*(5), 1651–1686.

Bartlett, P. L., Bousquet, O., Mendelson, S., et al. (2005). Local Rademacher complexities. *The Annals of Statistics*, *33*(4), 1497–1537.

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

Belkin, M., Hsu, D., Ma, S., & Mandal, S. (2019). Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*, *116*(32), 15849–15854.

Bishop, C. M. (1995). Training with noise is equivalent to Tikhonov regularization. *Neural Computation*, *7*(1), 108–116.

Black, E., Raghavan, M., & Barocas, S. (2022). Model multiplicity: Opportunities, concerns, and solutions. *2022 ACM Conference on Fairness, Accountability, and Transparency*, 850–863.

Bousquet, O., & Elisseeff, A. (2002). Stability and generalization. *Journal of Machine Learning Research*, *2*(Mar), 499–526.

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

Brunet, M.-E., Anderson, A., & Zemel, R. (2022). Implications of model indeterminacy for explanations of automated decisions. *Advances in Neural Information Processing Systems*, *35*, 7810–7823.

Burges, C. J. (1998). A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*, *2*(2), 121–167.

Cacoullos, T. (1966). Estimation of a multivariate density. *Annals of the Institute of Statistical Mathematics*, *18*(1), 179–189.

Caesar, H., Uijlings, J., & Ferrari, V. (2018). COCO-Stuff: Thing and stuff classes in context. *Computer vision and pattern recognition (CVPR), 2018 IEEE conference on.*

Cao, F., Xie, T., & Xu, Z. (2008). The estimate for approximation error of neural networks: A constructive approach. *Neurocomputing*, *71*(4-6), 626–630.

Cattaneo, M. D., Jansson, M., & Ma, X. (2019). Simple local polynomial density estimators. *Journal of the American Statistical Association*, 1–7.

Chaudhari, P., Choromanska, A., Soatto, S., LeCun, Y., Baldassi, C., Borgs, C., Chayes, J., Sagun, L., & Zecchina, R. (2019). Entropy-SGD: Biasing gradient descent into wide valleys. *Journal of Statistical Mechanics: Theory and Experiment*, *2019*(12), 124018.

Chen, T., Morris, J., & Martin, E. (2006). Probability density estimation via an infinite gaussian mixture model: Application to statistical process monitoring. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *55*(5), 699–715.

Chipman, H. A., George, E. I., & McCulloch, R. E. (2010). Bart: Bayesian additive regression trees. *The Annals of Applied Statistics*, *4*(1), 266–298.

Chomont, N., El-Far, M., Ancuta, P., Trautmann, L., Procopio, F. A., Yassine-Diab, B., Boucher, G., Boulassel, M. R., Ghattas, G., Brenchley, J. M., Schacker, T. W., Hill, B. J., Douek, D. C., Routy, J. P., Haddad, E. K., & Sékaly, R. P. (2009). HIV reservoir size and persistence are driven by T cell survival and homeostatic proliferation. *Nat Med*, *15*(8), 893–900.

Chun, T. W., Engel, D., Berrey, M. M., Shea, T., Corey, L., & Fauci, A. S. (1998). Early establishment of a pool of latently infected, resting CD4(+) T cells during primary HIV-1 infection. *Proc Natl Acad Sci U S A*, *95*(15), 8869–8873.

Chun, T. W., Stuyver, L., Mizell, S. B., Ehler, L. A., Mican, J. A., Baseler, M., Lloyd, A. L., Nowak, M. A., & Fauci, A. S. (1997). Presence of an inducible HIV-1 latent reservoir during highly active antiretroviral therapy. *Proc Natl Acad Sci U S A*, *94*(24), 13193–13197.

Coker, B., Rudin, C., & King, G. (2021). A theory of statistical inference for ensuring the robustness of scientific results. *Management Science*, *67*(10), 6174–6197.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine learning*, *20*(3), 273–297.

Coston, A., Rambachan, A., & Chouldechova, A. (2021). Characterizing fairness over the set of good models under selective labels. *Proceedings of the 38th International Conference on Machine Learning.*

Cover, T. M. (1965). Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition. *IEEE Trans. Electron. Comput.*, *14*, 326–334.

Cucker, F., & Smale, S. (2002). On the mathematical foundations of learning. *Bulletin of the American mathematical society*, *39*(1), 1–49.

Damian, A., Ma, T., & Lee, J. D. (2021). Label noise sgd provably prefers flat global minimizers. *Advances in Neural Information Processing Systems*, *34*, 27449–27461.

D'Amour, A., Heller, K., Moldovan, D., Adlam, B., Alipanahi, B., Beutel, A., Chen, C., Deaton, J., Eisenstein, J., Hoffman, M. D., et al. (2022). Underspecification presents challenges for credibility in modern machine learning. *The Journal of Machine Learning Research*, *23*(1), 10237–10297.

Davydov, O. (2011). Algorithms and error bounds for multivariate piecewise constant approximation. In *Approximation algorithms for complex systems* (pp. 27–45, Vol. 3). Springer.

DeVore, R. A. (1998). Nonlinear approximation. *Acta Numerica*, *7*, 51–150.

Devroye, L. (1991). Exponential inequalities in nonparametric estimation. In *Nonparametric functional estimation and related topics* (pp. 31–44). Springer.

Dinh, L., Pascanu, R., Bengio, S., & Bengio, Y. (2017). Sharp minima can generalize for deep nets. *Proceedings of the 34th International Conference on Machine Learning*, *70*, 1019–1028.

Dong, J., & Rudin, C. (2020). Exploring the cloud of variable importance for the set of all good models. *Nature Machine Intelligence*, *2*(12), 810–824.

Dua, D., & Graff, C. (2019). UCI machine learning repository.

Falcinelli, S. D., Kilpatrick, K. W., Read, J., Murtagh, R., Allard, B., Ghofrani, S., Kirchherr, J., James, K. S., Stuelke, E., Baker, C., Kuruc, J. D., Eron, J. J., Hudgens, M. G., Gay, C. L., Margolis, D. M., & Archin, N. M. (2021). Longitudinal Dynamics of Intact HIV Proviral DNA and Outgrowth Virus Frequencies in a Cohort of Individuals Receiving Antiretroviral Therapy. *J Infect Dis*, *224*(1), 92–100.

Falcinelli, S. D., Cooper-Volkheimer, A. D., Semenova, L., Wu, E., Richardson, A., Ashokkumar, M., Margolis, D. M., Archin, N. M., Rudin, C. D., Murdoch, D., & Browne, E. P. (2023). Impact of Cannabis Use on Immune Cell Populations and the Viral Reservoir in People With HIV on Suppressive Antiretroviral Therapy. *The Journal of Infectious Diseases*.

Finzi, D., Hermankova, M., Pierson, T., Carruth, L. M., Buck, C., Chaisson, R. E., Quinn, T. C., Chadwick, K., Margolick, J., Brookmeyer, R., Gallant, J., Markowitz, M., Ho, D. D., Richman, D. D., & Siliciano, R. F. (1997). Identification of a reservoir for HIV-1 in patients on highly active antiretroviral therapy. *Science*, *278*(5341), 1295–1300.

Fisher, A., Rudin, C., & Dominici, F. (2019). 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.

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

Gandhi, R. T., Cyktor, J. C., Bosch, R. J., Mar, H., Laird, G. M., Martin, A., Collier, A. C., Riddler, S. A., Macatangay, B. J., Rinaldo, C. R., Eron, J. J., Siliciano, J. D., McMahon, D. K., Mellors, J. W., & AIDS Clinical Trials Group A5321 Team. (2021). Selective Decay of Intact HIV-1 Proviral DNA on Antiretroviral Therapy. *The Journal of Infectious Diseases*, *223*(2), 225–233.

Ghosh, A., Kumar, H., & Sastry, P. S. (2017). Robust loss functions under label noise for deep neural networks. *Proceedings of the AAAI Conference on Artificial Intelligence*, *31*(1).

Goh, S. T., Semenova, L., & Rudin, C. (2024). Sparse density trees and lists: An interpretable alternative to high-dimensional histograms. *INFORMS Journal on Data Science*.

Hochreiter, S., & Schmidhuber, J. (1997). Flat minima. *Neural Computation*, *9*(1), 1–42.

Holte, R. C. (1993). Very simple classification rules perform well on most commonly used datasets. *Machine Learning*, *11*(1), 63–91.

Hsu, H., & Calmon, F. (2022). Rashomon capacity: A metric for predictive multiplicity in classification. *Advances in Neural Information Processing Systems*, *35*, 28988–29000.

Hu, W., Li, Z., & Yu, D. (2020). Simple and effective regularization methods for training on noisily labeled data with generalization guarantee. *International Conference on Learning Representations.*

Kakade, S. M., Sridharan, K., & Tewari, A. (2008). On the complexity of linear prediction: Risk bounds, margin bounds, and regularization. *Proceedings of the 21st International Conference on Neural Information Processing Systems*, 793–800.

Kearns, M. (1995). A bound on the error of cross validation using the approximation and estimation rates, with consequences for the training-test split. *Advances in Neural Information Processing Systems*, *8*.

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

Keskar, N. S., Mudigere, D., Nocedal, J., Smelyanskiy, M., & Tang, P. T. P. (2016). On large-batch training for deep learning: Generalization gap and sharp minima. *arXiv preprint arXiv:1609.04836 (appeared at ICLR 2017).*

Khozeimeh, F., Alizadehsani, R., Roshanzamir, M., Khosravi, A., Layegh, P., & Nahavandi, S. (2017). An expert system for selecting wart treatment method. *Computers in biology and medicine*, *81*, 167–175.

Khozeimeh, F., Jabbari Azad, F., Mahboubi Oskouei, Y., Jafari, M., Tehranian, S., Alizadehsani, R., & Layegh, P. (2017). Intralesional immunotherapy compared to cryotherapy in the treatment of warts. *International journal of dermatology*, *56*(4), 474–478.

Koltchinskii, V., & Panchenko, D. (2002). Empirical margin distributions and bounding the generalization error of combined classifiers. *The Annals of Statistics*, *30*(1), 1–50.

Kowal, D. R. (2022). Fast, optimal, and targeted predictions using parameterized decision analysis. *Journal of the American Statistical Association*, *117*(540), 1875–1886.

Langford, J., & Shawe-Taylor, J. (2002). PAC-Bayes & margins. *Proceedings of the 15th International Conference on Neural Information Processing Systems*, 439–446.

Lecué, G. (2011). *Interplay between concentration, complexity and geometry in learning theory with applications to high dimensional data analysis* [Doctoral dissertation, Université Paris-Est].

Lee, K.-H., He, X., Zhang, L., & Yang, L. (2018). Cleannet: Transfer learning for scalable image classifier training with label noise. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 5447–5456.

Lee, Y., Yao, H., & Finn, C. (2023). Diversify and disambiguate: Out-of-distribution robustness via disagreement. *The Eleventh International Conference on Learning Representations.*

Letham, B., Rudin, C., McCormick, T. H., & Madigan, D. (2015). Interpretable classifiers using rules and bayesian analysis: Building a better stroke prediction model. *Annals of Applied Statistics*, *9*(3), 1350–1371.

Li, D., Yang, K., & Wong, W. H. (2016). Density estimation via discrepancy based adaptive sequential partition. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, & R. Garnett (Eds.), *Advances in neural information processing systems 29* (pp. 1091–1099). Curran Associates, Inc.

Li, J. Q., & Barron, A. R. (2000). Mixture density estimation. In S. A. Solla, T. K. Leen, & K. Müller (Eds.), *Advances in neural information processing systems 12* (pp. 279–285). MIT Press.

Li, M., Soltanolkotabi, M., & Oymak, S. (2020). Gradient descent with early stopping is provably robust to label noise for overparameterized neural networks. *International conference on artificial intelligence and statistics*, 4313–4324.

Li, W., Wang, L., Li, W., Agustsson, E., & Van Gool, L. (2017). Webvision database: Visual learning and understanding from web data. *arXiv preprint arXiv:1708.02862.*

Lin, J., Zhong, C., Hu, D., Rudin, C., & Seltzer, M. (2020). Generalized and scalable optimal sparse decision trees. *International Conference on Machine Learning*, 6150–6160.

Liu, H., Lafferty, J., & Wasserman, L. (2007, 21–24 Mar). Sparse nonparametric density estimation in high dimensions using the rodeo. In M. Meila & X. Shen (Eds.), *Proceedings of the eleventh international conference on artificial intelligence and statistics* (pp. 283–290, Vol. 2). PMLR.

Liu, H., Xu, M., Gu, H., Gupta, A., Lafferty, J., & Wasserman, L. (2011). Forest density estimation. *Journal of Machine Learning Research*, *12*, 907–951.

Liu, Q., Xu, J., Jiang, R., & Wong, W. H. (2021). Density estimation using deep generative neural networks. *Proceedings of the National Academy of Sciences*, *118*(15).

Lu, L., Jiang, H., & Wong, W. H. (2013). Multivariate density estimation by Bayesian sequential partitioning. *Journal of the American Statistical Association*, *108*(504), 1402–1410.

Lugosi, G., & Wegkamp, M. (2004). Complexity regularization via localized random penalties. *The Annals of Statistics*, *32*(4), 1679–1697.

Luo, R., Liu, A., & Wang, Y. (2019). Combining smoothing spline with conditional gaussian graphical model for density and graph estimation. *arXiv preprint arXiv:1904.00204*.

MacWilliams, F. J., & Sloane, N. J. A. (1977). *The theory of error-correcting codes* (Vol. 16). Elsevier.

Mahapatruni, R. S. G., & Gray, A. (2011, November). Cake: Convex adaptive kernel density estimation. In G. Gordon, D. Dunson, & M. Dudík (Eds.), *Proceedings of the fourteenth international conference on artificial intelligence and statistics* (pp. 498–506, Vol. 15). PMLR.

Marx, C., Calmon, F., & Ustun, B. (2020). Predictive multiplicity in classification. *International Conference on Machine Learning*, 6765–6774.

Mason, B., Jain, L., Mukherjee, S., Camilleri, R., Jamieson, K., & Nowak, R. (2022). Nearly optimal algorithms for level set estimation. *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, *151*, 7625–7658.

Mata, K., Kanamori, K., & Arimura, H. (2022). Computing the collection of good models for rule lists. *arXiv preprint arXiv:2204.11285*.

Mendelson, S. (2003). A few notes on statistical learning theory. *Advanced Lectures on Machine Learning*.

Mukherjee, S., Niyogi, P., Poggio, T., & Rifkin, R. (2006). Learning theory: Stability is sufficient for generalization and necessary and sufficient for consistency of empirical risk minimization. *Advances in Computational Mathematics*, *25*, 161–193.

Müller, P., & Quintana, F. A. (2004). Nonparametric Bayesian data analysis. *Statistical Science*, *19*(1), 95–110.

Nadaraya, É. A. (1970). Remarks on non-parametric estimates for density functions and regression curves. *Theory of Probability & Its Applications*, *15*(1), 134–137.

Nakkiran, P., Kaplun, G., Bansal, Y., Yang, T., Barak, B., & Sutskever, I. (2021). Deep double descent: Where bigger models and more data hurt. *Journal of Statistical Mechanics: Theory and Experiment*, *2021*(12), 124003.

Natarajan, N., Dhillon, I. S., Ravikumar, P. K., & Tewari, A. (2013). Learning with noisy labels. *Advances in Neural Information Processing Systems*, *26*.

Newman, D., & Rivlin, T. (1976). Approximation of monomials by lower degree polynomials. *Aequationes Mathematicae*, *14*(3), 451–455.

Noh, H., You, T., Mun, J., & Han, B. (2017). Regularizing deep neural networks by noise: Its interpretation and optimization. *Advances in Neural Information Processing Systems*, *30*.

Ormoneit, D., & Tresp, V. (1996). Improved gaussian mixture density estimates using bayesian penalty terms and network averaging. In D. S. Touretzky, M. C. Mozer, & M. E. Hasselmo (Eds.), *Advances in neural information processing systems 8* (pp. 542–548). MIT Press.

Ormoneit, D., & Tresp, V. (1998). Averaging, maximum penalized likelihood and bayesian estimation for improving gaussian mixture probability density estimates. *IEEE Transactions on Neural Networks*, *9*(4), 639–650.

Parzen, E. (1962). On estimation of a probability density function and mode. *The Annals of Mathematical Statistics*, *33*(3), 1065–1076.

Patki, N., Wedge, R., & Veeramachaneni, K. (2016). The synthetic data vault. *2016 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, 399–410.

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

Pawelczyk, M., Broelemann, K., & Kasneci, G. (2020). On counterfactual explanations under predictive multiplicity. *Conference on Uncertainty in Artificial Intelligence*, 809–818.

Peluso, M. J., Bacchetti, P., Ritter, K. D., Beg, S., Lai, J., Martin, J. N., Hunt, P. W., Henrich, T. J., Siliciano, J. D., Siliciano, R. F., Laird, G. M., & Deeks, S. G. (2020). Differential decay of intact and defective proviral DNA in HIV-1-infected individuals on suppressive antiretroviral therapy. *JCI Insight*, *5*(4).

Qinyu Zhu, C., Tian, M., Semenova, L., Liu, J., Xu, J., Scarpa, J., & Rudin, C. (2023). Fast and interpretable mortality risk scores for critical care patients. *arXiv e-prints*, arXiv–2311.

Ram, P., & Gray, A. G. (2011). Density estimation trees. *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, 627–635.

Rehn, P., Ahmadi, Z., & Kramer, S. (2018). Forest of normalized trees: Fast and accurate density estimation of streaming data. *2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA)*, 199–208.

Rejtö, L., & Révész, P. (1973). Density estimation and pattern classification. *Problems of Control and Information Theory*, *2*(1), 67–80.

Rogers, W. H., & Wagner, T. J. (1978). A finite sample distribution-free performance bound for local discrimination rules. *The Annals of Statistics*, *6*(3), 506–514.

Rosenblatt, M., et al. (1956). Remarks on some nonparametric estimates of a density function. *The Annals of Mathematical Statistics*, *27*(3), 832–837.

Ross, A., Pan, W., Celi, L., & Doshi-Velez, F. (2020). Ensembles of locally independent prediction models. *Proceedings of the AAAI Conference on Artificial Intelligence*, *34*(04), 5527–5536.

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, *1*(5), 206–215.

Rudin, C., Chen, C., Chen, Z., Huang, H., Semenova, L., & Zhong, C. (2022). Interpretable machine learning: Fundamental principles and 10 grand challenges. *Statistics Surveys*, *16*, 1–85.

Rudin, C., Wang, C., & Coker, B. (2020). The age of secrecy and unfairness in recidivism prediction. *Harvard Data Science Review*, *2*(1).

Sasaki, H., & Hyvärinen, A. (2018). Neural-kernelized conditional density estimation. *arXiv preprint arXiv:1806.01754*.

Schapire, R. E., Freund, Y., Bartlett, P., Lee, W. S., et al. (1998). Boosting the margin: A new explanation for the effectiveness of voting methods. *The Annals of Statistics*, *26*(5), 1651–1686.

Scott, D. W. (1979). On optimal and data-based histograms. *Biometrika*, *66*(3), 605–610.

Seidl, T., Assent, I., Kranen, P., Krieger, R., & Herrmann, J. (2009). Indexing density models for incremental learning and anytime classification on data streams. *In 12th EDBT/ICDT*, 311–322.

Semenova, L., Chen, H., Parr, R., & Rudin, C. (2023). A path to simpler models starts with noise. *Thirty-seventh Conference on Neural Information Processing Systems.*

Semenova, L., Rudin, C., & Parr, R. (2022). On the existence of simpler machine learning models. *2022 ACM Conference on Fairness, Accountability, and Transparency*, 1827–1858.

Semenova, L., Wang, Y., Falcinelli, S. D., Archin, N., Cooper-Volkheimer, A. D., Goonetilleke, N., Murdoch, D. M., Rudin, C. D., Margolis, D. M., & Browne, E. P. (2023). Machine learning approaches identify immunologic signatures of total and intact HIV DNA during long-term antiretroviral therapy. *Submitted to eLIFE.*

Shahin Shamsabadi, A., Yaghini, M., Dullerud, N., Wyllie, S., Aïvodji, U., Alaagib, A., Gambs, S., & Papernot, N. (2022). Washing the unwashable : On the (im)possibility of fairwashing detection. *Advances in Neural Information Processing Systems*, *35*, 14170–14182.

Shallue, C. J., Lee, J., Antognini, J., Sohl-Dickstein, J., Frostig, R., & Dahl, G. E. (2018). Measuring the effects of data parallelism on neural network training. *arXiv preprint arXiv:1811.03600.*

Silverman, B. W. (1986). *Density estimation for statistics and data analysis* (Vol. 26). CRC press.

Smith, G., Mansilla, R., & Goulding, J. (2020). Model class reliance for random forests. *Advances in Neural Information Processing Systems*, *33*, 22305–22315.

Song, H., Kim, M., & Lee, J.-G. (2019). Selfie: Refurbishing unclean samples for robust deep learning. *International Conference on Machine Learning*, 5907–5915.

Song, H., Kim, M., Park, D., Shin, Y., & Lee, J.-G. (2022). Learning from noisy labels with deep neural networks: A survey. *IEEE Transactions on Neural Networks and Learning Systems.*

Srebro, N., Sridharan, K., & Tewari, A. (2010). Smoothness, low noise and fast rates. *Advances in Neural Information Processing Systems*, *23*, 2199–2207.

Teney, D., Peyrard, M., & Abbasnejad, E. (2022). Predicting is not understanding: Recognizing and addressing underspecification in machine learning. *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXIII*, 458–476.

Tulabandhula, T., & Rudin, C. (2014). Robust optimization using machine learning for uncertainty sets. *Proceedings of the International Symposium on Artificial Intelligence and Mathematics (ISAIM).*

Ustun, B., & Rudin, C. (2016). Supersparse linear integer models for optimized medical scoring systems. *Machine Learning*, *102*(3), 349–391.

Vapnik, V. (1999). *The nature of statistical learning theory.* Springer science & business media.

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

Varet, S., Lacour, C., Massart, P., & Rivoirard, V. (2023). Numerical performance of penalized comparison to overfitting for multivariate kernel density estimation. *ESAIM: Probability and Statistics*, *27*, 621–667.

Verwer, S., & Zhang, Y. (2017). Learning decision trees with flexible constraints and objectives using integer optimization. *Integration of AI and OR Techniques in Constraint Programming: 14th International Conference, CPAIOR 2017, Padua, Italy, June 5-8, 2017, Proceedings 14*, 94–103.

Wallace, C. S., & Boulton, D. M. (1968). An information measure for classification. *The Computer Journal*, *11*(2), 185–194.

Wand, M. (1997). Data-based choice of histogram bin width. *The American Statistician*, *51*(1), 59–64.

Wang, Z. J., Zhong, C., Xin, R., Takagi, T., Chen, Z., Chau, D. H., Rudin, C., & Seltzer, M. (2022). TimberTrek: Exploring and curating sparse decision trees with interactive visualization. *2022 IEEE Visualization and Visual Analytics (VIS)*, 60–64.

Wasserman, L. (2006). *All of nonparametric statistics*. Springer Science & Business Media.

Watson-Daniels, J., Parkes, D. C., & Ustun, B. (2023). Predictive multiplicity in probabilistic classification. *Proceedings of the AAAI Conference on Artificial Intelligence*, *37*(9), 10306–10314.

Wen, Y., Luk, K., Gazeau, M., Zhang, G., Chan, H., & Ba, J. (2019). An empirical study of large-batch stochastic gradient descent with structured covariance noise. *arXiv preprint arXiv:1902.08234*.

Wong, J. K., Hezareh, M., nthard, H. F., Havlir, D. V., Ignacio, C. C., Spina, C. A., & Richman, D. D. (1997). Recovery of replication-competent HIV despite prolonged suppression of plasma viremia. *Science*, *278*(5341), 1291–1295.

Wu, K., Hou, W., & Yang, H. (2018). Density estimation via the random forest method. *Communications in Statistics-Theory and Methods*, *47*(4), 877–889.

Wu, Y., Tjelmeland, H., & West, M. (2007). Bayesian CART: Prior specification and posterior simulation. *Journal of Computational and Graphical Statistics*, *16*(1), 44–66.

Xiao, T., Xia, T., Yang, Y., Huang, C., & Wang, X. (2015). Learning from massive noisy labeled data for image classification. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2691–2699.

Xin, R., Zhong, C., Chen, Z., Takagi, T., Seltzer, M., & Rudin, C. (2022). Exploring the whole rashomon set of sparse decision trees. *Advances in Neural Information Processing Systems*, *35*, 14071–14084.

Yan, T., & Zhang, C. (2022). Margin-distancing for safe model explanation. *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, *151*, 5104–5134.

Yang, H., Rudin, C., & Seltzer, M. (2017). Scalable Bayesian rule lists. *Proceedings of the 34th International Conference on Machine Learning (ICML).*

Yang, K., & Wong, W. H. (2014a). Density estimation via adaptive partition and discrepancy control. *arXiv preprint arXiv:1404.1425.*

Yang, K., & Wong, W. H. (2014b). Discovering and visualizing hierarchy in the data. *arXiv preprint arXiv:1403.4370.*

Zeng, J., Ustun, B., & Rudin, C. (2017). Interpretable classification models for recidivism prediction. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *180*(3), 689–722.

Zhong, C., Chen, Z., Liu, J., Seltzer, M., & Rudin, C. (2023). Exploring and interacting with the set of good sparse generalized additive models. *Advances in Neural Information Processing Systems.*

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

Zhuang, X., Huang, Y., Palaniappan, K., & Zhao, Y. (1996). Gaussian mixture density modeling, decomposition, and applications. *IEEE Transactions on Image Processing*, *5*(9), 1293–1302.

Biography

Lesia Semenova was born in Ukraine. She earned her Bachelor of Science and Master of Science degrees in Applied Mathematics from the Department of Computer Science and Cybernetics at the Taras Shevchenko National University of Kyiv in 2012 and 2014 respectively. She began her Ph.D. in Computer Science at Duke University in 2016 under the supervision of Professor Cynthia Rudin and Professor Ronald Parr. Her research focused on reliable and trustworthy AI and its applications. Student teams that she coached have won the ASA Data Challenge Expo twice and placed third in a 3C Shared Task competition on scholarly document processing. Before joining Duke, she worked at the Samsung Research and Development Institute Ukraine in the Augmented Reality team. During her Ph.D. studies, she did two research internships at Pinterest Labs, where her work focused on user modeling and understanding. She was selected as one of the 2024 Rising Stars in Computational and Data Sciences by the Oden Institute for Computational Engineering and Sciences at The University of Texas at Austin.