1. Introduce Definitions

Rashomon Effect	Many different explanations exist for the same phenomenon. In machine learning, the Rashomon Effect occurs when many accurate-but-different models exist to describe the same data.
Rashomon Set	It means the set of all accurate-but-different models.
Rashomon Volume	It means the quantitative form of the Rashomon Effect. In machine learning, it also means the size of the Rashomon Set.
Rashomon Radio	It means a ratio of Rashomon Volume to the volume of hypothesis space. Also, Rashomon Radio is a new measure related to the simplicity of model classes.

2. Prove the Rashomon Set is Realistic

We can prove the Rashomon Set is realistic by exploring the connections between the hypothesis space, the Rashomon Volume, the Rashomon Radio, and the Rashomon Parameter. An illustration of a possible Rashomon Set is shown in Figure 1.

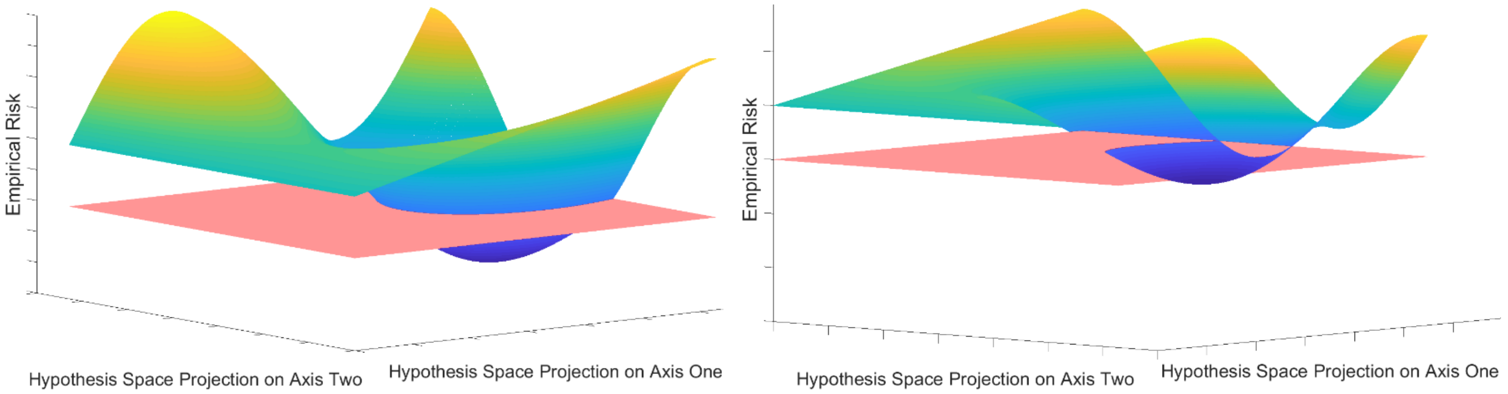

Figure 1: Red plane represents the Rashomon Parameter θ. Models below the plane belong to the Rashomon Set (in the case of two-dimensional hypothesis space).

3. More Specifically: The Definition of the Rashomon Set

Consider a training set of n data points $S=\{z_1,z_2,\cdots,z_n\}, z_i=(x_i,y_i)$ drawn i.i.d. from an unknown distribution $\mathcal{D}$ on a bounded set $\mathcal{Z}=\mathcal{X}\times\mathcal{Y}$, where $\mathcal{X}\subset R$ and $\mathcal{Y}\subset R$ are an input and an output space, respectively. For each point $z=(x,y)$, the loss function is $\phi(f(x),y)$. The *true risk* defines as $L(f)=E_{z\sim\mathcal{D}}[\phi(f(x),y)]$ and the *empirical risk* defines as $\hat{L}(f)=\frac{1}{n}\Sigma_{i=1}^{n}\phi(f(x_i),y_i)$. We define the *empirical Rashomon set* (or simply *Rashomon set*) as a subset of models of the hypothesis space $\mathcal{F}$. More precisely:

Definition 3.1. (*Rashomon Set*) Given $\theta\geq 0$, a dataset S, a hypothesis space $\mathcal{F}$, and a loss function ϕ, the *Rashomon Set* $\hat{R}_{set}(\mathcal{F},\theta)$ *is the subspace of the* hypothesis space defined as follows:

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

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

4. Prove the Rashomon Set Can Meaningfully Capture Explainable Models

While proving the existence of an *explainable model* (also called *interpretable model* or *simpler-but-accurate model*) in the Rashomon Set is a challenging practical problem because we never solve for such a model directly, which can also be practically hard. From my perspective, we can only prove it step by step. Firstly, we can find the property on the higher-complexity class, which can help us realize that the true anchored Rashomon Set is large (It is not a key question, so it will not elaborate. An illustration of hardly possible to capture an explainable model is shown as Figure 2, from my perspective). Secondly, after having done some analysis on the higher-complexity class, we have generalization guarantees that use only the lower-complexity class (from my perspective, it means extending the conclusion/properties on the higher-complexity class to the lower-complexity class).

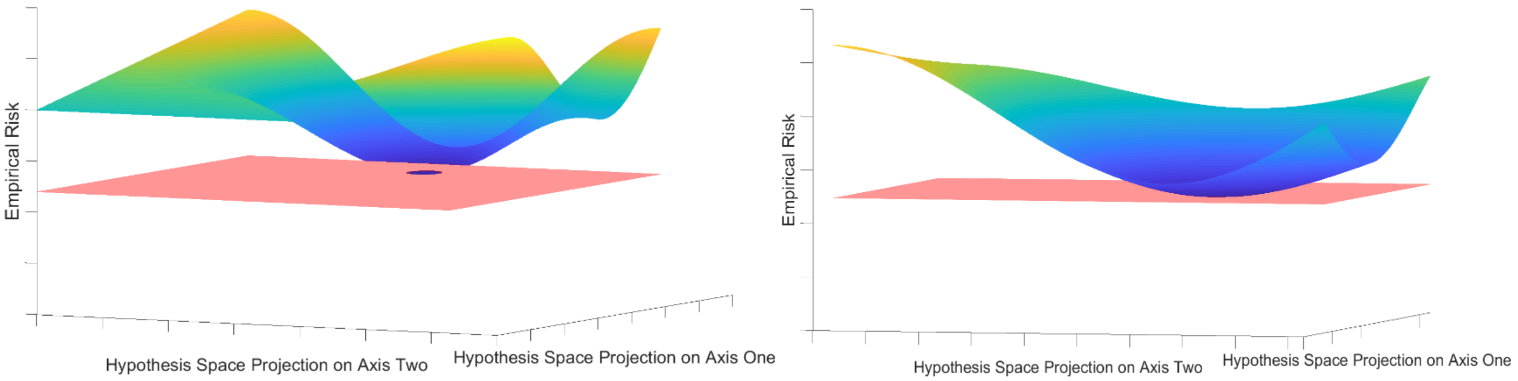

Figure 2: The true anchored Rashomon Set is not large enough.

5. More Specifically: The Existence of An Explainable Model

Consider the lower-complexity class $\mathcal{F}_1$ and the higher-complexity class $\mathcal{F}_2$. Firstly, we may want to optimize over $\mathcal{F}_2$ because it is less constrained. Secondly, we want to guarantee the existence of at least one function in $\mathcal{F}_1$ based on what we observe with $\mathcal{F}_2$. Thus, what we aim to prove in this section is the existence of functions in $\mathcal{F}_1$ that are in the Rashomon Set of $\mathcal{F}_2$.

The following theorem shows that, under certain conditions, if there is a function close to $\mathcal{F}_2$'s minimizer in function space that is also in $\mathcal{F}_1$, then it is a function we would be looking for, the explainable model. (As for proving the existence of multiple explainable models, it is not a key question, so it will not elaborate.)

Theorem 4.1. (Existence of **one** explainable model) For K-Lipschitz loss l bounded by b consider hypothesis spaces $\mathcal{F}_1$ and $\mathcal{F}_2$ such that $\mathcal{F}_1 \subset \mathcal{F}_2$. If there exists $\bar{f}_1 \in \mathcal{F}_1$ such that $\left\| \hat{f}_2 - \bar{f}_1 \right\|_p < \frac{\theta}{K}$, where $\hat{f}_2$ is the empirical risk minimizer within $\mathcal{F}_2$, then for a fixed parameter $\varepsilon \in (0,1)$:

1. $\bar{f}_1$ is in the Rashomon Set $\hat{R}_{set}(\mathcal{F}_2, \theta)$.

2. With probability greater than $1-\varepsilon$ with respect to the random draw of training data,

$$\left| L(\bar{f}_1) - \hat{L}(\bar{f}_1) \right| \le 2KR_n(\mathcal{F}_1) + b\sqrt{\frac{\log(2/\varepsilon)}{2n}}$$

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

Reference

[1] Lesia Semenova, Cynthia Rudin, Ronald Parr: A study in Rashomon curves and volumes: A new perspective on generalization and model simplicity in machine learning. arXiv:1908.01755. 2019;

[2] Laurent Dinh, Razvan Pascanu, Samy Bengio, Yoshua Bengio: Sharp Minima Can Generalize For Deep Nets. arXiv:1703.04933. 2017;

[3] 纪守领,李进锋,杜天宇: 机器学习模型可解释性方法、应用与安全研究综述. 计算机研究与发展. 2019;

[4] 吴飞,廖彬兵,韩亚洪: 深度学习的可解释性. 航空兵器. 2019;

[5] 成科扬,王宁,师文喜,詹永照: 深度学习可解释性研究进展.计算机研究与发展. 2020;