

The Double-Edged Nature of the Rashomon Set for Trustworthy Machine Learning

Ethan Hsu^{1*} Harry Chen^{2*} Chudi Zhong³ Lesia Semenova⁴
¹Duke University ²MIT ³UNC-Chapel Hill ⁴Rutgers University

Abstract

Real-world machine learning (ML) pipelines rarely produce a single model; instead, they produce a Rashomon set of many near-optimal ones. We show that this multiplicity reshapes key aspects of trustworthiness. At the individual-model level, sparse interpretable models tend to preserve privacy but are fragile to adversarial attacks. In contrast, the diversity within a large Rashomon set enables reactive robustness: even when an attack breaks one model, others often remain accurate. Rashomon sets are also stable under small distribution shifts. However, this same diversity increases information leakage, as disclosing more near-optimal models provides an attacker with progressively richer views of the training data. Through theoretical analysis and empirical studies of sparse decision trees and linear models, we characterize this robustness–privacy trade-off and highlight the dual role of Rashomon sets as both a resource and a risk for trustworthy ML.

1 Introduction

In high-stakes domains such as lending, criminal justice, and healthcare, the standard goal of finding a single “best” predictive model is no longer enough. A long-standing observation in statistical modeling challenges the idea that such a best model exists. Breiman’s Rashomon Effect [1] and follow-up work [2, 3, 4, 5, 6] show that many distinct models can achieve nearly indistinguishable predictive performance while relying on different features, logic, or decision boundaries. This multiplicity matters in modern high-stakes settings, where institutions require models that are not only accurate but also interpretable, stable under distribution shifts, and privacy-preserving. Recent algorithms [7, 8, 9, 10, 11, 12] now make it possible to construct or approximate these sets of near-optimal models, known as Rashomon sets, and to study and use them in practice.

Crucially, modern machine learning pipelines routinely produce such multiplicity even when practitioners do not think of themselves as computing a Rashomon set. Hyperparameter sweeps, fairness constraints, random seeds, feature restrictions, and automated model search all generate many near-optimal models. These models are often inspected, for example, in robustness audits or regulatory reporting. This perspective motivates a shift. Rather than viewing the Rashomon set as a theoretical construct, in realistic governance scenarios the *Rashomon set itself is the natural policy object*. It is the set of near-optimal models that institutions already generate during model development, that shape downstream decisions, and that regulators, auditors, or internal teams may query, compare, stress-test, or even partially disclose. Once the Rashomon set

*Equal contributions

is treated as a policy object, a natural question arises: *What are the positive and negative trustworthiness consequences of having a large, diverse Rashomon set?*

Although interpretability and fairness have been extensively studied within Rashomon sets [2, 13, 14, 15, 16, 17, 18], the relationships between model multiplicity and robustness, privacy, and stability remain far less understood. Recent work has begun to explore privacy in this context through differential privacy [19] and on robustness through active learning over Rashomon sets [20], but a systematic understanding of these properties in large, diverse Rashomon sets is still missing. In this work, we focus on these under-explored aspects of trustworthiness and examine how large, diverse Rashomon sets shape robustness, stability, and information leakage. Importantly, these sets introduce both opportunities and risks.

On the positive side, diversity within the Rashomon set can be a resource. When monitoring, audits, or red-teaming reveal a problematic region of the input space, institutions need not retrain from scratch. Instead, they can select a different near-optimal model in the Rashomon set that behaves more favorably on the flagged inputs. Furthermore, the Rashomon set can be leveraged for moving target techniques [21, 22], where models can be rotated out either on a regular basis or in response to adversarial attacks. A diverse Rashomon set can guarantee that attacks do not transfer between released models or other models in the wild. We refer to this ability to switch to a differently behaved, near-optimal model as *reactive robustness*. It arises because the Rashomon set contains many models that are equally accurate yet rely on different features or decision boundaries, increasing the chance that at least one model avoids the vulnerability or failure mode discovered in deployment.

On the negative side, the same diversity can be a liability. Releasing or internally exposing many near-optimal models, even individually sparse and interpretable ones, can accelerate information leakage about the training data. Each additional model offers a new “view” of the dataset, tightening an attacker’s ability to reconstruct features or approximate the underlying distribution.

An empirical illustration of this duality appears in Figure 1, which uses the COMPAS recidivism dataset. Each point represents a collection of sparse decision trees selected from the Rashomon set, varying both how many and which trees are included. As we select more diverse models, adversarial accuracy improves as we can find models that withstand attacks targeted at the original tree. Yet, the dataset reconstruction error monotonically decreases, indicating increased privacy risk. Together, these trends reveal a *robustness–privacy trade-off induced by model multiplicity*: diversity provides robustness but simultaneously increases information leakage. This trade-off is not specific to COMPAS and we observe similar behavior across multiple datasets (Section 6). The remainder of this paper explains *how this phenomenon arises* and discusses its implications.

To do so, we take a conceptual approach supported by theoretical analysis and empirical evidence, studying reactive robustness, stability, and information-theoretic privacy using sparse decision trees and linear models. Our main findings are: (1) We show that inherently interpretable models are fundamentally fragile to adversarial attacks, while sparsity naturally limits information leakage (Section 4). In other words, a single sparse model can be relatively private but not

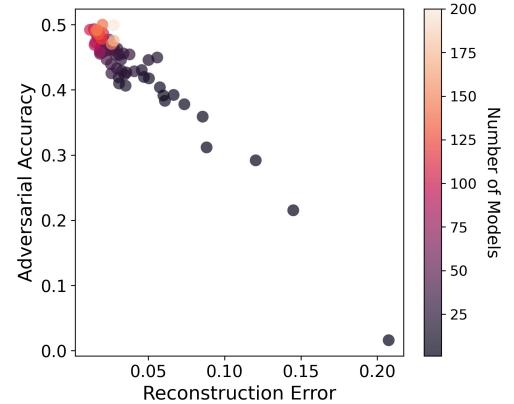


Figure 1: The robustness–privacy trade-off on the COMPAS dataset. As more diverse models from the Rashomon set are included, adversarial accuracy increases (greater reactive robustness), while reconstruction error decreases (greater information leakage).

robust, motivating a move beyond the single-model view. (2) When the Rashomon set is sufficiently diverse, it contains near-optimal models that remain accurate under adversarial attacks targeted at a specific model in the set (Section 5.1). (3) We prove that Rashomon sets are stable under small distribution and dataset shifts. Models that are near-optimal before the shift remain near-optimal within a slightly relaxed tolerance after the shift (Section 5.2). (4) We show that as more models from the Rashomon set are disclosed, information leakage grows, which means that the attacker can more accurately approximate the data distribution and reconstruct training examples (Section 5.3). (5) We support these findings with experiments on sparse decision trees across multiple datasets, illustrating the robustness–privacy trade-off due to model multiplicity.

Taken together, these results reveal a dual role of the Rashomon set: it is stable and provides a source for reactive robustness, yet it also presents a risk of increased privacy leakage when too many models are released. Our analysis offers a conceptual basis for understanding how diverse Rashomon sets shape robustness, stability, and privacy, and provides a foundation for future work on selecting, managing, or disclosing near-optimal models in practice, with the ultimate goal of supporting governance practices that operate at the Rashomon set level rather than the single-model level.

2 Related Work

Rashomon Effect. Rashomon Effect [23] occurs when there are multiple models that can explain data equally well. When these models produce different classifications, it is also known as model multiplicity [3, 4]. Recently, multiple methods have been proposed to measure the Rashomon Effect [2, 24, 3] or compute the Rashomon set (the set of near-optimal models) [7, 10, 12, 9]. In terms of trustworthy measures, the Rashomon Effect has been studied for sparsity [13, 14], fairness [25, 18, 26], explainability [27], variable importance [28, 29, 30], robustness [20] and differential privacy [19]. Among them, Nguyen et al. [20] introduces an active learning approach that selectively ensembles distinct, high-performing models from the Rashomon set, and Kulynych et al. [19] shows that the randomization inherent in differentially private training mechanisms leads to predictive multiplicity. Instead, we focus on how the existence of a diverse Rashomon set influences robust model selection and information leakage risks.

Adversarial Robustness. Many machine learning models are sensitive to perturbations in the inputs that lead to changes in the model outputs [31, 32]. Much work has been done to analyze and provide methods to reduce this vulnerability to adversarial perturbations. One technique is to introduce random noise either to input data [33, 34] or to the model weights [35, 36]. Other methods include explicit regularization to encourage robust behavior [37, 38], the removal of adversarial noise within the inputs [39], and the detection of adversarial inputs [40]. Theoretically, Zhang et al. [38] decomposes the adversarial error into a natural error as well as a boundary error controlled by the distance between data and the decision boundary, and Wu et al. [41] demonstrates that wider and more complex neural networks are more vulnerable to robust attacks. Bousquet and Elisseeff [42], Feldman and Vondrak [43] study stability of learning algorithms and its impact on generalization. Of particular interest to our work is the observation that adversarial perturbations transfer well between models [44, 45]. In Section 5.1, we demonstrate that having a Rashomon set with diverse models is necessary to block this transferability of adversarial attacks.

Privacy and information leakage. Privacy in ML refers to the requirement that models do not leak sensitive information about individual data points during training or inference [46]. Several mechanisms such as k-anonymity [47], L-diversity [48], and differential privacy [49] are used to address this issue. Empirically, privacy risks are often evaluated using membership inference attacks [50] and model inversion attacks [51, 52, 53]. From an information-theoretic perspective, privacy can be measured by the mutual information between the model and its training data [54, 55]. A research area explores the relationship between differential

privacy and information leakage [56, 57, 58]. In this paper, we study how the structure of interpretable models within the Rashomon set affects information leakage, and how sparsity constraints naturally reduce such leakage.

Robustness-Privacy trade-off. Studying the trade-offs among properties such as accuracy, robustness, and privacy is an active area of research in trustworthy ML [59], where the trade-offs between accuracy and privacy and between accuracy and robustness have especially been widely studied [60, 61, 62, 63, 38]. The trade-off between robustness and privacy remains an open problem. Some studies suggest that the two properties can benefit each other and that it is possible to achieve both simultaneously [64, 65, 66, 67]. However, other work has shown that training models with differential privacy can reduce adversarial robustness [68, 69], and adversarial training, which improves robustness, has been found to increase privacy risks [70, 71]. Most prior work focuses on neural networks, whereas in this paper, we explore the interplay between robustness and privacy for near-optimal inherently interpretable models.

The rest of the paper is organized as follows. Section 3 introduces necessary notations. Section 4 analyzes the robustness and privacy properties of individual models. Section 5 studies how the existence of a Rashomon set influences stability, robustness, and information leakage.

3 Notations

Consider n i.i.d. samples $S = \{(x_i, y_i)\}$ from an unknown distribution \mathcal{D} on $\mathcal{X} \times \mathcal{Y}$, where $\mathcal{X} \subset \mathbb{R}^p$ and $\mathcal{Y} \subset \mathbb{R}$ are an input and an output space respectively. Let \mathcal{F} be a hypothesis space. In this work we will mainly focus on the hypothesis space of interpretable models, including but not limited to linear models, sparse decision trees or rule lists, however, some of our results are hypothesis space agnostic and apply to any hypothesis space (such as theorems in Sections 5.2 and 5.3). Denote $\Omega(f)$ as a regularization term with a parameter $\lambda \in \mathbb{R}_{\geq 0}$. For example, $\Omega(\cdot)$ can be sparsity constraints that penalize the number of leaves in the decision tree or the length of the rule lists, or ℓ_0, ℓ_1, ℓ_2 norms. As a true risk, consider $L_{\mathcal{D}}(f) = \mathbb{E}_{\mathcal{D}}[\phi(f(x), y)]$, where $\phi : \mathbb{R}^2 \rightarrow \mathbb{R}_{\geq 0}$ is a loss function bounded on some fixed region (e.g. $[0, 1]$). We will mainly consider 0-1 loss ($\phi(f(x), y) = \mathbb{1}_{[f(x) \neq y]}$) for discrete hypothesis spaces and exponential loss ($\phi(f(x), y) = e^{-yf(x)}$) for the hypothesis space of linear models.

We aim to learn a model f^* from a hypothesis space \mathcal{F} that minimizes the objective $obj_{\mathcal{D}}(f) = L_{\mathcal{D}}(f) + \lambda\Omega(f)$. This is approximated by minimizing the empirical objective, $\hat{obj}_S(f) = \hat{L}_S(f) + \lambda\Omega(f)$, where $\hat{L}_S(f) = \frac{1}{n} \sum \phi(f(x_i), y_i)$ is the empirical risk. Correspondingly, $\hat{f} \in \arg \min_{f \in \mathcal{F}} \hat{L}_S(f)$ is an empirical risk minimizer (ERM). When the distribution or dataset is clear, we use the shorthand notation $L(f), \hat{L}(f), obj(f), \hat{obj}(f)$. When $\lambda = 0$, we will directly optimize the true or empirical risks without the regularization penalty.

Following [2, 13, 7, 28, 14], given $\epsilon > 0$, we define Rashomon set $\hat{\mathcal{R}}(\epsilon)$ as a set of near-optimal models, such that:

$$\hat{\mathcal{R}}(\epsilon) = \{f \in \mathcal{F} : \hat{obj}(f) \leq \hat{obj}(\hat{f}) + \epsilon\},$$

where ϵ is the Rashomon parameter that determines the tolerance for near-optimality. Generally, ϵ is a small value. It might correspond to a 1% relative increase over the optimal model's objective value (i.e., $\epsilon = 0.01 \cdot obj(\hat{f})$), or an absolute increase that corresponds to a small drop in accuracy (e.g., 1-3%). Note that when regularization parameter $\lambda = 0$, then $\hat{\mathcal{R}}(\epsilon) = \{f \in \mathcal{F} : \hat{L}(f) \leq \hat{L}(\hat{f}) + \epsilon\}$. Correspondingly, we also define the true Rashomon set, based on the true objective: $\mathcal{R}(\epsilon) = \{f \in \mathcal{F} : obj(f) \leq obj(f^*) + \epsilon\}$. Likewise, when $\lambda = 0$, we have $\mathcal{R}(\epsilon) = \{f \in \mathcal{F} : L(f) \leq L(f^*) + \epsilon\}$.

In this paper, we study how the existence of large Rashomon sets influences stability, privacy, and robustness. To motivate the set-level analysis, we first analyze how these criteria behave for a single near-

Criterion	Single Sparse Model	Rashomon Set
Robustness	✗ vulnerable (no alternatives)	✓ reactive robustness (has alternatives)
Stability	✓ algorithmic stability (via regularization)	✓ stable set (models remain near-optimal under shifts)
Privacy	✓ private (sparse model)	✗ leakier (more models reveal more information)

Table 1: Comparison of trustworthiness criteria at the single-model level versus the Rashomon-set level.

optimal model. Our results are illustrative rather than exhaustive. They identify the mechanisms behind robustness, stability, and privacy patterns using interpretable model classes, rather than delivering fully general characterizations. Table 1 summarizes the conceptual differences that our analysis will formalize.

4 A Sparse Model Can Be Private and Stable, but It May Not Be Robust

When relying on a single model, a data practitioner may hope to achieve both robustness and privacy. This section shows that achieving both simultaneously is difficult, even for interpretable or sparse models. While sparsity, which is often linked to interpretability on tabular data, can serve as a built-in privacy mechanism, we prove that these models are nonetheless inherently vulnerable to adversarial attacks.

4.1 Sparser Models are More Private and Algorithmically Stable

A model leaks information about its training data through the parameters or decision paths. We consider the information-theoretic perspective, where the leakage is quantified by mutual information between the learned model and the training data, denoted $I(f; S)$ [54]. Prior work shows that regularization can decrease $I(f; S)$ [72], motivating us to explore similar phenomena in sparse decision trees.

Theorem 1 (Sparsity controls mutual information in a single tree). *Let $S = \{(x_i, y_i)\}_{i=1}^n$ be a dataset of n i.i.d. samples from distribution \mathcal{D} over $\mathcal{X} \times \mathcal{Y}$, where $\mathcal{X} = \{0, 1\}^d$ and $\mathcal{Y} = \{0, 1\}$. Let \mathcal{F} be the class of binary classification decision trees with l_f leaves, and let $f \in \mathcal{F}$ be a tree fit on S through a possibly-random training algorithm. Then the mutual information between the learned tree f and the dataset S satisfies: $I(f; S) = O(l_f \log d)$.*

Theorem 1 shows that mutual information increases roughly linearly with the number of leaves, so sparser trees leak less information and are therefore more privacy-preserving. Intuitively, fewer splits mean fewer distinct models a deterministic learner can output, which limits the entropy of f . An analogous counting argument applies to linear models with an ℓ_0 penalty, where the number of nonzero coefficients plays the same role.

Another lens on privacy is membership inference attacks, where an adversary tries to decide whether a point was in the training set. Yeom et al. [73] show that the adversary’s advantage is bounded by the model’s generalization gap scaled by a constant. Because sparse models often generalize better than their non-regularized counterparts [74, 75, 76, 72], they naturally offer stronger protection.

Note that low mutual information (high privacy) and high algorithmic stability [42] are often observed together because they are a shared consequence of regularization. Indeed, Bousquet and Elisseeff [42] shows that, for reproducing kernel Hilbert spaces, regularization strength bounds stability, meaning that stronger regularization leads to better uniform stability (see Theorem 22 [42] for more details). Despite these privacy and stability benefits, a single sparse model remains inherently fragile to adversarial perturbations.

4.2 A Single Model is Inherently Vulnerable to the Adversarial Attack

Adversarial robustness revolves around an adversarial example, which, given a sample (x_i, y_i) , is defined as $x'_i = x_i + \delta$ such that $f(x'_i) \neq y_i$. To prevent trivial samples, perturbations are constrained within a bounded set defined by a L_p -norm. Formally, the set of permissible perturbation $\mathcal{S}_p = \{\delta \mid \|\delta\|_p \leq \eta\}$, where η specifies the maximum allowed perturbation magnitude. In this work, we will use \mathcal{S}_2 or \mathcal{S}_∞ when considering continuous samples and \mathcal{S}_0 for binary samples. The adversarial data \mathcal{D}' then can be constructed by taking $x_i \in \mathcal{X}$ and perturbed into adversarial sample x'_i .

In this setting, we analyze the vulnerability of a single rule list, a type of logical model composed of if-then-else statements. It's also viewed as a one-sided decision tree. Formally, a rule list with K rules is defined as a quadruple $d = (d_p, \delta_p, q_0, K)$,

where $d_p = (p_1, \dots, p_K)$ is the vector of antecedents (decision split nodes on the rule list path), $\delta_p = (q_1, \dots, q_K)$ is the vector of predictions corresponding to each decision split, and q_0 is the default prediction for samples that are captured by none of the antecedents (see Angelino et al. [77] for more details). The following theorem characterizes the adversarial risk of such models under simple binary perturbations.

Theorem 2 (Inherent vulnerability of single models). *Consider a binary dataset $S = \{(x_i, y_i)\}_{i=1}^n$ that has binary features and binary labels, where n_+ denotes the number of data points with positive labels in S . Let $d = (d_p, \delta_p, q_0, K)$ be a rule list such that $q_1 = 0$ and each rule predicts the majority label of the points captured by that rule. Further, let I be the smallest index i such that $q_i = 1$. Let $S' = \{x'_i, y_i\}_{i=1}^n$ be an adversarial dataset constructed by flipping up to one feature in each x_i (i.e., an L_0 -bounded perturbation with $\eta = 1$ restricted to binary features). Let \hat{L} be the 0-1 loss. If \bar{n}_+ is the number of positive data points captured by one of the first $I - 1$ leaves, then $\hat{L}_{S'}(d) - \hat{L}_S(d) \geq \frac{n_+ - \bar{n}_+}{n}$.*

From Theorem 2, we can see that the gap between the error and adversarial error increases with the number of positive points, meaning that a rule list with low robust error must have both low error and a high class imbalance. In other words, balanced datasets are unlikely to permit robust models. Note that this theorem provides a minimum guaranteed impact, irrespective of the specific attack strength η . We also note that a similar vulnerability exists for linear models under L_2 attack on datasets where much of the data is not well-separated, as the attack can simply push the data across the decision boundary. For both rule lists and linear models, a strong assumption on the data must be made in order for single models to be robust.

While a single model is vulnerable to adversarial perturbations, an attack designed for one model may not transfer to others in the Rashomon set. These limitations of single-model robustness naturally motivate analyzing the full set of near-optimal models next.

5 The Duality of the Rashomon Set

The existence of multiple, equally accurate models in the Rashomon set can present a fundamental trade-off. This section explores it, demonstrating how the existence of a large, diverse Rashomon set enables robustness and stability, but can also be exploited to create risks to data privacy. We provide formally stated results, including adversarial risk bounds, stability theorems, and a KL-based privacy bound, that together illustrate how the Rashomon set diversity shapes these trustworthy properties.

5.1 Diversity of the Rashomon Set Helps Robust Model Selection

In real deployments, vulnerabilities can be discovered reactively through performance changes on recent data, audits that show systematic inconsistencies, red-teaming exercises, or domain-expert complaints about

specific failure cases. These issues often affect a local region of the input space rather than the entire distribution. When such a problematic region is identified, retraining a model from scratch is the default response in many ML pipelines. However, retraining can be slow and computationally expensive. A more efficient alternative would be to keep a diverse approximation of the Rashomon set gathered either naturally as part of modern machine learning pipelines or through Rashomon set estimation techniques. Then, instead of retraining, an institution can switch to another near-optimal model in the set whose behavior differs on the problematic points while maintaining comparable predictive performance. Moreover, having a diverse Rashomon set might enable an option to apply a moving target defense, which defends against attacks by rotating through a set of diverse models or decision boundaries during deployment [21, 22].

In this section, we show that when the Rashomon set is sufficiently diverse, such reactive robustness becomes possible. Here, diversity refers to the presence of multiple near-optimal models that make meaningfully different decisions, while the precise definition of such difference is problem-dependent. For discrete hypothesis spaces such as decision trees or rule lists, we will capture diversity through prediction differences (e.g., Hamming distance between classifications). For continuous models such as linear classifiers, we will look at the diversity through geometric differences in the parameter space, such as the angle between weight vectors. Despite these specific formalizations, the takeaway on the reactive robustness is the same: if the Rashomon set contains models that disagree with the model on the attacked points, *the adversarial vulnerability of one model will not transfer to the others*.

Our theoretical results below formalize this intuition. For hypothesis spaces that optimize 0–1 loss, consider a dataset S with ERM \hat{f} . For theoretical clarity, we consider the worst case where every example in S is perturbed to form an adversarial dataset, which upper-bounds the impact of any localized attack. Let S' be an adversarial dataset constructed by modifying each sample (x_i, y_i) to attack \hat{f} . Intuitively, models that are similar to \hat{f} should be similarly vulnerable to this attack and thus should perform poorly on S' . To formalize this, we measure diversity between f and \hat{f} through their weighted prediction difference, $H(f, \hat{f}) = \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{f(x_i) \neq \hat{f}(x_i)}$. This Hamming distance can be considered as a diversity measure for discrete hypothesis spaces such as decision trees, rule lists, or scoring systems. Models with higher disagreement have more distinct decision boundaries and therefore may fail differently. Then, under 0–1 loss, the triangle inequality immediately gives a robustness bound:

$$L_{S'}(f) = \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{f(x'_i) \neq y'_i} \leq \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{f(x'_i) \neq \hat{f}(x'_i)} + \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{\hat{f}(x'_i) \neq y'_i} = H(f, \hat{f}) + L_{S'}(\hat{f}). \quad (1)$$

The inequality in (1) highlights the core mechanism behind reactive robustness: any model f can only outperform the attacked model \hat{f} on the adversarial dataset if it disagrees with \hat{f} sufficiently often. If $H(f, \hat{f})$ is small, then the bound forces $L_{S'}(f)$ to be close to $L_{S'}(\hat{f})$, meaning that f inherits the same vulnerability as \hat{f} . Conversely, a model that disagrees with \hat{f} on the attacked points can achieve better performance on S' . In other words, robustness against an attack targeted at \hat{f} requires diversity as models that mimic \hat{f} 's decisions will fail in the same way. This intuition applies broadly to any hypothesis space trained under 0–1 loss. Next, we provide more detailed theoretical evidence in the linear setting, where diversity, captured through geometry, similarly enables robustness through non-transferability of adversarial attacks.

Consider the hypothesis space of linear models $f(x) = w^T x$ where $w \in \mathbb{R}^p$ and let \hat{w}_S be ERM model for some dataset S . As before, let S' be an adversarial dataset generated from S using the L_2 norm and targeted to maximize the classification error of \hat{w}_S . That is, $x' = x + \delta$, where $\|\delta\|_2 \leq \eta$ and δ is chosen to make $y \cdot (\hat{w}_S)^T x'$ as small or negative as possible (e.g., $\delta \approx -\eta y \frac{\hat{w}_S}{\|\hat{w}_S\|_2}$). Here, we use the angle between weight vectors as our measure of diversity. We will show that models whose weight vectors form a larger angle have more distinct decision boundaries and are therefore less likely to share the same adversarial vulnerabilities. In

this sense, angular distance plays the same role for linear models that prediction disagreement played for 0-1 loss in the discussion above. For margin-based losses (i.e. losses that depend on functional margin $yf(x)$, such as exponential loss), we can compute the loss of an arbitrary linear model on the adversarial dataset as follows:

Theorem 3 (Risk on adversarial dataset). *Suppose that $\hat{L}_S(w) = \frac{1}{n} \sum_{i=1}^n \phi(y_i \cdot w^T x_i)$ where ϕ is a loss that is a function of the margin ($y_i f(x_i)$). For an L_2 attack on the optimal model \hat{w}_S with budget η , the loss of w on the adversarial dataset S' is $\hat{L}_{S'}(w) = \frac{1}{n} \sum_{i=1}^n \phi(y_i \cdot w^T x_i - \eta \|w\|_2 \cos(w, \hat{w}_S))$.*

For any reasonable objective, the loss ϕ should decrease with the margin. Thus, we intuitively have that, as the angle between a given model w and the optimal model increases (as measured in $\cos(w, \hat{w}_S)$), the margin should decrease faster, so the loss should increase faster. The next corollary formalizes this intuition for the exponential loss.

Corollary 1. *Suppose that we have the exponential loss $\phi(y \cdot w^T x) = e^{-y \cdot w^T x}$. Then, for any unit weights w_1 and w_2 satisfying $\cos(w_1, \hat{w}_S) > \cos(w_2, \hat{w}_S)$, we have that $\frac{\hat{L}_{S'}(w_1)}{\hat{L}_S(w_1)} > \frac{\hat{L}_{S'}(w_2)}{\hat{L}_S(w_2)}$. In other words, the adversarial attack is most effective on models more similar to \hat{w}_S .*

Further, as the strength of the adversarial attack grows, models that are of greater angle with the optimal model will eventually surpass the performance of lower angle models.

Corollary 2. *Suppose that $\phi(\cdot)$ is decreasing with respect to the margin, and let w_1 and w_2 be unit weights so that $\cos(w_1, \hat{w}_S) > \cos(w_2, \hat{w}_S)$. Then, for large enough η (e.g. when $\eta \rightarrow \infty$), $\hat{L}_{S'}(w_1) > \hat{L}_{S'}(w_2)$.*

From Corollaries 1 and 2, we see that models that make a greater angle with the optimal model are more robust to adversarial attacks. Therefore, if the Rashomon set is diverse enough to contain these models, they will perform well on both datasets S and S' . Our results also indicate that the type of adversarial attack and the definition of the Rashomon set diversity have to go hand in hand for a robust model to exist in the Rashomon set. For example, if the Rashomon set is diverse only through parallel shifts of the boundary, keeping w aligned with \hat{w} , then this diversity will likely not guarantee the existence of a robust model in the Rashomon set. Note that we observe similar results in the setting of least-square regression as we discuss in Appendix B.

5.2 Rashomon Sets are Stable to Small Changes in Distribution

After establishing reactive robustness benefits, we now will show that the Rashomon set as a whole remains stable under a small distribution shift. More specifically, we consider the scenario where the underlying data distribution changes slightly under covariate shift. If this shift is small, models that were good under the original distribution remain good under the new distribution, within a slightly relaxed performance threshold as we show in the next theorem.

Theorem 4 (Rashomon set is robust under small distribution shift). *Consider bounded loss function ϕ , $\phi \in [0, 1]$, and two data distributions \mathcal{D} and \mathcal{D}' , such that $\mathcal{D}(x) \neq \mathcal{D}'(x)$ and $\mathcal{D}(y|x) = \mathcal{D}'(y|x)$. If $KL(\mathcal{D} \parallel \mathcal{D}') \leq \frac{\epsilon^2}{8}$, then if a function f is in the true Rashomon set for the data distribution \mathcal{D} , $f \in \mathcal{R}_{z \sim \mathcal{D}}(\frac{\epsilon}{2})$, it is also in the true Rashomon set for the data distribution \mathcal{D}' , $f \in \mathcal{R}_{z \sim \mathcal{D}'}(\epsilon)$.*

Theorem 4 shows that the Rashomon set is stable: if a model performs well on one data distribution, it will still perform well, relative to the new best model, even after a small shift in the distribution. This doesn't

mean the model’s absolute accuracy is guaranteed as if the shift makes the task itself harder, performance may drop for all models. But the theorem ensures the drop is smooth because the set of good models changes gradually, so models that were strong before the shift remain among the best options after the shift. In other words, one doesn’t need to start the search for a good model all over again and can simply look for it within the existing Rashomon set.

Interestingly, the same stability occurs empirically, if the two datasets, S and S' , differ by only a small number of data points. In this case, the empirical Rashomon sets defined on these two datasets are similar. Specifically, any model belonging to the Rashomon set for one dataset is guaranteed to belong to the Rashomon set of the other dataset if we slightly increase the tolerance threshold by an amount proportional to the number of differing samples (K/n).

Theorem 5 (Two Rashomon sets constructed on neighboring datasets are indistinguishable). *For 0-1 loss, let S and S' be two datasets, each of size n . Let $\hat{\mathcal{R}}_S(\epsilon) := \{f \in \mathcal{F} | \hat{\text{obj}}(f, S) \leq \hat{\text{obj}}(\hat{f}, S) + \epsilon\}$ and $\hat{\mathcal{R}}_{S'}(\epsilon) := \{f \in \mathcal{F} | \hat{\text{obj}}(f, S') \leq \hat{\text{obj}}(\hat{f}', S') + \epsilon\}$. Suppose S and S' differ in at most K samples. Then $\hat{\mathcal{R}}_S(\epsilon) \subseteq \hat{\mathcal{R}}_{S'}(\epsilon + \frac{2K}{n})$ and $\hat{\mathcal{R}}_{S'}(\epsilon) \subseteq \hat{\mathcal{R}}_S(\epsilon + \frac{2K}{n})$.*

Theorem 5 establishes the stability of the Rashomon set when the dataset undergoes minor modifications, such as adding, removing, or changing a few examples. Note that while Theorem 2 suggests that such perturbations may decrease the absolute performance of the optimal model(s), Theorems 4 and 5 show that the Rashomon set as a whole can remain stable. We also verify this dataset-level stability empirically. Using four datasets and pre-computed Rashomon sets, we modify K samples in each dataset and recompute the Rashomon set on the modified data. When we increase the Rashomon tolerance from ϵ to $\epsilon + 2K/n$, the new Rashomon sets remain highly overlapping with the original. Even under a 6% dataset modification, more than 80% of the models remain, and for smaller perturbations (e.g., 2%), the sets are nearly identical (Figure 10). This confirms that small perturbations to the dataset do not radically change the pool of near-optimal models, providing practical evidence for Rashomon-set stability.

For an auditor, our observation means that if empirical Rashomon sets remain overlapping under small dataset perturbations, the institution’s choice among near-optimal interpretable models is reproducible as re-running the learning pipeline on new data will not radically change the pool of plausible models. The near-invariance in Theorem 5 also has implications for privacy as we discuss next.

5.3 Larger Rashomon Sets Increase Information Leakage

While the diversity of the Rashomon set can be a powerful defense, it also creates new risks. Imagine an organization that shares several of its top-performing interpretable models to promote transparency or comply with regulations. Release of each model on its own may seem safe, especially given privacy-preserving property of sparse models. But together, they can reveal more than intended. In this section, we explore how an adversary might combine together information from the released set of “safe” models to construct a privacy attack that is capable of reconstructing sensitive information from the original training data. We consider the Rashomon set for the arbitrary model class and focus on binary classification.

Theorem 6 (KL divergence bound for random ensembles from the Rashomon set). *Let $S = \{(x_i, y_i)\}_{i=1}^n$ be a dataset of n i.i.d. samples from distribution \mathcal{D} over $\mathcal{X} \times \mathcal{Y}$, where $\mathcal{Y} = \{0, 1\}$, and define $p(x) := P(y = 1|x)$ based on S . Let $\mathcal{R}(\epsilon)$ be the Rashomon set trained on S , containing $N = |\mathcal{R}(\epsilon)|$ models $\{f_1, \dots, f_N\}$. Assume each model outputs a probability $f_i(x) \in [\delta, 1 - \delta]$ for some constant $\delta \in (0, 1/2]$ for the given input x . Let $\mu(x) = \frac{1}{N} \sum_{i=1}^N f_i(x)$ be the mean prediction and $\sigma^2(x) = \frac{1}{N} \sum_{i=1}^N (f_i(x) - \mu(x))^2$ be the variance of predictions in the Rashomon set for input x . Suppose we sample m models without replacement from*

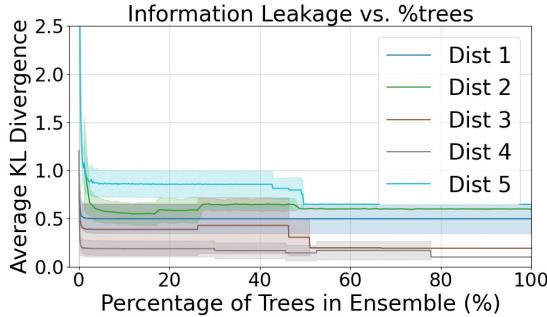


Figure 2: KL divergence between $p(x)$ and $q_{\Pi}(x)$ decreases as more trees are released into the ensemble.

$\mathcal{R}(\epsilon)$, where $m \leq N$, with sample indices $\Pi = (\pi_1, \pi_2, \dots, \pi_m)$ chosen uniformly without replacement from $\{1, 2, \dots, N\}$. Define the ensemble prediction as $q_{\Pi}(x) := \frac{1}{m} \sum_{i=1}^m f_{\pi_i}(x)$. Then, the expected KL divergence between $p(x)$ and the ensemble prediction $q_{\Pi}(x)$ is bounded by:

$$E_{\Pi}[KL(p(x)||q_{\Pi}(x))] \leq \frac{(p(x) - \mu(x))^2 + \frac{(N-m)\sigma^2(x)}{(N-1)m}}{\delta(1-\delta)}.$$

Theorem 6 reveals a privacy trade-off induced by Rashomon multiplicity. Even if each model in the Rashomon set is sparse and therefore individually privacy-preserving, releasing many such models increases aggregate leakage. The ensemble of models provides a more accurate approximation of the data distribution, as reflected by the decreasing KL divergence between $p(x)$ and $q_{\Pi}(x)$ (Figure 2). In this sense, larger Rashomon sets offer more “views” of the data, which collectively reveal more information about the underlying distribution.

It is important to distinguish this distributional leakage from *membership privacy*, which concerns whether an individual training example can be inferred from the released models. The distributional leakage does not automatically imply stronger membership-inference risk for individual data points. Membership privacy concerns whether releasing a model reveals whether a specific example was contained in the training set. Theorem 5 shows that Rashomon sets constructed on neighboring datasets (differing in only a few samples) are nearly identical as any model in one set is very likely to appear in the other. For large datasets, releasing the Rashomon set therefore reveals almost the same information for two neighboring datasets, limiting an adversary’s ability to infer the presence or absence of individual records. This aligns with the intuition behind differential privacy.

Finally, we note that diversity in the Rashomon set also has implications for ensemble-based robustness. Appendix B shows that forming an ensemble from sufficiently diverse models can improve robustness by reducing the chance that an adversarial perturbation affects all models simultaneously. This form of robustness is different from the reactive robustness studied in Section 5.1. Instead of switching to a new near-optimal model after an attack is detected, ensemble methods combine predictions from multiple diverse models to limit the transferability of the attack. This provides a mechanism for robustness that complements the reactive robustness view in Section 5.1. Importantly, the trade-off we discussed still holds because releasing many diverse models, whether individually or as part of an ensemble, increases distributional information leakage as characterized by Theorem 6.

Next, we focus on our empirical findings.

6 Experiments

We now present experimental results that support our conceptual framework and theoretical findings. We focus on sparse decision trees and use TreeFARMS [7] to construct the tree Rashomon set. Our evaluation aims to answer the following questions: (1) How does the diversity within the Rashomon set benefit adversarial robustness? (2) How does such diversity affect privacy? (3) What does the robustness–privacy trade-off look like?

6.1 Diversity in the Rashomon Set Benefits Adversarial Robustness

Given that sparse decision trees are inherently interpretable, we consider white-box attacks. We adopt the evasion algorithm proposed in [78] to generate adversarial examples by enumerating all possible perturbations that lead to incorrect predictions for the optimal tree in the Rashomon set, and study how other trees in the Rashomon set respond to these adversarial examples. We report the adversarial score defined as the accuracy on the adversarial dataset and the distance to the optimal tree calculated by the Hamming distance between the classification patterns. For binary classification, given a dataset S , a classification pattern $C_f(S)$ is a n -tuple of predicted labels $C_f(S) = (f(x_1), f(x_2), \dots, f(x_n))$. We use these classification patterns to measure diversity in (1) in Section 5.1 as well as in our experiments.

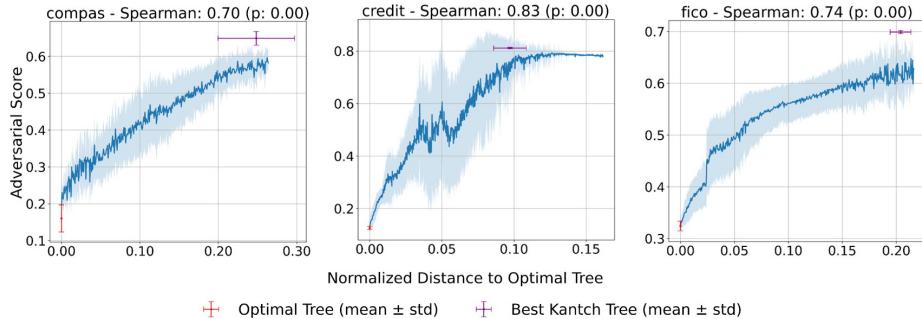


Figure 3: Adversarial score (accuracy) of trees in the Rashomon set vs. their distance to the optimal tree. Results are aggregated over five folds. The optimal trees (in red) are attacked. The most robust trees (in purple) are far from the optimal tree. Trees with the same distance to optimal trees are grouped, and mean and standard deviation of their adversarial score are shown as line plots with shaded uncertainty.

Figure 3 shows a strong positive trend between a tree’s distance from the optimal model and its adversarial score. Trees with classification patterns similar to the optimal tree (bottom left) are more vulnerable to adversarial examples, while those that differ more in their predictions (top right) are more robust. This suggests that diversity of the Rashomon set benefits adversarial robustness, as a subset of trees can maintain high adversarial accuracy even when others fail.

6.2 Diversity in the Rashomon Set Accelerates Information Leakage

While our theoretical analysis quantifies privacy leakage via mutual information, computing it exactly is often infeasible in practice. In our experiments, we therefore adopt a more tractable and widely used proxy: reconstruction error from a dataset reconstruction attack. This approach reflects an adversary’s ability to recover training data from released models and serves as an operational measure of privacy risk. DRAFT

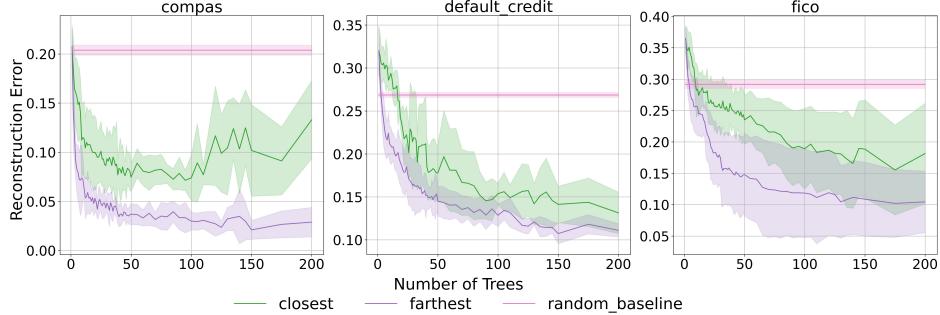


Figure 4: Comparison of reconstruction error between different selection strategies. The random baseline randomly guesses the feature values for each data point.

[79] can reconstruct the dataset used by a random forest by solving a constraint programming problem. We use the reconstruction error from this attack as a proxy for information leakage, where a lower reconstruction error indicates greater leakage. Note that the dataset reconstruction error differs from the membership privacy, which is protected by Rashomon-set stability (Theorem 5).

Following the setup in [79], we sample 100 data points to train a Rashomon set. Trees are then sequentially selected from the Rashomon set and passed to DRAFT. We run DRAFT multiple times as more trees are added. Specifically, DRAFT is trained after each additional tree from 1 to 50, every 5 trees from 50 to 150, and again at 175 and 200 trees. In total, we consider up to 200 trees from the Rashomon set. We run this process five times with different random seeds for sampling data points. We consider two strategies to select trees. By default, the first tree selected is the optimal model. The *closest* strategy then iteratively selects the tree whose classification pattern has the smallest Hamming distance to that of the optimal tree. The *farthest* strategy greedily selects the tree whose classification pattern has the largest Hamming distance from those of the previously selected trees.

Figure 4 shows that the “farthest” strategy (in purple) has lower reconstruction error, indicating greater information leakage. The “closest” strategy (in green) has better privacy, as it has a higher reconstruction error. These results suggest that more diversity in the Rashomon set and disclosing more diverse models lead to increased privacy risk.

6.3 Robustness-Privacy Tradeoffs under Rashomon Set

To study the relationship between robustness and privacy, we design a new strategy to sample trees from the Rashomon set. We first sort trees by Hamming distance of their classification patterns to the optimal tree, then select trees at evenly spaced intervals. For instance, given a Rashomon set of 1000 trees, an ensemble of 3 trees would include those ranked 1, 500, and 1000, while an ensemble of 100 trees would include every 10th tree in the sorted list. To evaluate robustness-privacy tradeoffs, we report both the reconstruction error and the best adversarial accuracy among the selected trees.

Figure 5 shows the reconstruction error versus adversarial accuracy for ensembles constructed by selected trees. Each point represents an ensemble of a specific size, with color indicating the number of trees included. When only a limited number of models are released, the privacy is preserved but the models are not diverse enough to avoid an adversarial attack targeted on the optimal trees. As more trees are selected, robustness improves but at the cost of greater information leakage. This result aligns with our previous findings and provides direct evidence of a robustness-privacy trade-off when releasing the Rashomon set in the wild. We

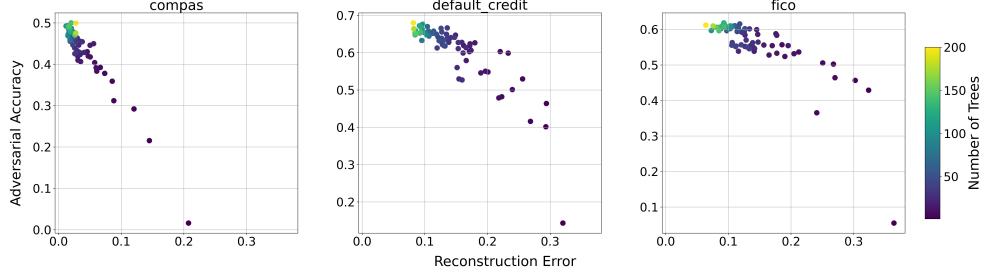


Figure 5: Reconstruction error vs. adversarial accuracy for ensembles constructed with different numbers of evenly sampled trees from the Rashomon set.

provide more experimental insights in Appendix C.

The reconstruction–robustness curves often display an intermediate regime where robustness increases substantially while reconstruction error has not yet sharply decreased. This pattern indicates a potential “sweet spot” in which releasing a modest number of diverse models provides robustness benefits with only limited additional privacy risk. The location and width of this intermediate region vary across datasets, reflecting differences in the diversity structure of each Rashomon set. Understanding what determines this region is a promising direction for future work.

7 Conclusions and Policy Implications

This work shows that diversity within the Rashomon set has a dual effect: it can provide robustness and stability, yet it also increases information leakage when many near-optimal models are exposed. Across theory and experiments, we demonstrate that these phenomena arise from the same mechanism. More specifically, models that differ in their predictions or parameters fail differently under adversarial perturbations but collectively reveal more about the underlying data distribution.

Our results highlight that trustworthiness is shaped not only by the properties of a single deployed model but by the structure of the Rashomon set as a whole, explored during model development. Even when only one model is ultimately selected, the existence of many near-optimal alternatives affects opportunities for reactive robustness, the reproducibility of model choice, and the aggregate privacy risk if these alternatives are released or inspected. Our analysis opens a direction for studying trustworthy ML at the level of sets of models rather than individual ones, particularly focusing on how different notions of diversity influence robustness, stability, and information leakage.

Our findings motivate a shift from single-model governance to set-level governance. Organizations may benefit from retaining a diverse Rashomon set internally to support robustness and reproducibility, while limiting external disclosure of near-optimal models to reduce privacy risk. Stability results such as Theorem 5 can serve as indicators of reproducibility across retraining or data updates. Rather than releasing entire Rashomon sets, institutions may disclose representative models or summary statistics, balancing transparency with privacy. These observations motivate future work on diversity-aware disclosure policies and practical tools for auditing Rashomon sets in real deployments.

References

- [1] Leo Breiman. Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical Science*, 16(3):199–231, 2001.
- [2] Lesia Semenova, Cynthia Rudin, and Ronald Parr. On the existence of simpler machine learning models. In *ACM Conference on Fairness, Accountability, and Transparency (ACM FAccT)*, 2022.
- [3] Charles Marx, Flavio Calmon, and Berk Ustun. Predictive multiplicity in classification. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 6765–6774, 2020.
- [4] Emily Black, Manish Raghavan, and Solon Barocas. Model multiplicity: Opportunities, concerns, and solutions. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 850–863, 2022.
- [5] Prakhar Ganesh, Afaf Taik, and Golnoosh Farnadi. The curious case of arbitrariness in machine learning. *arXiv preprint arXiv:2501.14959*, 2025.
- [6] Lucas Monteiro Paes, Rodrigo Cruz, Flavio P Calmon, and Mario Diaz. On the inevitability of the rashomon effect. In *2023 IEEE International Symposium on Information Theory (ISIT)*, pages 549–554. IEEE, 2023.
- [7] Rui Xin, Chudi Zhong, Zhi Chen, Takuya Takagi, Margo Seltzer, and Cynthia Rudin. Exploring the whole Rashomon set of sparse decision trees. In *Neural Information Processing Systems (NeurIPS)*, volume 35, pages 14071–14084, 2022.
- [8] Chudi Zhong, Zhi Chen, Jiachang Liu, Margo Seltzer, and Cynthia Rudin. Exploring and interacting with the set of good sparse generalized additive models. In *Neural Information Processing Systems (NeurIPS)*, 2023.
- [9] Jiachang Liu, Chudi Zhong, Boxuan Li, Margo Seltzer, and Cynthia Rudin. FasterRisk: Fast and accurate interpretable risk scores. In *Neural Information Processing Systems (NeurIPS)*, 2022.
- [10] Hsiang Hsu, Guihong Li, Shaohan Hu, et al. Dropout-based rashomon set exploration for efficient predictive multiplicity estimation. *arXiv preprint arXiv:2402.00728*, 2024.
- [11] Jon Donnelly, Zhicheng Guo, Alina Jade Barnett, Hayden McTavish, Chaofan Chen, and Cynthia Rudin. Rashomon sets for prototypical-part networks: Editing interpretable models in real-time. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 4528–4538, 2025.
- [12] Hsiang Hsu, Ivan Brugere, Shubham Sharma, Freddy Lecue, and Richard Chen. Rashomongb: Analyzing the rashomon effect and mitigating predictive multiplicity in gradient boosting. *Advances in Neural Information Processing Systems*, 37:121265–121303, 2024.
- [13] Lesia Semenova, Harry Chen, Ronald Parr, and Cynthia Rudin. A path to simpler models starts with noise. In *Neural Information Processing Systems (NeurIPS)*, 2023.
- [14] Zachery Boner, Harry Chen, Lesia Semenova, Ronald Parr, and Cynthia Rudin. Using noise to infer aspects of simplicity without learning. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL <https://openreview.net/forum?id=b172ac0R4L>.

- [15] Amanda Coston, Ashesh Rambachan, and Alexandra Chouldechova. Characterizing fairness over the set of good models under selective labels. In *Proceedings of the International Conference on Machine Learning (ICML)*, volume 139, pages 2144–2155, 18–24 Jul 2021.
- [16] Benjamin Laufer, Manish Raghavan, and Solon Barocas. What constitutes a less discriminatory algorithm? In *Proceedings of the 2025 Symposium on Computer Science and Law*, pages 136–151, 2025.
- [17] Emily Black, John Logan Koepke, Pauline T Kim, Solon Barocas, and Mingwei Hsu. Less discriminatory algorithms. *Geo. LJ*, 113:53, 2024.
- [18] Gordon Dai, Pavan Ravishankar, Rachel Yuan, Daniel B Neill, and Emily Black. Be intentional about fairness!: Fairness, size, and multiplicity in the rashomon set. *arXiv preprint arXiv:2501.15634*, 2025.
- [19] Bogdan Kulynych, Hsiang Hsu, Carmela Troncoso, and Flavio P Calmon. Arbitrary decisions are a hidden cost of differentially private training. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, pages 1609–1623, 2023.
- [20] Simon Nguyen, Kentaro Hoffman, and Tyler McCormick. Unique rashomon sets for robust active learning. *arXiv preprint arXiv:2503.06770*, 2025.
- [21] Abderrahmen Amich and Birhanu Eshete. Morphence: Moving target defense against adversarial examples. In *Proceedings of the 37th Annual Computer Security Applications Conference*, pages 61–75, 2021.
- [22] Ke He, Dan Dongseong Kim, and Muhammad Rizwan Asghar. Mtd-ad: Moving target defense as adversarial defense. *IEEE Transactions on Dependable and Secure Computing*, 22(5):5047–5059, 2025. doi: 10.1109/TDSC.2025.3560246.
- [23] L Breiman, JH Friedman, R Olshen, and CJ Stone. *Classification and Regression Trees*. Wadsworth, 1984.
- [24] Hsiang Hsu and Flavio Calmon. Rashomon capacity: A metric for predictive multiplicity in classification. In *Neural Information Processing Systems (NeurIPS)*, volume 35, pages 28988–29000, 2022.
- [25] Lucas Langlade, Julien Ferry, Gabriel Laberge, and Thibaut Vidal. Fairness and sparsity within rashomon sets: Enumeration-free exploration and characterization. *arXiv preprint arXiv:2502.05286*, 2025.
- [26] Anna P Meyer, Yea-Seul Kim, Aws Albarghouthi, and Loris D’Antoni. Perceptions of the fairness impacts of multiplicity in machine learning. *arXiv preprint arXiv:2409.12332*, 2024.
- [27] Sebastian Müller, Vanessa Toborek, Katharina Beckh, Matthias Jakobs, Christian Bauckhage, and Pascal Welke. An empirical evaluation of the rashomon effect in explainable machine learning. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 462–478. Springer, 2023.
- [28] Aaron Fisher, Cynthia Rudin, and Francesca Dominici. All models are wrong, but many are useful: Learning a variable’s importance by studying an entire class of prediction models simultaneously. *Journal of Machine Learning Research*, 20(177):1–81, 2019.

- [29] Jon Donnelly, Srikanth Katta, Cynthia Rudin, and Edward P Browne. The rashomon importance distribution: Getting RID of unstable, single model-based variable importance. In *Neural Information Processing Systems (NeurIPS)*, 2023.
- [30] Jiayun Dong and Cynthia Rudin. Exploring the cloud of variable importance for the set of all good models. *Nature Machine Intelligence*, 2(12):810–824, 2020.
- [31] Anh Nguyen, Jason Yosinski, and Jeff Clune. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 427–436, 2015.
- [32] Samuel G Finlayson, John D Bowers, Joichi Ito, Jonathan L Zittrain, Andrew L Beam, and Isaac S Kohane. Adversarial attacks on medical machine learning. *Science*, 363(6433):1287–1289, 2019.
- [33] Bai Li, Changyou Chen, Wenlin Wang, and Lawrence Carin. Certified adversarial robustness with additive noise, 2019. URL <https://arxiv.org/abs/1809.03113>.
- [34] Chuan Guo, Mayank Rana, Moustapha Cisse, and Laurens van der Maaten. Countering adversarial images using input transformations, 2018. URL <https://arxiv.org/abs/1711.00117>.
- [35] Rafael Pinot, Laurent Meunier, Alexandre Araujo, Hisashi Kashima, Florian Yger, Cédric Gouy-Pailler, and Jamal Atif. Theoretical evidence for adversarial robustness through randomization, 2019. URL <https://arxiv.org/abs/1902.01148>.
- [36] Zhezhi He, Adnan Siraj Rakin, and Deliang Fan. Parametric noise injection: Trainable randomness to improve deep neural network robustness against adversarial attack. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.
- [37] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples, 2015. URL <https://arxiv.org/abs/1412.6572>.
- [38] Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric Xing, Laurent El Ghaoui, and Michael Jordan. Theoretically principled trade-off between robustness and accuracy. In *International conference on machine learning*, pages 7472–7482. PMLR, 2019.
- [39] Fangzhou Liao, Ming Liang, Yinpeng Dong, Tianyu Pang, Xiaolin Hu, and Jun Zhu. Defense against adversarial attacks using high-level representation guided denoiser, 2018. URL <https://arxiv.org/abs/1712.02976>.
- [40] Jan Hendrik Metzen, Tim Genewein, Volker Fischer, and Bastian Bischoff. On detecting adversarial perturbations, 2017. URL <https://arxiv.org/abs/1702.04267>.
- [41] Boxi Wu, Jinghui Chen, Deng Cai, Xiaofei He, and Quanquan Gu. Do wider neural networks really help adversarial robustness?, 2021. URL <https://arxiv.org/abs/2010.01279>.
- [42] Olivier Bousquet and André Elisseeff. Stability and generalization. *Journal of machine learning research*, 2(Mar):499–526, 2002.
- [43] Vitaly Feldman and Jan Vondrák. Generalization bounds for uniformly stable algorithms. *Advances in Neural Information Processing Systems*, 31, 2018.

- [44] Hossein Hosseini, Yize Chen, Sreeram Kannan, Baosen Zhang, and Radha Poovendran. Blocking transferability of adversarial examples in black-box learning systems. *arXiv preprint arXiv:1703.04318*, 2017.
- [45] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.
- [46] Bo Liu, Ming Ding, Sina Shaham, Wenny Rahayu, Farhad Farokhi, and Zihuai Lin. When machine learning meets privacy: A survey and outlook. *ACM Computing Surveys (CSUR)*, 54(2):1–36, 2021.
- [47] Latanya Sweeney. k-anonymity: A model for protecting privacy. *International journal of uncertainty, fuzziness and knowledge-based systems*, 10(05):557–570, 2002.
- [48] Ashwin Machanavajjhala, Daniel Kifer, Johannes Gehrke, and Muthuramakrishnan Venkitasubramaniam. l-diversity: Privacy beyond k-anonymity. *Acm transactions on knowledge discovery from data (tkdd)*, 1(1):3–es, 2007.
- [49] Cynthia Dwork. Differential privacy. In *International colloquium on automata, languages, and programming*, pages 1–12. Springer, 2006.
- [50] Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership inference attacks against machine learning models. In *2017 IEEE symposium on security and privacy (SP)*, pages 3–18. IEEE, 2017.
- [51] Matt Fredrikson, Somesh Jha, and Thomas Ristenpart. Model inversion attacks that exploit confidence information and basic countermeasures. In *Proceedings of the 22nd ACM SIGSAC conference on computer and communications security*, pages 1322–1333, 2015.
- [52] Liwei Song and Prateek Mittal. Systematic evaluation of privacy risks of machine learning models. In *30th USENIX Security Symposium (USENIX Security 21)*, pages 2615–2632, 2021.
- [53] Sasi Kumar Murakonda and Reza Shokri. MI privacy meter: Aiding regulatory compliance by quantifying the privacy risks of machine learning. In *Workshop on Hot Topics in Privacy Enhancing Technologies (HotPETs)*, 2020.
- [54] Raef Bassily, Shay Moran, Ido Nachum, Jonathan Shafer, and Amir Yehudayoff. Learners that use little information. In *Algorithmic Learning Theory*, pages 25–55. PMLR, 2018.
- [55] Tianhao Wang, Yuheng Zhang, and Ruoxi Jia. Improving robustness to model inversion attacks via mutual information regularization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 11666–11673, 2021.
- [56] Mário S Alvim, Miguel E Andrés, Konstantinos Chatzikokolakis, and Catuscia Palamidessi. On the relation between differential privacy and quantitative information flow. In *International Colloquium on Automata, Languages, and Programming*, pages 60–76. Springer, 2011.
- [57] Mário S Alvim, Miguel E Andrés, Konstantinos Chatzikokolakis, Pierpaolo Degano, and Catuscia Palamidessi. Differential privacy: on the trade-off between utility and information leakage. In *International workshop on formal aspects in security and trust*, pages 39–54. Springer, 2011.

- [58] Borzoo Rassouli and Deniz Gündüz. Optimal utility-privacy trade-off with total variation distance as a privacy measure. *IEEE Transactions on Information Forensics and Security*, 15:594–603, 2019.
- [59] Alex Gittens, Bülent Yener, and Moti Yung. An adversarial perspective on accuracy, robustness, fairness, and privacy: Multilateral-tradeoffs in trustworthy ML. *IEEE access : practical innovations, open solutions*, 10:120850–120865, 2022.
- [60] Raef Bassily, Adam Smith, and Abhradeep Thakurta. Private empirical risk minimization: Efficient algorithms and tight error bounds. In *2014 IEEE 55th Annual Symposium on Foundations of Computer Science*, pages 464–473. IEEE, 2014.
- [61] Di Wang, Minwei Ye, and Jinhui Xu. Differentially private empirical risk minimization revisited: Faster and more general. *Advances in Neural Information Processing Systems*, 30, 2017.
- [62] Ali Shafahi, W Ronny Huang, Christoph Studer, Soheil Feizi, and Tom Goldstein. Are adversarial examples inevitable? *arXiv preprint arXiv:1809.02104*, 2018.
- [63] Dimitris Tsipras, Shibani Santurkar, Logan Engstrom, Alexander Turner, and Aleksander Madry. Robustness may be at odds with accuracy. *arXiv preprint arXiv:1805.12152*, 2018.
- [64] Cynthia Dwork and Jing Lei. Differential privacy and robust statistics. In *Proceedings of the forty-first annual ACM symposium on Theory of computing*, pages 371–380, 2009.
- [65] Mathias Lecuyer, Vaggelis Atlidakis, Roxana Geambasu, Daniel Hsu, and Suman Jana. On the connection between differential privacy and adversarial robustness in machine learning. *stat*, 1050(9), 2018.
- [66] Mathias Lecuyer, Vaggelis Atlidakis, Roxana Geambasu, Daniel Hsu, and Suman Jana. Certified robustness to adversarial examples with differential privacy. In *2019 IEEE symposium on security and privacy (SP)*, pages 656–672. IEEE, 2019.
- [67] Hai Phan, My T Thai, Han Hu, Ruoming Jin, Tong Sun, and Dejing Dou. Scalable differential privacy with certified robustness in adversarial learning. In *International Conference on Machine Learning*, pages 7683–7694. PMLR, 2020.
- [68] Nurislam Tursynbek, Aleksandr Petushko, and Ivan Oseledets. Robustness threats of differential privacy. *arXiv preprint arXiv:2012.07828*, 2020.
- [69] Franziska Boenisch, Philip Sperl, and Konstantin Böttinger. Gradient masking and the underestimated robustness threats of differential privacy in deep learning. *arXiv preprint arXiv:2105.07985*, 2021.
- [70] Liwei Song, Reza Shokri, and Prateek Mittal. Privacy risks of securing machine learning models against adversarial examples. In *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security*, pages 241–257, 2019.
- [71] Fengxiang He, Shaopeng Fu, Bohan Wang, and Dacheng Tao. Robustness, privacy, and generalization of adversarial training. *arXiv preprint arXiv:2012.13573*, 2020.
- [72] Aolin Xu and Maxim Raginsky. Information-theoretic analysis of generalization capability of learning algorithms. *Advances in neural information processing systems*, 30, 2017.

- [73] Samuel Yeom, Irene Giacomelli, Matt Fredrikson, and Somesh Jha. Privacy risk in machine learning: Analyzing the connection to overfitting. In *2018 IEEE 31st computer security foundations symposium (CSF)*, pages 268–282. IEEE, IEEE, 2018.
- [74] Trevor Hastie, Robert Tibshirani, and Martin Wainwright. Statistical learning with sparsity. *Monographs on statistics and applied probability*, 143(143):8, 2015.
- [75] Vladimir N Vapnik. An overview of statistical learning theory. *IEEE transactions on neural networks*, 10(5):988–999, 1999.
- [76] Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.
- [77] Elaine Angelino, Nicholas Larus-Stone, Daniel Alabi, Margo Seltzer, and Cynthia Rudin. Learning certifiably optimal rule lists for categorical data. *Journal of Machine Learning Research*, 18:1–78, 2018.
- [78] Alex Kantchelian, J Doug Tygar, and Anthony Joseph. Evasion and hardening of tree ensemble classifiers. In *International conference on machine learning*, pages 2387–2396. PMLR, 2016.
- [79] Julien Ferry, Ricardo Fukasawa, Timothée Pascal, and Thibaut Vidal. Trained random forests completely reveal your dataset. In *International Conference on Machine Learning*, pages 13545–13569. PMLR, 2024.
- [80] Dheeru Dua and Casey Graff. UCI machine learning repository, 2017. URL <http://archive.ics.uci.edu/ml>.
- [81] J. Larson, S. Mattu, L. Kirchner, and J. Angwin. How we analyzed the COMPAS recidivism algorithm. *ProPublica*, 2016.
- [82] FICO, Google, Imperial College London, MIT, University of Oxford, UC Irvine, and UC Berkeley. Explainable Machine Learning Challenge, 2018.
- [83] Hayden McTavish, Chudi Zhong, Reto Achermann, Ilias Karimalis, Jacques Chen, Cynthia Rudin, and Margo Seltzer. Fast sparse decision tree optimization via reference ensembles. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 9604–9613, 2022.

A Proofs for Theoretical Results in Section 4

In this appendix, we provide proofs for the theoretical results presented in Section 4.

Theorem 2 (Inherent vulnerability of single models). *Consider a binary dataset $S = \{(x_i, y_i)\}_{i=1}^n$ that has binary features and binary labels, where n_+ denotes the number of data points with positive labels in S . Let $d = (d_p, \delta_p, q_0, K)$ be a rule list such that $q_1 = 0$ and each rule predicts the majority label of the points captured by that rule. Further, let I be the smallest index i such that $q_i = 1$. Let $S' = \{x'_i, y'_i\}_{i=1}^n$ be an adversarial dataset constructed by flipping up to one feature in each x_i (i.e., an L_0 -bounded perturbation with $\eta = 1$ restricted to binary features). Let \hat{L} be the 0-1 loss. If \bar{n}_+ is the number of positive data points captured by one of the first $I - 1$ leaves, then $\hat{L}_{S'}(d) - \hat{L}_S(d) \geq \frac{n_+ - \bar{n}_+}{n}$.*

Proof. Let a_k and b_k denote the number of negative (zero) and positive points captured by rule k . Then, note that all data points captured by rules with index at least I are vulnerable to the adversarial attack. Thus, we have the following:

$$\begin{aligned} K_+ &= \sum_{k=1}^K b_k \\ \hat{L}_S(d) &= \frac{1}{n} \left[\sum_{k=1}^{I-1} b_k + \sum_{k=I}^K (a_k q_k + b_k (1 - q_k)) \right] \\ \hat{L}_{S'}(d) &\geq \frac{1}{n} \left[\sum_{k=1}^{I-1} b_k + \sum_{k=I}^K (a_k + b_k) \right] \\ \bar{K}_+ &= \sum_{k=1}^{I-1} b_k. \end{aligned}$$

From here, note that $q_k = \mathbb{1}_{a_k \leq b_k}$ since each rule predicts the majority label. Then, we can write

$$\begin{aligned} (K_+ - \bar{K}_+) + (n\hat{L}_S(d) - \bar{K}_+) &= \sum_{k=I}^K b_k + \sum_{k=I}^K (a_k q_k + b_k (1 - q_k)) \\ &= \sum_{k=I}^K (a_k + b_k) + \sum_{k=I}^K (1 - q_k)(b_k - a_k) \\ &\leq \sum_{k=I}^K (a_k + b_k) \\ &\leq n\hat{L}_{S'}(d) - \bar{K}_+, \end{aligned}$$

where $(1 - q_k)(b_k - a_k) = \mathbb{1}_{a_k > b_k}(b_k - a_k) \leq 0$. Therefore, we get that $\hat{L}_{S'}(d) - \hat{L}_S(d) \geq \frac{K_+ - \bar{K}_+}{n}$. \square

The theorem above states that there is an inherent cost to robustness when attacking a single model. Therefore, having the whole Rashomon set can be useful for model selection if adversarial attacks are expected.

Theorem 1 (Sparsity controls mutual information in a single tree). *Let $S = \{(x_i, y_i)\}_{i=1}^n$ be a dataset of n i.i.d. samples from distribution \mathcal{D} over $\mathcal{X} \times \mathcal{Y}$, where $\mathcal{X} = \{0, 1\}^d$ and $\mathcal{Y} = \{0, 1\}$. Let \mathcal{F} be the class of*

binary classification decision trees with l_f leaves, and let $f \in \mathcal{F}$ be a tree fit on S through a possibly-random training algorithm. Then the mutual information between the learned tree f and the dataset S satisfies: $I(f; S) = O(l_f \log d)$.

Proof. The mutual information between the model f and the dataset S can be expressed as $I(f, S) = H(f) - H(f|S)$, and the upper bound $I(f, S) \leq H(f)$. It is well known that entropy is bounded by the logarithm of the size of the alphabet, so we next bound the size of the hypothesis space, $|\mathcal{F}|$.

Let $T(l_f)$ denote the number of binary trees with l_f leaves. A full binary tree with l_f leaves has exactly $l_f - 1$ internal nodes and $2l_f - 1$ total nodes. It is well known that the number of such tree shapes is the $(l_f - 1)$ -st Catalan number, so $T(l_f) = C_{l_f-1}$.

To bound $|\mathcal{F}|$, note that each decision tree can be obtained by taking a binary tree, assigning splits to each internal node, and assigning a prediction to each leaf. Each internal node has at most d features to split on, and it selects a direction (less than or greater than). Each leaf node can likewise choose to either predict 0 or 1. Thus, since there are $l_f - 1$ internal nodes with at most $2d$ choices at each internal node, and there are l_f leaves with 2 choices at each leaf, we have that $|\mathcal{F}| \leq T(l_f) \cdot (2d)^{l_f-1} 2^{l_f} = C_{l_f-1} \cdot (2d)^{l_f-1} 2^{l_f}$. Since C_{l_f-1} grows at most exponentially in l_f , we have that $\log |\mathcal{F}| = O(l_f \log d)$. Thus, $I(f; S) = O(l_f \log d)$. \square

B Proofs for Theoretical Results in Section 5

In this appendix, we provide proofs for the theoretical results presented in Section 5.

B.1 Proof for Theorem 4

We state and prove Theorem 4 below. Note that for the purposes of the next theorem, the model belonging to the Rashomon set is based on the risk $L_{\mathcal{D}}$, meaning that $\mathcal{R}(\epsilon) = \{f \in \mathcal{F} : L(f) \leq L(f^*) + \epsilon\}$, where f^* is optimal model.

Theorem 4 (Rashomon set is robust under small distribution shift). *Consider bounded loss function ϕ , $\phi \in [0, 1]$, and two data distributions \mathcal{D} and \mathcal{D}' , such that $\mathcal{D}(x) \neq \mathcal{D}'(x)$ and $\mathcal{D}(y|x) = \mathcal{D}'(y|x)$. If $KL(\mathcal{D}\|\mathcal{D}') \leq \frac{\epsilon^2}{8}$, then if a function f is in the true Rashomon set for the data distribution \mathcal{D} , $f \in \mathcal{R}_{z \sim \mathcal{D}}(\frac{\epsilon}{2})$, it is also in the true Rashomon set for the data distribution \mathcal{D}' , $f \in \mathcal{R}_{z \sim \mathcal{D}'}(\epsilon)$.*

Proof. Given the bounded loss function $\phi(f(x), y)$ for a data point $z = (x, y)$, we will overload definition of the loss and also consider $\phi(f, z) : \mathcal{F} \times (\mathcal{X} \times \mathcal{Y}) \rightarrow [0, 1]$. Let $f_{\mathcal{D}}^* = \arg \inf_{g \in \mathcal{F}} \mathbb{E}_{z \sim \mathcal{D}}[\phi(g, z)]$ and $f_{\mathcal{D}'}^* = \arg \inf_{g \in \mathcal{F}} \mathbb{E}_{z \sim \mathcal{D}'}[\phi(g, z)]$ be optimal models for distributions \mathcal{D} and \mathcal{D}' respectively. Also, let $d_{TV}(\mathcal{D}, \mathcal{D}')$ be the total variational distance defined as:

$$d_{TV}(\mathcal{D}, \mathcal{D}') = \sup_{B \in \mathcal{B}} |\Pr_{\mathcal{D}}[B] - \Pr_{\mathcal{D}'}[B]| = \frac{1}{2} \int |p_{\mathcal{D}}(z) - p_{\mathcal{D}'}(z)| dz,$$

where \mathcal{B} is the set of measurable subsets under \mathcal{D} and \mathcal{D}' . Then since $\phi(\cdot, \cdot) \in [0, 1]$, we have $|\phi(g, z) - 1/2| \leq$

$1/2$ for any $g \in \mathcal{F}$. Thus:

$$\begin{aligned}
|\mathbb{E}_{z \sim \mathcal{D}} \phi(g, z) - \mathbb{E}_{z \sim \mathcal{D}'} \phi(g, z)| &= \left| \int_z \phi(g, z) (p_{\mathcal{D}}(z) - p_{\mathcal{D}'}(z)) dz \right| \\
&= \left| \int_z (\phi(g, z) - \frac{1}{2})(p_{\mathcal{D}}(z) - p_{\mathcal{D}'}(z)) dz \right| \\
&\leq \int_z |\phi(g, z) - \frac{1}{2}| \cdot |p_{\mathcal{D}}(z) - p_{\mathcal{D}'}(z)| dz \\
&\leq \int_z \frac{1}{2} \cdot |p_{\mathcal{D}}(z) - p_{\mathcal{D}'}(z)| dz \\
&= \frac{1}{2} \int |p_{\mathcal{D}}(z) - p_{\mathcal{D}'}(z)| dz \\
&= d_{TV}(\mathcal{D}, \mathcal{D}') \tag{2}
\end{aligned}$$

Note that (2) holds for $g = f$ and $g = f_{\mathcal{D}}^*$ as well.

By Pinsker's inequality $d_{TV}(\mathcal{D}, \mathcal{D}') \leq \sqrt{\frac{1}{2} KL(\mathcal{D} \parallel \mathcal{D}')}}$. Since $f \in \mathcal{R}_{z \sim \mathcal{D}}(\frac{\epsilon}{2})$ and $KL(\mathcal{D} \parallel \mathcal{D}') \leq \frac{\epsilon^2}{8}$, we have that $d_{TV}(\mathcal{D}, \mathcal{D}') \leq \sqrt{\frac{1}{2} KL(\mathcal{D} \parallel \mathcal{D}')} \leq \sqrt{\frac{1}{2} \frac{\epsilon^2}{8}} = \sqrt{\frac{\epsilon^2}{16}} = \frac{\epsilon}{4}$. Therefore, we can bound the expected risks' difference for distribution \mathcal{D}' :

$$\begin{aligned}
\mathbb{E}_{z \sim \mathcal{D}'} \phi(f, z) - \mathbb{E}_{z \sim \mathcal{D}'} \phi(f_{\mathcal{D}'}^*, z) &= (\mathbb{E}_{z \sim \mathcal{D}} \phi(f, z) - \mathbb{E}_{z \sim \mathcal{D}} \phi(f_{\mathcal{D}}^*, z)) \\
&\quad + (\mathbb{E}_{z \sim \mathcal{D}'} \phi(f, z) - \mathbb{E}_{z \sim \mathcal{D}} \phi(f, z)) \\
&\quad + (\mathbb{E}_{z \sim \mathcal{D}} \phi(f_{\mathcal{D}}^*, z) - \mathbb{E}_{z \sim \mathcal{D}'} \phi(f_{\mathcal{D}'}^*, z)) \\
&\leq \frac{\epsilon}{2} \quad (\text{Since } f \in \mathcal{R}_{z \sim \mathcal{D}}(\frac{\epsilon}{2})) \\
&\quad + |\mathbb{E}_{z \sim \mathcal{D}'} \phi(f, z) - \mathbb{E}_{z \sim \mathcal{D}} \phi(f, z)| + (\mathbb{E}_{z \sim \mathcal{D}} \phi(f_{\mathcal{D}}^*, z) - \mathbb{E}_{z \sim \mathcal{D}'} \phi(f_{\mathcal{D}'}^*, z)) \\
&\leq \frac{\epsilon}{2} + |\mathbb{E}_{z \sim \mathcal{D}'} \phi(f, z) - \mathbb{E}_{z \sim \mathcal{D}} \phi(f, z)| \\
&\quad + (\mathbb{E}_{z \sim \mathcal{D}} \phi(f_{\mathcal{D}}^*, z) - \mathbb{E}_{z \sim \mathcal{D}'} \phi(f_{\mathcal{D}}^*, z)) \quad (\text{Since } f_{\mathcal{D}'}^* \text{ minimizes for } \mathcal{D}') \\
&\leq \frac{\epsilon}{2} + |\mathbb{E}_{z \sim \mathcal{D}'} \phi(f, z) - \mathbb{E}_{z \sim \mathcal{D}} \phi(f, z)| + |\mathbb{E}_{z \sim \mathcal{D}} \phi(f_{\mathcal{D}}^*, z) - \mathbb{E}_{z \sim \mathcal{D}'} \phi(f_{\mathcal{D}}^*, z)| \\
&\leq \frac{\epsilon}{2} + d_{TV}(\mathcal{D}, \mathcal{D}') + d_{TV}(\mathcal{D}, \mathcal{D}') \quad (\text{Using result from (2)}) \\
&= \frac{\epsilon}{2} + 2d_{TV}(\mathcal{D}, \mathcal{D}') \leq \frac{\epsilon}{2} + 2 \left(\frac{\epsilon}{4} \right) \quad (\text{Since } d_{TV}(\mathcal{D}, \mathcal{D}') \leq \frac{\epsilon}{4}) \\
&= \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon.
\end{aligned}$$

Therefore, $f \in \mathcal{R}_{z \sim \mathcal{D}'}(\epsilon)$. \square

Note that for ridge regression, we can derive a more specific condition on the total variational distance as we show in Lemma 1.

B.2 Proof for Theorem 5

We state and prove Theorem 5 below. Note that while the theorem is proved for the objective defined Rashomon set, the statement holds for the risk defined Rashomon set when the regularization parameter is

zero.

Theorem 5 (Two Rashomon sets constructed on neighboring datasets are indistinguishable). *For 0-1 loss, let S and S' be two datasets, each of size n . Let $\hat{\mathcal{R}}_S(\epsilon) := \{f \in \mathcal{F} | \hat{obj}(f, S) \leq \hat{obj}(\hat{f}, S) + \epsilon\}$ and $\hat{\mathcal{R}}_{S'}(\epsilon) := \{f \in \mathcal{F} | \hat{obj}(f, S') \leq \hat{obj}(\hat{f}', S') + \epsilon\}$. Suppose S and S' differ in at most K samples. Then $\hat{\mathcal{R}}_S(\epsilon) \subseteq \hat{\mathcal{R}}_{S'}(\epsilon + \frac{2K}{n})$ and $\hat{\mathcal{R}}_{S'}(\epsilon) \subseteq \hat{\mathcal{R}}_S(\epsilon + \frac{2K}{n})$.*

Proof. For any model $f \in \mathcal{F}$, the difference in 0-1 loss due to change of the dataset from S to S' that differ in at most K samples is at most $\frac{K}{n}$. Hence,

$$|\hat{obj}_S(f) - \hat{obj}_{S'}(f)| = |\hat{L}_S(f) - \hat{L}_{S'}(f)| \leq \frac{K}{n}. \quad (3)$$

By definition, since \hat{f} is ERM, $\hat{obj}_S(\hat{f}) \leq \hat{obj}_S(\hat{f}')$. Plugging in \hat{f}' in (3), we get that:

$$\hat{obj}_S(\hat{f}') \leq \hat{obj}_{S'}(\hat{f}') + \frac{K}{n}. \text{ Therefore, we get that}$$

$$\hat{obj}_{S'}(\hat{f}') \geq \hat{obj}_S(\hat{f}') - \frac{K}{n} \geq \hat{obj}_S(\hat{f}) - \frac{K}{n}. \quad (4)$$

Next, we show for the empirical Rashomon sets that $\hat{\mathcal{R}}_S(\epsilon) \subseteq \hat{\mathcal{R}}_{S'}(\epsilon + \frac{2K}{n})$. For any model $f \in \hat{\mathcal{R}}_S(\epsilon)$, by definition, $\hat{obj}_S(f) \leq \hat{obj}_S(\hat{f}) + \epsilon$. Based on (3) we get that,

$$\hat{obj}_{S'}(f) \leq \hat{obj}_S(f) + \frac{K}{n} \leq \hat{obj}_S(\hat{f}) + \epsilon + \frac{K}{n}. \quad (5)$$

Finally, combining (4) and (5) together, we see that

$$\hat{obj}_{S'}(f) - \hat{obj}_{S'}(\hat{f}') \leq \hat{obj}_S(\hat{f}) + \epsilon + \frac{K}{n} - \hat{obj}_S(\hat{f}) + \frac{K}{n} = \epsilon + \frac{2K}{n}, \quad (6)$$

which means that $f \in \hat{\mathcal{R}}_{S'}(\epsilon + \frac{2K}{n})$ and correspondingly, $\hat{\mathcal{R}}_S(\epsilon) \subseteq \hat{\mathcal{R}}_{S'}(\epsilon + \frac{2K}{n})$. Following similar logic we can show that $\hat{\mathcal{R}}_{S'}(\epsilon) \subseteq \hat{\mathcal{R}}_S(\epsilon + \frac{2K}{n})$ yielding the statement of the theorem. \square

B.3 Proofs for Theorem 3 and Corollaries 1 and 2

We state and prove Theorem 3 as well as two corollaries from it below.

Theorem 3. *Suppose that $\hat{L}_S(w) = \frac{1}{n} \sum_{i=1}^n \phi(y_i \cdot w^T x_i)$ where ϕ is a loss that is a function of the margin ($y_i f(x_i)$). For an L_2 attack on the optimal model \hat{w}_S with budget η , the loss of w on the adversarial dataset S' is $\hat{L}_{S'}(w) = \frac{1}{n} \sum_{i=1}^n \phi(y_i \cdot w^T x_i - \eta \|w\|_2 \cos(w, \hat{w}_S))$.*

Proof. For each sample x_i , under the adversarial attack, we have that $x'_i = x_i + \delta_i$ where $\delta_i = -\eta y_i \frac{\hat{w}_S}{\|\hat{w}_S\|_2}$. Then, the margin for a given model w on the adversarial input x_i is:

$$\begin{aligned} y_i \cdot w^T x'_i &= y_i \cdot w^T (x_i + \delta_i) \\ &= y_i \cdot w^T x_i + y_i \cdot w^T \delta_i \\ &= y_i \cdot w^T x_i + y_i \cdot w^T \left(-\eta y_i \frac{\hat{w}_S}{\|\hat{w}_S\|_2} \right) \\ &= y_i \cdot w^T x_i - \eta y_i^2 \frac{w^T \hat{w}_S}{\|\hat{w}_S\|_2} \\ &= y_i \cdot w^T x_i - \eta \frac{w^T \hat{w}_S}{\|\hat{w}_S\|_2} \\ &= y_i \cdot w^T x_i - \eta \|w\|_2 \cos(w, \hat{w}_S). \end{aligned}$$

Therefore, we get that $\hat{L}_{S'}(w) = \frac{1}{n} \sum_{i=1}^n \phi(y_i \cdot w^T x_i - \eta \|w\|_2 \cos(w, \hat{w}_S))$. \square

In the next corollary, we show that if the model $w^T x$ is closer to the ERM model, then the adversarial attack has more effect on it in terms of the loss.

Corollary 1. Suppose that we have the exponential loss $\phi(y \cdot w^T x) = e^{-y \cdot w^T x}$. Then, for any unit weights w_1 and w_2 satisfying $\cos(w_1, \hat{w}_S) > \cos(w_2, \hat{w}_S)$, we have that $\frac{\hat{L}_{S'}(w_1)}{\hat{L}_S(w_1)} > \frac{\hat{L}_{S'}(w_2)}{\hat{L}_S(w_2)}$. In other words, the adversarial attack is most effective on models more similar to \hat{w}_S .

Proof. For any unit vector w , we have by Theorem 3 that

$$\hat{L}_{S'}(w) = \frac{1}{n} \sum_{i=1}^n \exp(-y_i \cdot w^T x_i + \eta \cos(w, \hat{w}_S)) = \exp(\eta \cos(w, \hat{w}_S)) \hat{L}_S(w).$$

Then, we can write

$$\begin{aligned} \frac{\hat{L}_{S'}(w_1)}{\hat{L}_S(w_1)} &= \frac{\exp(\eta \cos(w_1, \hat{w}_S)) \hat{L}_S(w_1)}{\hat{L}_S(w_1)} \\ &= \exp(\eta \cos(w_1, \hat{w}_S)) \\ &> \exp(\eta \cos(w_2, \hat{w}_S)) \\ &= \frac{\hat{L}_{S'}(w_2)}{\hat{L}_S(w_2)} \end{aligned}$$

\square

Our next corollary makes this point even more explicit in the case of a strong attack, showing that if the diversity within the Rashomon set is higher in terms of angular distance, then models with higher adversarial risk can exist.

Corollary 2. Suppose that $\phi(\cdot)$ is decreasing with respect to the margin, and let w_1 and w_2 be unit weights so that $\cos(w_1, \hat{w}_S) > \cos(w_2, \hat{w}_S)$. Then, for large enough η (e.g. when $\eta \rightarrow \infty$), $\hat{L}_{S'}(w_1) > \hat{L}_{S'}(w_2)$.

Proof. Since S is finite, it suffices to show that, for each i , there exists some N_i such that

$$\phi(y_i \cdot w_1^T x_i - \eta \cos(w_1, \hat{w}_S)) > \phi(y_i \cdot w_2^T x_i - \eta \cos(w_2, \hat{w}_S))$$

for all $\eta \geq N_i$. However, this is true since $\eta \cos(w_1, \hat{w}_S)$ is arbitrarily larger than $\eta \cos(w_2, \hat{w}_S)$ as $\eta \rightarrow \infty$, so $y_i \cdot w_1^T x_i - \eta \cos(w_1, \hat{w}_S) < y_i \cdot w_2^T x_i - \eta \cos(w_2, \hat{w}_S)$ for sufficiently large η . \square

Next, we focus on a different loss function for linear models, specifically the least-squares loss used in regression tasks.

B.4 Proof for Theorem 7

For our second setting, we consider least-squares regression, where now the loss is $\phi(f(x), y) = (f(x) - y)^2$. Our hypothesis space is still linear models $w^T x$. Then the Rashomon set is ellipsoid $(w - \hat{w}) \mathbb{E}[xx^T] (w - \hat{w})^T \leq \epsilon$, where singular values of matrix $\mathbb{E}[xx^T]$ determine its shape. Let $TV(\mathcal{D}, \mathcal{D}')$ be the total variational distance between two distributions, then under label shift, we can bound the minimum singular value σ_{min} of the matrix $\mathbb{E}[xx^T]$ as follows:

Theorem 7. Let \mathcal{D} and \mathcal{D}' be two data distributions in $\mathcal{X} \times \mathcal{Y}$ such that $\mathcal{D}(x) = \mathcal{D}'(x)$ but $\mathcal{D}(y|x) \neq \mathcal{D}'(y|x)$. Furthermore, suppose that $y \in [a, b]$ for all $y \in \mathcal{Y}$ and suppose that $\|x\|_2 \leq C$ for all $x \in \mathcal{X}$. Then, there exists some model in both true Rashomon sets $\mathcal{R}_{\mathcal{D}}(\epsilon)$ and $\mathcal{R}_{\mathcal{D}'}(\epsilon)$ if

$$TV(\mathcal{D}, \mathcal{D}') \leq \frac{2\sqrt{\epsilon}}{(b-a) \cdot C} \cdot \sqrt{\sigma_{\min}(E[xx^T])}.$$

To prove Theorem 7, we use Lemma 1 which we prove below.

Lemma 1. Suppose that $y \in [a, b]$ for all $y \in \mathcal{Y}$, and suppose that $\|x\|_2 \leq C$ for all $x \in \mathcal{X}$. Furthermore, suppose that we undergo a distribution shift to \mathcal{D}' where $\mathcal{D}(x) = \mathcal{D}'(x)$ but $\mathcal{D}(y|x) \neq \mathcal{D}'(y|x)$. Then, if $\hat{w}_{\mathcal{D}'}$ is the optimal linear model for \mathcal{D}' under the least-squares objective, we have that

$$\|E_{x \sim \mathcal{X}}[xx^T](\hat{w}_{\mathcal{D}} - \hat{w}_{\mathcal{D}'})\|_2 \leq (b-a) \cdot C \cdot TV(\mathcal{D}, \mathcal{D}').$$

Proof. For a data distribution \mathcal{D} , the optimal weights for a linear model under the least-squares objective is

$$\hat{w}_{\mathcal{D}} = (E_{x \sim \mathcal{X}}[xx^T])^{-1} E_{(x,y) \sim \mathcal{D}}[xy]. \quad (7)$$

Then, we can write

$$\begin{aligned} \|E[xx^T](\hat{w}_{\mathcal{D}} - \hat{w}_{\mathcal{D}'})\|_2 &= \|E_{(x,y) \sim \mathcal{D}}[xy] - E_{(x,y) \sim \mathcal{D}'}[xy]\|_2 \\ &= \left\| \int_x x \mathcal{D}(x) \int_y y (\mathcal{D}(y|x) - \mathcal{D}'(y|x)) \right\|_2 \\ &= \left\| \int_x x \mathcal{D}(x) \int_y \left(y - \frac{a+b}{2}\right) (\mathcal{D}(y|x) - \mathcal{D}'(y|x)) \right\|_2 \\ &\leq \int_x \|x\|_2 \mathcal{D}(x) \int_y \left|y - \frac{a+b}{2}\right| |\mathcal{D}(y|x) - \mathcal{D}'(y|x)| \\ &\leq \frac{b-a}{2} \cdot \int_x \|x\|_2 \mathcal{D}(x) \int_y |\mathcal{D}(y|x) - \mathcal{D}'(y|x)| \\ &\leq \frac{(b-a) \cdot C}{2} \cdot \int_x \mathcal{D}(x) \int_y |\mathcal{D}(y|x) - \mathcal{D}'(y|x)| \\ &= \frac{(b-a) \cdot C}{2} \cdot \int_x \int_y |\mathcal{D}(x, y) - \mathcal{D}'(x, y)| \\ &= (b-a) \cdot C \cdot TV(\mathcal{D}, \mathcal{D}') \end{aligned}$$

as desired. \square

Now we prove Theorem 7, which bounds the total variational distance with the value of the minimum singular value of the expected data matrix.

Proof. From Theorem 10 of [2], we know that $\mathcal{R}_{\mathcal{D}}(\epsilon)$ is the ellipsoid described by the equation

$$(w - \hat{w}_{\mathcal{D}})^T \frac{E[xx^T]}{\epsilon} (w - \hat{w}_{\mathcal{D}}) \leq 1.$$

We have a similar equation for $R_{\mathcal{D}'}(\epsilon)$. Equivalently, if $M = (\frac{E[xx^T]}{\epsilon})^{\frac{1}{2}}$ and $S(0, 1)$ is the unit ball centered at the origin, then $\mathcal{R}_{\mathcal{D}}(\epsilon) = M^{-1}S(0, 1) + \hat{w}_{\mathcal{D}} = \{w: \|M(w - \hat{w}_{\mathcal{D}})\|_2 \leq 1\}$. Consider $\bar{w} = \frac{\hat{w}_{\mathcal{D}} + \hat{w}_{\mathcal{D}'}}{2}$, then

$$\|M(\bar{w} - \hat{w}_{\mathcal{D}})\|_2 = \left\| \frac{M(\hat{w}_{\mathcal{D}'} - \hat{w}_{\mathcal{D}})}{2} \right\|_2 = \frac{1}{2} \|M(\hat{w}_{\mathcal{D}'} - \hat{w}_{\mathcal{D}})\|_2.$$

Next we will show that $\bar{w} \in \mathcal{R}_{\mathcal{D}}(\epsilon)$. If the bound on the total variational distance in the theorem assumption holds and given Lemma 1, we have that:

$$\begin{aligned} \|M(\hat{w}_{\mathcal{D}'} - \hat{w}_{\mathcal{D}})\|_2 &= \left\| \frac{1}{\sqrt{\epsilon}} E[xx^T]^{-\frac{1}{2}} E[xx^T](\hat{w}_{\mathcal{D}'} - \hat{w}_{\mathcal{D}}) \right\|_2 \\ &\leq \frac{1}{\sqrt{\epsilon}} \left\| E[xx^T]^{-\frac{1}{2}} \right\|_2 \|E[xx^T](\hat{w}_{\mathcal{D}'} - \hat{w}_{\mathcal{D}})\|_2 \\ &= \frac{1}{\sqrt{\epsilon} \sqrt{\sigma_{\min}(E[xx^T])}} \|E[xx^T](\hat{w}_{\mathcal{D}'} - \hat{w}_{\mathcal{D}})\|_2 \\ &\leq \frac{(b-a) \cdot C}{\sqrt{\epsilon} \sqrt{\sigma_{\min}(E[xx^T])}} \cdot TV(\mathcal{D}, \mathcal{D}') \\ &\leq 2. \end{aligned}$$

Therefore,

$$\|M(\bar{w} - \hat{w}_{\mathcal{D}})\|_2 = \frac{1}{2} \|M(\hat{w}_{\mathcal{D}'} - \hat{w}_{\mathcal{D}})\|_2 \leq 1,$$

which means that $w \in \mathcal{R}_{\mathcal{D}}(\epsilon)$. Similar argument shows that $w \in R_{\mathcal{D}'}(\epsilon)$, which proves the theorem. \square

Theorem 7 means that for a model to remain robust across larger distribution shifts (i.e., increased $TV(\mathcal{D}, \mathcal{D}')$), the data's covariance matrix $E[\mathbf{x}\mathbf{x}^\top]$ must have a higher minimum singular value, $\sigma_{\min}(E[\mathbf{x}\mathbf{x}^\top])$. A relatively high $\sigma_{\min}(E[\mathbf{x}\mathbf{x}^\top])$, particularly when the overall spectrum of singular values for $E[\mathbf{x}\mathbf{x}^\top]$ is well-conditioned contributes to the “roundness” of the Rashomon set. In turn, this roundness can be viewed as a form of structural diversity, meaning that the set contains models whose parameters reflect more uniform importance across different feature directions. Therefore, such diverse sets are more likely to contain models that can be selected for their robustness.

B.5 Diverse ensemble of models from the Rashomon set is more adversarially robust

The proofs in Section 5.1 shows that models that are diverse (e.g. rely on the different logic) are less vulnerable to the adversarial attack of the optimal model. We can generalize this intuition to an ensemble. More specifically, we consider a majority-vote ensemble of models from the Rashomon set. First, we consider independent models in Theorem 8. If the models are sufficiently diverse, their failures on a given adversarial input might be treated as largely independent events. We relax this assumption to allow weak correlations between models in Theorem 9.

Theorem 8 (Independent ensemble). *Let $\{f_1, f_2, \dots, f_k\}$ be a subset of models in the Rashomon set where k is odd (to prevent ties). Let $\delta: \mathbb{R}^d \rightarrow \mathbb{R}^d$ be a function that takes in a data point and outputs a (possibly random) perturbation for that point. For a random data point (x, y) drawn from the distribution \mathcal{D} , let $Z_i = \mathbf{1}_{[f_i(x+\delta(x)) \neq y]}$ be a random variable indicating whether model f_i predicts the perturbed data point incorrectly. Let $p_i = \Pr_{(x,y)}(Z_i = 1)$. Assume that there exists $p < \frac{1}{2}$ such that $p_i \leq p$ for all i . Let*

$S_k = \sum_{i=1}^k Z_i$ be the random count of individual models that the attack fools on a single input. Since k is odd, the probability $\Pr_{(x,y)}(S_k \geq k/2)$ is exactly the chance that at least half of the k models are wrong, which means the majority-vote ensemble is also wrong on that adversarially-perturbed input. Suppose Z_1, Z_2, \dots, Z_k are independent. Then,

$$\Pr_{(x,y)}(S_k \geq k/2) \leq e^{-kD_{KL}(\frac{1}{2}||p)}.$$

Proof. Based on the Chernoff bound, we can get, for $t > 0$,

$$\Pr(S_k \geq k/2) = \Pr(e^{tS_k} \geq e^{tk/2}) \leq \mathbb{E}[e^{tS_k}]e^{-tk/2}.$$

Since, Z'_i s are independent,

$$\mathbb{E}[e^{tS_k}] = \mathbb{E}[e^{t(\sum_{i=1}^k Z_i)}] = \mathbb{E}[\prod_{i=1}^k e^{tZ_i}] = \prod_{i=1}^k \mathbb{E}[e^{tZ_i}].$$

Since Z_i is a Bernoulli variable, $e^{tZ_i} = 1 + (e^t - 1)Z_i$. Then,

$$\mathbb{E}[e^{tS_k}] = \prod_{i=1}^k \mathbb{E}[1 + (e^t - 1)Z_i] = \prod_{i=1}^k (1 + (e^t - 1)\mathbb{E}[Z_i]) = \prod_{i=1}^k (1 + (e^t - 1)p_i).$$

Since $p_i \leq p, \forall i$,

$$\mathbb{E}[e^{tS_k}] = \prod_{i=1}^k (1 + (e^t - 1)p_i) \leq \prod_{i=1}^k (1 + (e^t - 1)p) = (1 + (e^t - 1)p)^k.$$

Then, $\mathbb{E}[e^{tS_k}]e^{-tk/2} \leq (1 - p + pe^t)^k e^{-tk/2}$. Let $h(t) = \ln(1 - p + pe^t) - \frac{t}{2}$. Then,

$$(1 - p + pe^t)^k e^{-tk/2} = e^{kh(t)}.$$

Since the bound holds for any $t > 0$, let's find the value of t that gives us the tightest bound.

$$\frac{dh}{dt} = \frac{pe^t}{1 - p + pe^t} - \frac{1}{2}.$$

Set it to 0, we get $t^* = \ln \frac{1-p}{p}, e^{t^*} = \frac{1-p}{p}$. Then we know,

$$\begin{aligned} h(t^*) &= \ln(1 - p + 1 - p) - \frac{1}{2} \ln \frac{1-p}{p} \\ &= \ln 2 + \ln(1 - p) - \frac{1}{2} \ln(1 - p) + \frac{1}{2} \ln p \\ &= \ln 2 + \frac{1}{2} \ln((1 - p)p) \\ &= \ln(2\sqrt{(1 - p)p}). \end{aligned}$$

Now, we know $\mathbb{E}[e^{tS_k}]e^{-tk/2} \leq (1 - p + pe^{t^*})^k e^{-t^*k/2} = e^{kh(t^*)}$. Since $p \in [0, 1/2]$, $(1 - p)p \in [0, 1/4]$ and $h(t^*) \in (-\infty, 0)$. Hence, $e^{kh(t^*)}$ decreases as k increases.

Also, note that $KL(\frac{1}{2}||p) = -\ln(2\sqrt{(1 - p)p})$. Therefore, $e^{kh(t^*)} = e^{-k \cdot KL(\frac{1}{2}||p)}$. \square

Theorem 8 shows that even if individual models have a non-trivial probability of being fooled by an attack, the probability that a majority-vote ensemble of these models fails decreases exponentially with the size of the ensemble.

We can also consider the dependent case, when there is correlation between models in the ensemble to get similar bounds:

Theorem 9 (Dependent ensemble - this is not helpful). *Let $\{f_1, f_2, \dots, f_k\}$ be a subset of models in the Rashomon set, $k > 2$, and k is odd (to prevent ties). Let $Z_i = \mathbf{1}_{[f_i(x+\delta) \neq y]}$ be a random variable whether model f_i can predict a perturbed data incorrectly for fixed $\delta \in \mathbb{R}^d$. Let $p_i = \Pr(Z_i = 1)$. Assume that there exists $p < \frac{1}{2}$ such that $p_i \leq p$ for all i . Let $S_k = \sum_{i=1}^k Z_i$ be the random count of individual models that the attack fools on a single input. Since k is odd, the probability $\Pr(S_k \geq k/2)$ is exactly the chance that at least half of the k models are wrong, which means the majority-vote ensemble is also wrong on that adversarially-perturbed input. Assume that pairwise correlation between models is bounded, $|\text{corr}(Z_i, Z_j)| \leq \rho$ for all $i \neq j$, where $0 \leq \rho < 1$. Then*

$$\Pr(S_k \geq k/2) \leq \frac{p(1-p)[1 + (k-1)\rho]}{p(1-p)[1 + (k-1)\rho] + k(\frac{1}{2}-p)^2}.$$

Proof. The variance of Z_i is at most $p(1-p)$. The variance of S_k is at most $kp(1-p) + k(k-1)\rho \cdot p(1-p) = p(1-p)[k + k(k-1)\rho]$. The expectation of S_k is at most kp . Then by Cantelli's Inequality,

$$\begin{aligned} P(S_k \geq k/2) &= P(S_k - E[S_k] \geq k/2 - E[S_k]) \\ &\leq P(S_k - E[S_k] \geq k(0.5 - p)) \\ &\leq \frac{\sigma^2}{\sigma^2 + k^2(\frac{1}{2}-p)^2} \\ &\leq \frac{p(1-p)[k + k(k-1)\rho]}{p(1-p)[k + k(k-1)\rho] + k^2(\frac{1}{2}-p)^2} \\ &= \frac{p(1-p)[1 + (k-1)\rho]}{p(1-p)[1 + (k-1)\rho] + k(\frac{1}{2}-p)^2} \\ &= \frac{p(1-p)[1 + (k-1)\rho]}{p(1-p)[1 + (k-1)\rho + k] + \frac{k}{4}} \end{aligned}$$

□

As $k \rightarrow \infty$, we can see that the bound approaches $\frac{1}{1 + \frac{(\frac{1}{2}-p)^2}{p(1-p) \cdot \rho}}$, which increases with both ρ and p . Thus, as long as the subset of models sampled from the Rashomon set are diverse and therefore decorrelated, an ensemble created from these models is robust to perturbations.

B.6 Proof for Theorem 6

Theorem 6 (KL divergence bound for random ensembles from the Rashomon set). *Let $S = \{(x_i, y_i)\}_{i=1}^n$ be a dataset of n i.i.d. samples from distribution \mathcal{D} over $\mathcal{X} \times \mathcal{Y}$, where $\mathcal{Y} = \{0, 1\}$, and define $p(x) := P(y = 1|x)$ based on S . Let $\mathcal{R}(\epsilon)$ be the Rashomon set trained on S , containing $N = |\mathcal{R}(\epsilon)|$ models $\{f_1, \dots, f_N\}$. Assume each model outputs a probability $f_i(x) \in [\delta, 1-\delta]$ for some constant $\delta \in (0, 1/2]$ for the given input x . Let $\mu(x) = \frac{1}{N} \sum_{i=1}^N f_i(x)$ be the mean prediction and $\sigma^2(x) = \frac{1}{N} \sum_{i=1}^N (f_i(x) - \mu(x))^2$ be the variance of predictions in the Rashomon set for input x . Suppose we sample m models without replacement from $\mathcal{R}(\epsilon)$, where $m \leq N$, with sample indices $\Pi = (\pi_1, \pi_2, \dots, \pi_m)$ chosen uniformly without replacement from $\{1, 2, \dots, N\}$. Define the ensemble prediction as $q_\Pi(x) := \frac{1}{m} \sum_{i=1}^m f_{\pi_i}(x)$. Then, the expected KL divergence*

between $p(x)$ and the ensemble prediction $q_\Pi(x)$ is bounded by:

$$E_\Pi[KL(p(x)||q_\Pi(x))] \leq \frac{(p(x) - \mu(x))^2 + \frac{(N-m)\sigma^2(x)}{(N-1)m}}{\delta(1-\delta)}.$$

Proof. The KL divergence between two Bernoulli distributions with parameters $p = p(x)$ and $q = q_\Pi(x)$ is given by $KL(p||q) = p \log \frac{p}{q} + (1-p) \log \frac{1-p}{1-q}$. Using the Chi-squared divergence upper bound on KL divergence ($KL(p||q) \leq \chi^2(p||q)$), we get:

$$KL(p||q) \leq \frac{(p-q)^2}{q(1-q)}, \quad \text{provided } q \in (0,1).$$

By assumption, $f_i(x) \in [\delta, 1-\delta]$ for all $i = 1, \dots, N$ and some $\delta \in (0, 1/2]$. This implies the ensemble prediction $q_\Pi(x) = \frac{1}{m} \sum_{i=1}^m f_{\pi_i}(x)$ also lies in $[\delta, 1-\delta]$. Therefore, the denominator $q_\Pi(x)(1-q_\Pi(x))$ is bounded below: $q_\Pi(x)(1-q_\Pi(x)) \geq \delta(1-\delta) > 0$.

Applying this bound to the KL inequality we get that:

$$KL(p(x)||q_\Pi(x)) \leq \frac{(p(x) - q_\Pi(x))^2}{q_\Pi(x)(1-q_\Pi(x))} \leq \frac{(p(x) - q_\Pi(x))^2}{\delta(1-\delta)}.$$

Now, we take the expectation over the random sampling Π of m models:

$$\mathbb{E}_\Pi[KL(p(x)||q_\Pi(x))] \leq \mathbb{E}_\Pi \left[\frac{(p(x) - q_\Pi(x))^2}{\delta(1-\delta)} \right] = \frac{1}{\delta(1-\delta)} \mathbb{E}_\Pi[(p(x) - q_\Pi(x))^2].$$

Let $\mu(x) = E_\Pi[q_\Pi(x)]$ be the mean prediction over the entire Rashomon set. We decompose the expected squared error term:

$$\begin{aligned} \mathbb{E}_\Pi[(p(x) - q_\Pi(x))^2] &= \mathbb{E}_\Pi[(p(x) - \mu(x) + \mu(x) - q_\Pi(x))^2] \\ &= \mathbb{E}_\Pi[(p(x) - \mu(x))^2 + (\mu(x) - q_\Pi(x))^2 + 2(p(x) - \mu(x))(\mu(x) - q_\Pi(x))] \\ &= (p(x) - \mu(x))^2 + \mathbb{E}_\Pi[(\mu(x) - q_\Pi(x))^2] \\ &= (p(x) - \mu(x))^2 + Var_\Pi(q_\Pi(x)), \end{aligned}$$

where because of the linearity of expectation we used that $\mathbb{E}_\Pi[\mu(x) - q_\Pi(x)] = \mu(x) - \mathbb{E}_\Pi[q_\Pi(x)] = \mu(x) - \mu(x) = 0$ and $\mathbb{E}_\Pi[(\mu(x) - q_\Pi(x))^2]$ is the variance of the sample mean $q_\Pi(x)$, denoted as $Var_\Pi(q_\Pi(x))$. For sampling m items without replacement from a finite population of size N with variance $\sigma^2(x) = \frac{1}{N} \sum_{i=1}^N (f_i(x) - \mu(x))^2$, the variance of the sample mean is:

$$Var_\Pi(q_\Pi(x)) = \frac{\sigma^2(x)}{m} \left(\frac{N-m}{N-1} \right) = \frac{(N-m)\sigma^2(x)}{(N-1)m}.$$

Substituting this variance back into the expression for the expected squared error:

$$\mathbb{E}_\Pi[(p(x) - q_\Pi(x))^2] = (p(x) - \mu(x))^2 + \frac{(N-m)\sigma^2(x)}{(N-1)m},$$

which gives us the inequality for the expected KL divergence:

$$\mathbb{E}_\Pi[KL(p(x)||q_\Pi(x))] \leq \frac{(p(x) - \mu(x))^2 + \frac{(N-m)\sigma^2(x)}{(N-1)m}}{\delta(1-\delta)}.$$

□

From the theorem above we know that as m increases towards N , the variance term in the numerator decreases (since $N - m \geq 0$), thus reducing the upper bound on the expected KL divergence between the empirical probability $p(x)$ and the ensemble prediction $q_{\Pi}(x)$. Note that this bound is not limited to the Rashomon set of decision trees. It applies to the Rashomon set of other hypothesis spaces.

C Experimental Setup and Results

C.1 Computation Resources

We performed experiments on a 2.7Ghz (768GB RAM 48 cores) Intel Xeon Gold 6226 processor. Each model is trained individually on one core per dataset. We requested 32GB memory for each parallel run.

C.2 Dataset

We present results for 6 datasets: four are from the UCI Machine Learning Repository [80] (Adult, Bank, Credit, and Diabetes), a recidivism dataset (COMPAS) [81], and the Fair Isaac (FICO) credit risk dataset [82] used for the Explainable ML Challenge. We predict which individuals are arrested within two years of release on the COMPAS dataset, and whether an individual will default on a loan for the FICO dataset. The detailed experimental setups are provided in Appendix C.3.

C.3 More Experimental Results

In this appendix, we present additional experimental results for the robustness and privacy analysis.

Table 2: Summary of parameters for adversarial robustness experiment and number of trees averaged over five fold.

Dataset	Adult	Bank	COMPAS	Credit	Diabetes	FICO
Rashomon adder ϵ	0.01	0.02	0.01	0.01	0.04	0.01
Average # Trees	595508.0	1525531.2	846640.0	127601.4	99961.6	300473.2

Diversity in the Rashomon Set Benefits Adversarial Robustness

Collection and Setup: We ran this experiment on 6 datasets. Since TreeFARMS takes binary input, we binarized all datasets using the threshold guessing technique proposed in [83] with `n_estimator` = 30, `max_depth` = 2. To run TreeFARMS, we set regularization to 0.01 and `depth_budget` to 5. The value of ϵ is tuned to ensure that the constructed Rashomon set contains a sufficient number of trees, usually more than 100,000. The exact ϵ values and the corresponding number of trees for each dataset are provided in Table 2. We adopt the algorithm proposed in [78] to attack the optimal tree, the tree with the lowest objective value. The attack uses the S_{∞} set with $\eta = 0.1$ as defined in section 5.1. We report the performance of other trees in the Rashomon set on the adversarial dataset. For visualization purposes, we group the trees based on their prediction patterns on the validation set and measure their Hamming distance to the prediction pattern of the optimal tree.

Results: Figure 6, a more comprehensive version of Figure 3, shows that trees with classification patterns similar to the optimal tree (bottom left corner) are more vulnerable to adversarial examples, while those that differ more in their predictions (top right corner) are more robust. The reported Spearman correlation coefficients are all positive, and for 4 out of 6 datasets, the coefficients are above 0.7, indicating a strong

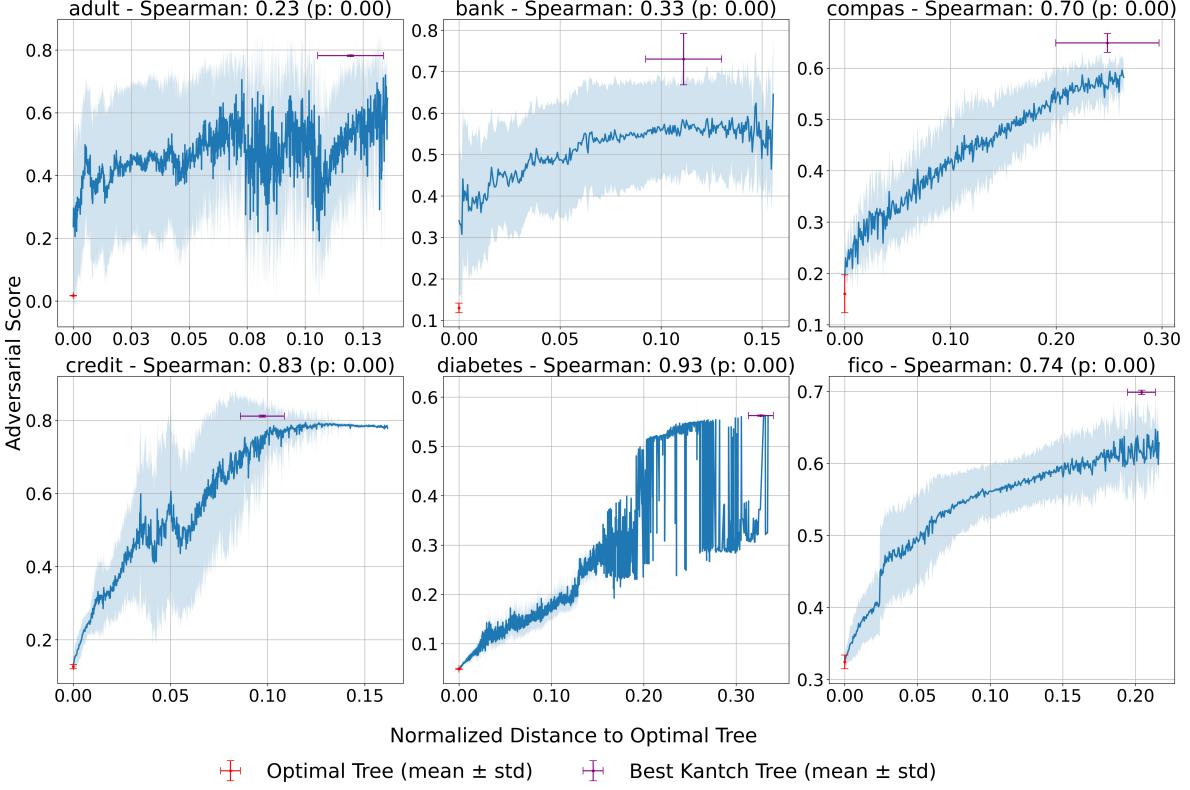


Figure 6: Adversarial score of trees in Rashomon set vs. their distance to optimal tree. Results are aggregated over five folds. The optimal trees (in red) are attacked. The most robust trees (in purple) are far from optimal tree. Trees with the same distance to optimal trees are grouped, and mean and standard deviation of their adversarial score are shown as line plots with shaded uncertainty.

positive correlation between diversity and adversarial robustness. The relatively lower Spearman correlation for the Adult and Bank datasets may be due to the limited diversity within their Rashomon sets, which might be caused by the data distributions.

Figure 7 shows a scatterplot of each tree’s distance to the optimal tree versus its adversarial score. As we can see, for the COMPAS, Credit, Diabetes, and FICO datasets, we observe a strong positive trend. While for the Adult and Bank datasets, some trees with zero distance from the attacked tree still perform well on the adversarial examples, an overall increasing trend remains visible in the scatter plots, supporting our main conclusion.

Diversity in the Rashomon Set Accelerates Information Leakage

Collection and Setup: We ran this experiment on 6 datasets. Following the setup in [79], we binarized each dataset and subsampled 100 data points to form the training set. We ran TreeFARMS on these 100-sample datasets using $\epsilon = 0.02$ and $\text{depth_budget} = 5$. We tuned the regularization parameter instead of ϵ to control the size of the Rashomon set, since the set size is less sensitive to changes in the regularizer. Controlling the size is necessary because the DRAFT algorithm can only feasibly compute ensembles with a few hundred estimators [79].

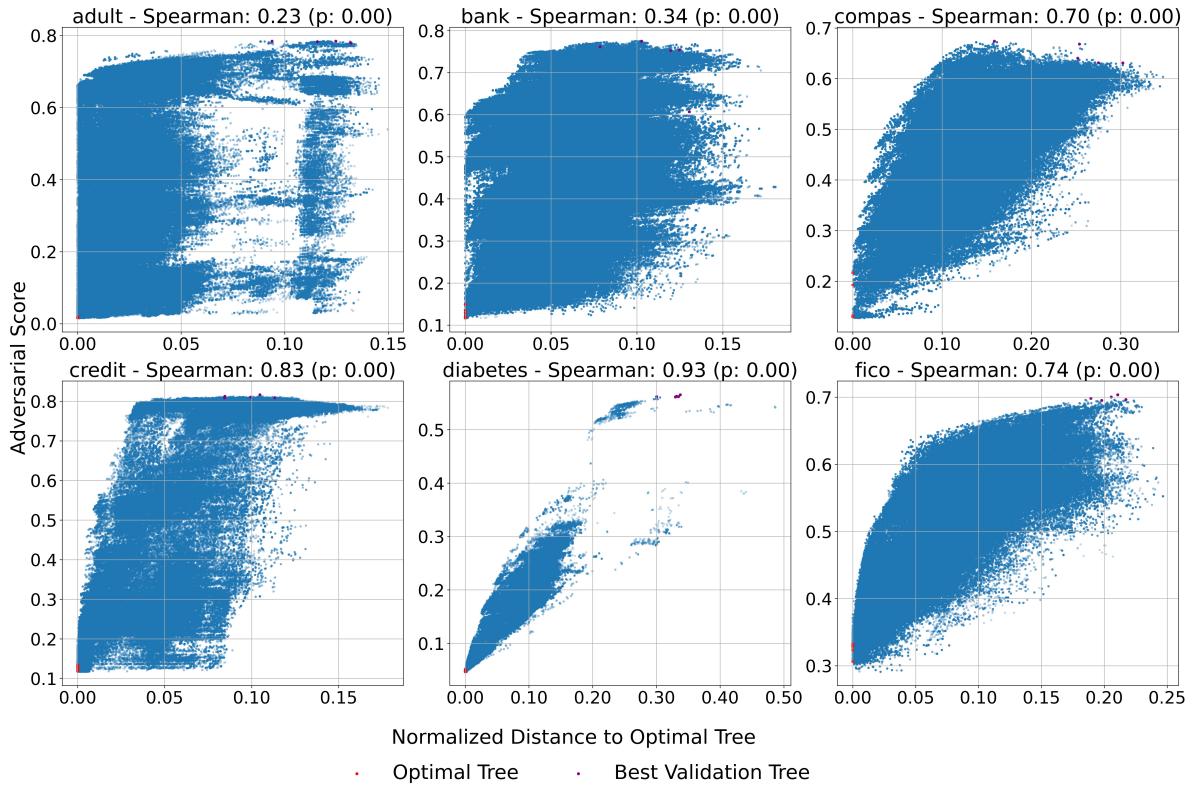


Figure 7: Adversarial score of trees in Rashomon set vs. their distance to optimal tree. Results are aggregated over five folds. The optimal trees (in red) are attacked. The most robust trees (in purple) are far from optimal tree.

The exact values of the regularizer and the average number of trees are reported in Table 3. Once the Rashomon set is constructed, trees are sequentially selected from the Rashomon set and passed to DRAFT. We run DRAFT multiple times as more trees are added. Specifically, DRAFT is trained after each additional tree from 1 to 50, every 5 trees from 50 to 150, and again at 175 and 200 trees. In total, we consider up to 200 trees from the Rashomon set. We run this process five times with different random seeds for sampling data points. We consider two strategies to select trees. By default, the first tree selected is the optimal model. The *closest* strategy then iteratively selects the tree whose classification pattern has the smallest Hamming distance to that of the optimal tree. The *farthest* strategy greedily selects the tree whose classification pattern has the largest Hamming distance from those of the previously selected trees.

Table 3: Summary of parameters for the data reconstruction experiment and the average number of trees across five runs.

Dataset	adult	bank	compas	credit	diabetes	fico
Regularization λ	0.01	0.013	0.01	0.0165	0.0125	0.02
Average # Trees	29428.6	12267.4	27432.8	6806.8	6418.4	7327.0

Results: Figure 8 shows the comparison of reconstruction error between different selection strategies. For all datasets, the “farthest” strategy (in purple) has lower reconstruction error, indicating greater information leakage. The “closest” strategy (in green) has better privacy, as it has higher reconstruction error. These results suggest that more diversity in the Rashomon set and disclosing more diverse models lead to increased privacy risk. The random baseline randomly guesses the feature values for each data point. It’s a very conservative attack and usually results in high reconstruction error.

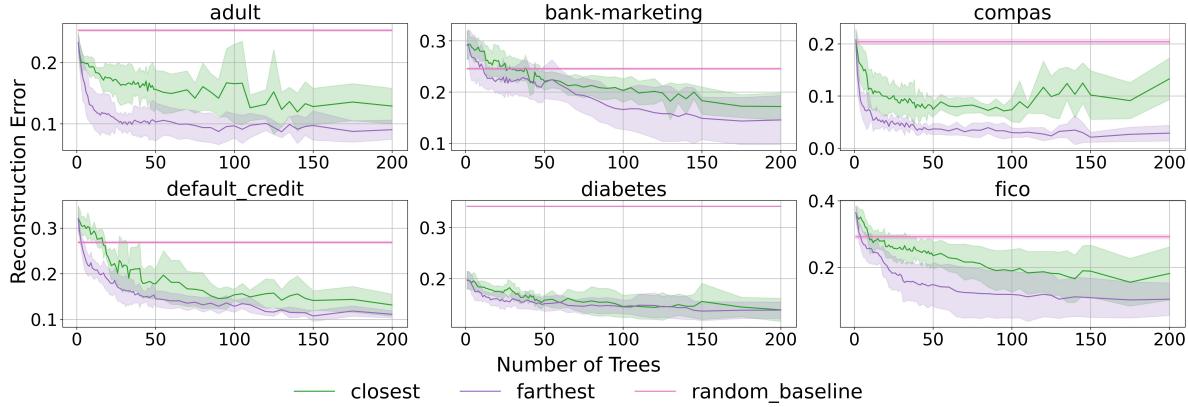


Figure 8: Comparison of reconstruction error between different selection strategies. The random baseline randomly guesses the feature values for each data point.

Robustness-Privacy Tradeoffs under the Rashomon Set

Collection and Setup: We evaluate the robustness-privacy tradeoff directly in this experiment. First, we construct multiple groups of trees of varying sizes from a Rashomon set to represent different levels of diversity. Then, we assess each group’s performance under both reconstruction and robustness attacks. To form these groups, we sort trees in the Rashomon set by Hamming distance of their classification patterns

to the optimal tree, then select trees at evenly spaced intervals. For instance, given a Rashomon set of 1000 trees, an ensemble of 3 trees would include those ranked 1, 500, and 1000, while an ensemble of 100 trees would include every 10th tree in the sorted list. As in previous experiments, we use DRAFT to perform the reconstruction attack and report the reconstruction error for each group. For the robustness evaluation, we apply an adversarial attack targeting the optimal tree in the Rashomon set using the \mathcal{S}_0 set with $\eta = 1$, which allows a single binary feature flip per data point. This setup is used because DRAFT requires binary features. We then evaluate all trees within each group and record the adversarial accuracy of the best-performing tree in the group.

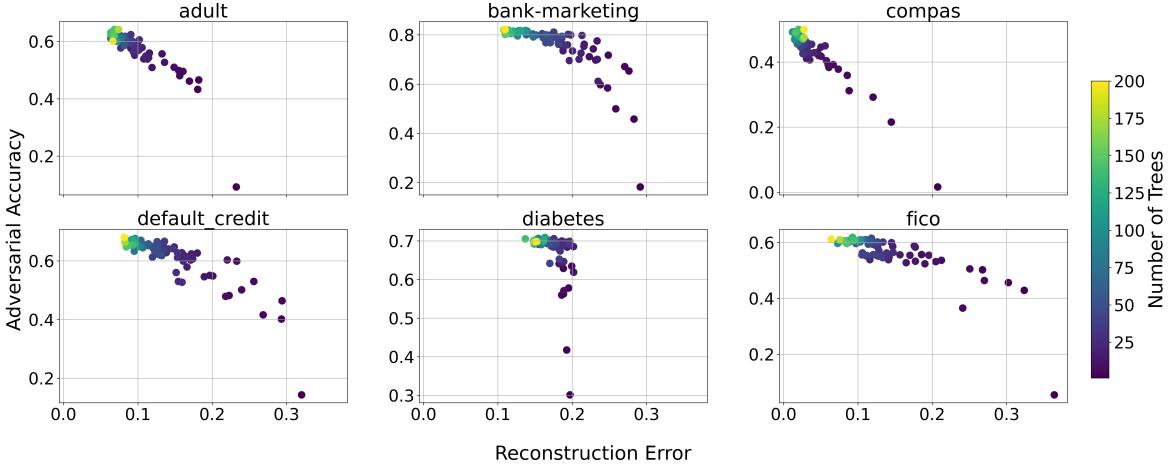


Figure 9: Reconstruction error vs. adversarial accuracy for ensembles constructed with different numbers of evenly sampled trees from the Rashomon set.

Results: Figure 9 shows the reconstruction error versus adversarial accuracy for ensembles constructed by selected trees. Each point represents an ensemble of a specific size, with color indicating the number of trees included. When only a limited number of models are released, the privacy is preserved but the models are not diverse enough to avoid an adversarial attack targeted on the optimal trees. As more trees are selected, robustness improves but at the cost of greater information leakage.

C.4 Study for Theorem 5

We provide in this section empirical verification of the theoretical results of Theorem 5. To do this, we compute the Rashomon sets for 4 binarized datasets using TreeFARMS with regularization equal to 0.01, Rashomon parameter $\epsilon = 0.03$, and $\text{depth_budget} = 5$. For each dataset and 24 values of K uniformly ranging from 0.25% to 6% of the number of points of the dataset, we modify K samples of the dataset. Specifically, the modification begins by using the k -means clustering algorithm to compute 5 clusters of the dataset. We randomly select a cluster with over K samples, and we uniformly and at random remove K samples from that cluster. We then randomly select a different cluster, from which we randomly and with replacement select K samples to duplicate. Lastly, we flip the label of each of the duplicate samples with probability 50%. The resulting dataset sees a targeted shift in both the feature distribution as well as the label distribution.

On the modified dataset, we compute the modified Rashomon set with the same value of regularization and depth_budget as before, as well as Rashomon parameter ϵ' . We use two different values of ϵ' , where

$\epsilon' = \epsilon$ reuses the same Rashomon parameter as before, and $\epsilon' = \epsilon + \frac{2K}{n}$ uses the Rashomon parameter stated in Theorem 5. We lastly compute the percentage of models in the original Rashomon set that remain in the modified Rashomon set. We repeat this process on 5 modified datasets in total and average the results.

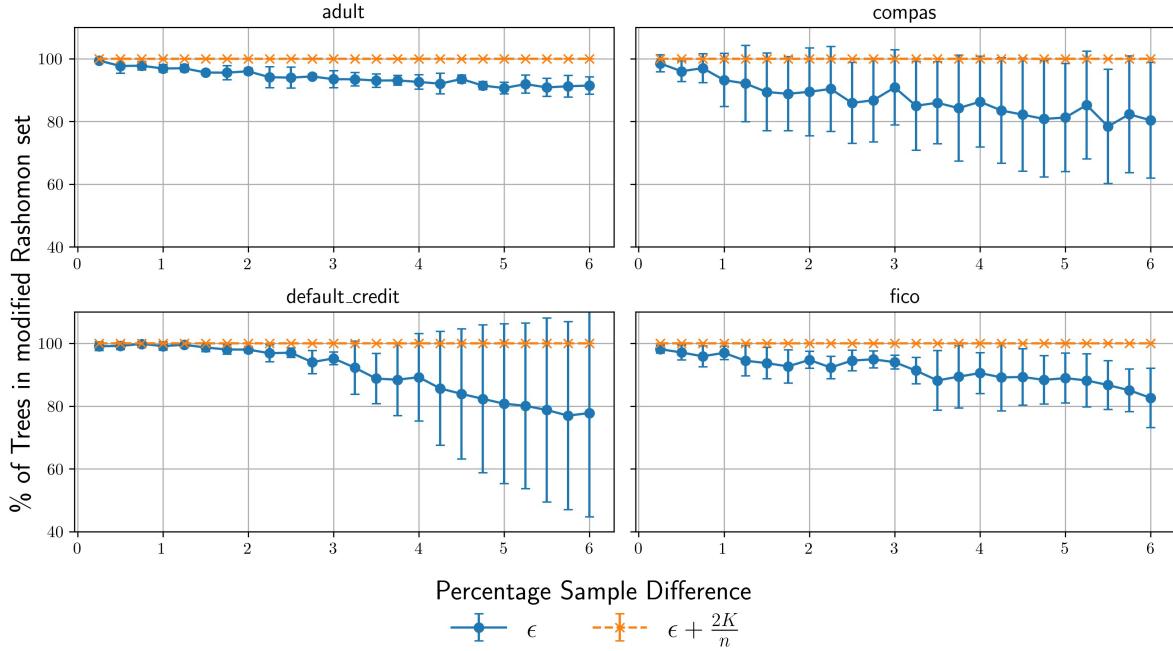


Figure 10: Percentage of trees in the Rashomon set that remain in the modified Rashomon set after modifying K samples.

As we observe in Figure 10, we can see that every model in the original Rashomon set remains in the modified Rashomon set when $\epsilon' = \epsilon = \frac{2K}{n}$. This agrees with Theorem 5, which predicts that the original Rashomon set should be contained within the modified Rashomon set. When we fix $\epsilon' = \epsilon$, we instead see that some models exit the Rashomon set after we modify the dataset, with a greater change in the dataset resulting in a smaller overlap between the two Rashomon sets.

C.5 Synthetic Study for Theorem 6

In Section 5.3, to support analytical conclusions of Theorem 6, we provided simulation experiment where we measure the KL divergence between the synthetic distribution and the predicted distribution. The details of the simulation setup are described in this section.

We use five different data distributions for this experiment. Let $\mathbf{x} = (x_0, x_1, \dots, x_{d-1}) \in \{0, 1\}^d$ be a binary feature vector of dimension d . We set $d = 4$ in this experiment. Let $P(Y = 1 | \mathbf{x})$ denote the true conditional probability of label $Y = 1$ given \mathbf{x} . Also, let the sum of features be $s = \sum_{i=0}^{d-1} x_i$. The five distributions are defined as follows:

- Distribution 1 (Parity): $P(Y = 1 | \mathbf{x}) = 0.1 + 0.8 \cdot \mathbb{1}[s = 0]$.
- Distribution 2: $P(Y = 1 | \mathbf{x}) = 0.15 + 0.7 \cdot \mathbb{1}[x_0 = 1 \wedge x_1 = 1]$.
- Distribution 3 (XOR): $P(Y = 1 | \mathbf{x}) = 0.2 + 0.6 \cdot \mathbb{1}[(x_0 \oplus x_1 = 1) \wedge (x_2 \oplus x_3 = 0)]$.

- Distribution 4 (Random): $P(Y = 1 | \mathbf{x}) = 0.3 + 0.4 \cdot r, r \sim \mathcal{U}(0, 1)$.

- Distribution 5: $P(Y = 1 | \mathbf{x}) = \begin{cases} 0.05 & \text{if } s \leq 1 \\ 0.95 & \text{if } s \geq 3 \\ 0.5 + 0.4(x_0 - 0.5) & \text{otherwise.} \end{cases}$

We set $d = 4$, the distribution $P(\mathbf{x})$ is uniform. For each of these five true distributions $P(Y = 1 | \mathbf{x})$, we sample a dataset of 100 points 5 times. These train sets are then used to train the Rashomon set. The TreeFARMS configuration includes a regularization parameter of 0.001 and a Rashomon bound multiplier of 0.03.

For each training data, we consider ensembles of models from the Rashomon set. We start from one tree and increase counter J that corresponds to the ensemble size until we reach the ensemble that consists of all trees in the Rashomon set. We estimate the expected KL divergence between the true distribution and an ensemble's average prediction. Our procedure is as follows for each J : (1) We sample an ensemble of size J 20 times without replacement from the models in the Rashomon set. (2) For each ensemble and for every data point, we estimate the ensemble's average predicted probability of $P(Y = 1 | \mathbf{x})$ by averaging predictions of models in the ensemble. (3) For each \mathbf{x} , we compute the pointwise KL divergence between the true conditional distribution $P_{\text{true}}(Y = 1 | \mathbf{x})$ and the ensemble's predicted conditional distribution $P(Y = 1 | \mathbf{x})$. (4) We compute the expected KL divergence by taking the empirical mean of pointwise KL divergences over 20 samples of the ensembles. (5) Finally, we report the expected KL divergence averaged over 5 dataset sampled from the same true distribution.

As we observe in Figure 2, the KL-divergences decreases as the number of trees in ensemble increases, verifying the results proved in Theorem 6.

D Additional Studies

D.1 Analysis of Ensemble Construction Strategies

In Section 6 and Appendix C.3, we have introduced several strategies for constructing ensembles using trees from the Rashomon set. Here, we briefly summarize each strategy along with its motivation. We also introduce a few additional strategies that enable further interesting analyses.

Random versus Evenly-spaced Sampling: As mentioned in section 6.3, we introduce *increment* sampling strategy, which selects trees from the Rashomon set at evenly spaced intervals based on their predicting patterns, depending on the size of the ensemble we want to construct. The goal is to evenly represent the Rashomon set in order to observe its default properties. We compare this strategy to the *random* sampling method, where trees are randomly selected to form the ensemble. For both methods, we include the optimal tree as the first selected tree by default. Figure 11 shows this comparison. We can observe that both strategies produce closely aligned error values and follow similar trends. Therefore, using the increment sampling can help us capture the variation in the Rashomon set.

Closest, Farthest, and Evenly-spaced Sampling: In section 6.2, we introduced the *closest* and *farthest* sampling strategies. The closest strategy selects the next tree that is most similar to the optimal tree in terms of prediction patterns. In contrast, the *farthest* strategy aims to maximize diversity by greedily selecting trees that have the highest average prediction distance from those already selected. We compare these strategies with the *increment* strategy.

Figure 12 shows that the increment strategy leaks more information than the other two strategies, indicating that evenly spaced sampling may pose a higher privacy risk. This is not particularly surprising, as

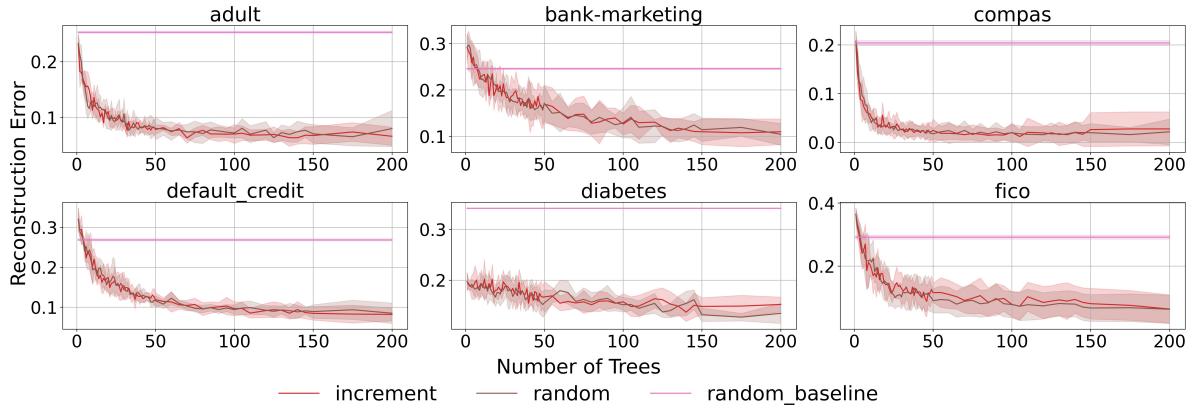


Figure 11: Comparison of reconstruction error between *increment* and *random* strategy. The random baseline guesses the feature values for each data point.

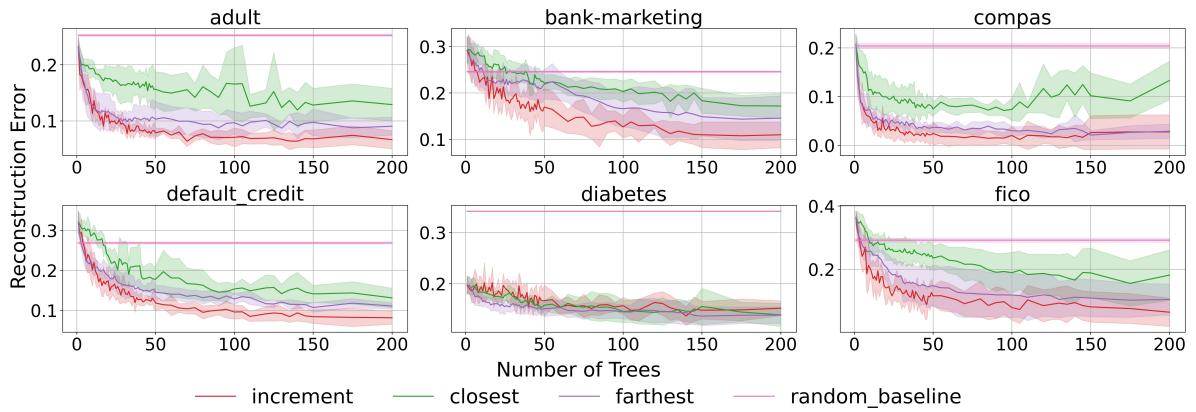


Figure 12: Comparison of reconstruction error between *increment*, *closest*, and *farthest* strategy. The random baseline guesses the feature values for each data point.

selecting only similar or highly dissimilar trees tends to concentrate on specific regions of the Rashomon set, whereas the increment strategy more effectively captures the full landscape. This finding also motivates future research into alternative definitions of diversity beyond prediction patterns.

Sampling Strategies Based on Tree Sparsity: In addition to diversity-based and evenly-spaced sampling strategies, we also explore other approaches based on sparsity. In decision trees, sparsity is usually measured by the number of leaves. We study two sparsity-based strategies: the *sparsest* strategy selects trees with the fewest leaves in ascending order, while the *densest* strategy selects trees with the most leaves in descending order.

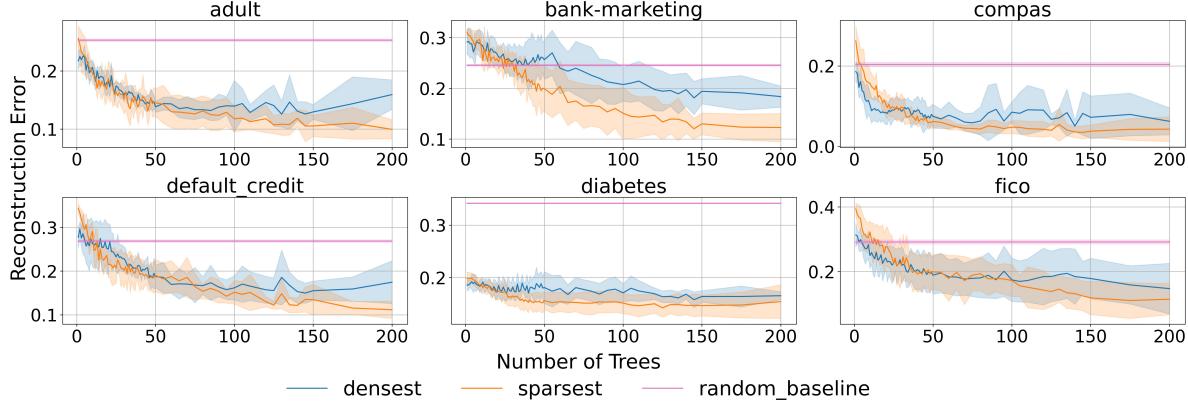


Figure 13: Comparison of the reconstruction error between *sparsest* and *densest* strategy. The random baseline guesses the feature values for each data point.

Theorem 1 proves that individual sparse models leak less information. This is reflected in Figure 13, where the sparsest tree (in orange) has higher reconstruction error than the densest tree (in blue) when the x-axis is one. However, as more trees are added, the orange curve drops below the blue curve. This may be because sparse trees, while individually less expressive, offer greater structural diversity and better generalization when aggregated. As a result, ensembles of sparse trees can more effectively cover the input space without overfitting, leading to lower reconstruction error compared to ensembles composed of dense trees, which may be more redundant or over-specialized. This observation suggests that sparsity may have a more complex impact on privacy leakage and motivates further study of how structural factors influence privacy under the Rashomon set setting.

Comparing Robustness-Privacy Trade-offs for Different Sampling Strategies In Section 6 Figure 5, we empirically observed the robustness-privacy trade-offs when the Rashomon set is present. In this figure, we used incremental sampling procedure. Here, we verify that other sampling strategies lead to this trade-off as well. Experimentally, we use the same setup as in Figure 5.

In Figure 14 we plot the robustness-privacy trade-off for six datasets under different sampling strategies. First of all, we notice that the trade-off is preserved for all three strategies. However, the extent of it differs, most likely due to the diversity of the constructed Rashomon sets (for example, we expect farthest strategy to produce less diverse sets as compared to incremental one). This difference in the extent of the trade-off between sampling strategies is especially evident in datasets like COMPAS and Adult.

Besides the trade-off, we also saw another interesting pattern in this figure. For some datasets like bank-marketing and diabetes, we observe that one can construct a Rashomon set that contains models that can achieve low privacy risk and high robustness (as indicated by the presence of points in the top right of the plots). Overall, our empirical findings show that this trade-off is an interesting phenomenon that is influenced

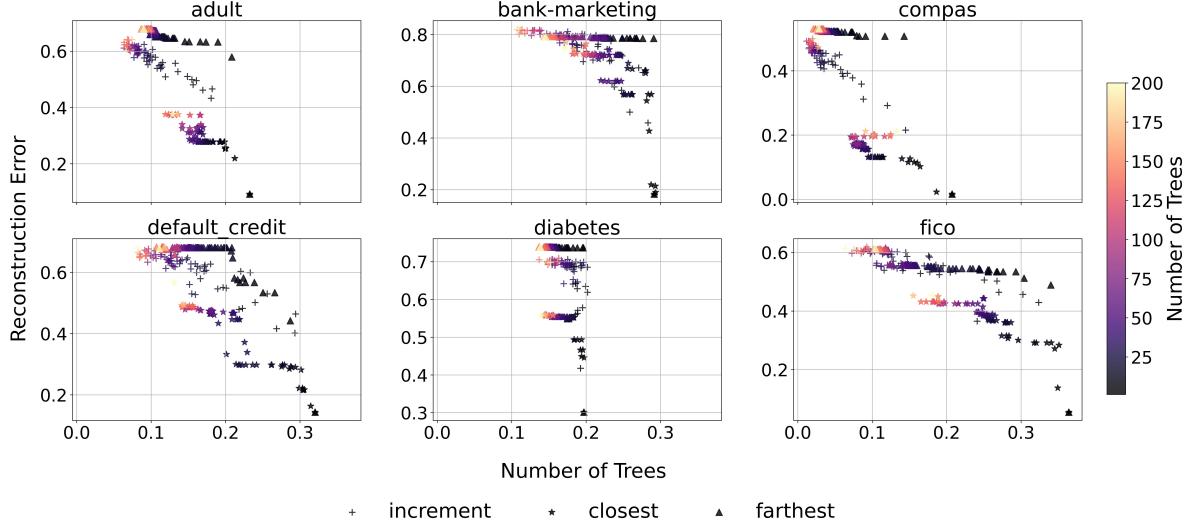


Figure 14: Reconstruction error vs. adversarial accuracy for ensembles constructed with *increment*, *farthest*, and *closest* strategy. Each strategy is plotted with different markers as shown in the legend.

by the diversity of the Rashomon set and can be a rich direction for future research.

D.2 Single Tree is Vulnerable to Adversarial Attack

In Section 5.1 in Theorem 2, we discussed the inherent vulnerability of a single model, using rule lists, which is a one-sided decision tree. Here, we verify empirically this vulnerability for the Rashomon set of sparse decision trees, which include rule lists as well. Our setup is similar to the one in adversarial robustness experiments that we described above. For COMPAS, default-credit, and FICO, we used binarized datasets as described in Section 6.2. We did not perform subsampling for the Rashomon set. We divided the data into five folds for cross-validation. We trained TreeFARMS with regularization of 0.01, depth budget of 4, and the Rashomon adder ϵ set to 0.01. We performed ℓ_1 attack with $\eta = 1$, allowing a single binary flip to each row. We attacked each tree separately.

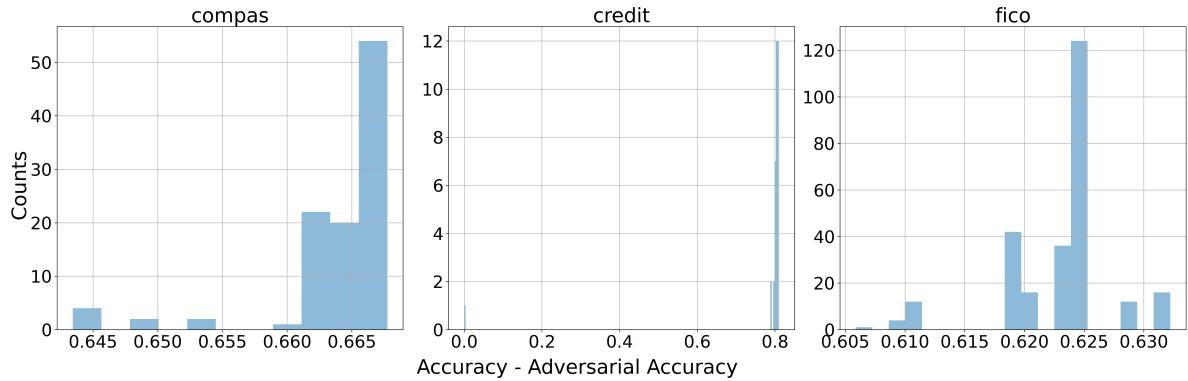


Figure 15: Accuracy gap ($\hat{L}_S(f) - \hat{L}_{S'}(f)$) under ℓ_1 attack on the individual trees of the Rashomon set of sparse decision trees. The results presented in histogram are averaged over five folds.

The results of this experiment are presented in Figure 15. There was only one tree in the credit dataset for which the attack was ineffective. For all other trees across the three datasets, we observed a steep drop in accuracy of at least 60% between the original empirical risk, $\hat{L}_S(tree)$, and the adversarial risk, $\hat{L}_{S'}(tree)$, where $tree$ is a model from the Rashomon set. Therefore, within the hypothesis space of sparse decision trees, there is an inherent vulnerability when each tree is attacked individually (similar to Theorem 2). As we show in Section 5.1, diverse Rashomon sets allow for model selection that is more resilient to the attack, provided the notion of diversity aligns with the type of attack.