Supplemental materials
A Path to Simpler Models Starts With Noise

A Proof for Theorem 2

We state and prove Theorem 2 below.

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

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

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

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

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

$$\mathbb{E}_{S_\rho}\,\hat{L}_{S_\rho}(f) = (1-2\rho)\hat{L}_S(f) + \rho. \tag{3}$$

Recall that we take expectation with respect to different ways of adding noise to labels, therefore S_ρ and S have the same x, but different y. We do not use (3) for the proof of Theorem 2, but use it in Appendix J.

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

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

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

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

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

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

□

In Theorem 2, the statement of the theorem is correct for any fixed $f \in \mathcal{F}$. Corollary 3 follows directly from Theorem 2. Here, instead of a fixed model $f \in \mathcal{F}$, we consider models in the Rashomon sets that maximize expected variance.

Corollary 3 (Maximum variance increases with label noise)**.** *Under the same assumptions as in Theorem 2, we have that*

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

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

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

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

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

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

□

Next, we generalize the statement of Theorem 2 to the non-uniform label noise case, where each sample $z = (x, y)$ is flipped with probability ρ_x that depends on x. We show that under this non-uniform label noise, the variance of the loss increases in Theorem 12.

Theorem 12 (Variance increases with non-uniform label noise)**.** *Consider 0-1 loss l, infinite true data distribution $\mathcal{D}$, and a hypothesis space $\mathcal{F}$. Assume that there exists at least one function $\bar{f} \in \mathcal{F}$ such that $L_{\mathcal{D}}(\bar{f}) < \frac{1}{2} - \gamma$. For a fixed $f \in \mathcal{F}$, let $\sigma^2(f, \mathcal{D})$ be the variance of the loss: $\sigma^2(f, \mathcal{D}) = Var_{z \sim \mathcal{D}} l(f,z)$ on data distribution $\mathcal{D}$. Consider non-uniform label noise, where each label y is flipped independently with probability ρ_x, $(x, y) \sim \mathcal{D}$. Let $\mathcal{D}_\rho$ denote the noisy version of $\mathcal{D}$. For any $\delta > 0$, let $\mathcal{D}_{\rho^\delta}$ be a noisier data distribution than $\mathcal{D}_\rho$, meaning that for every sample*

(x, y) *the probabilities of labels being flipped are higher by* δ*:* $\rho_x^\delta = \rho_x + \delta$*. If for a fixed model* $f \in \mathcal{F}$*,* $L_{\mathcal{D}_{\rho^\delta}}(f) < 0.5$*, then*

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

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

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

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

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

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

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

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

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

□

We also show that the variance of losses increases under margin noise for data that come from Gaussian distributions (in Theorem 15). We model margin noise by moving two Gaussians closer together along the vector that connects the two means. Before stating and proving the theorem we discuss two lemmas that are helpful for the proof.

Lemma 13. *Consider distribution* $\mathcal{X} \in \mathbb{R}^m$ *and a linear model* $f = \omega^T x + b$*, where* $\omega \in \mathbb{R}^m$*,* $\omega \neq \bar{0}$ *and* $b \in \mathbb{R}$*. Let* $x \mapsto Ax + c$ *be a bijective affine transformation, where* $A \in \mathbb{R}^{m\times m}$ *and* $c \in \mathbb{R}^m$*. For the linear model* $g(x) = f(A^{-1}(x - c))$ *and the distribution* $\mathcal{Z} = A\mathcal{X} + c$*, we have that:*

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

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

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

□

Lemma 14. *Consider a Gaussian distribution $\mathcal{X} \sim \mathcal{N}(\mu, I)$, where $\mu \in \mathbb{R}^m$, and a linear model $f = \omega^T x + b$, where $\omega \in \mathbb{R}^m$, $\omega \neq \bar{0}$ and $b \in \mathbb{R}$. Let $r = \frac{\omega^T \mu + b}{\|\omega\|}$ be the signed distance from μ to the decision boundary of f. Then,*

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

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

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

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

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

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

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

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

From Lemma 13 and definitions of O, g, $\mathcal{Z}$, we get that $P_{x \sim \mathcal{X}}(f(x) > 0) = P_{z \sim \mathcal{Z}}(g(z) > 0)$. Therefore:

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

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

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

Theorem 15. *Consider data distribution $\mathcal{D} = \mathcal{X} \times \mathcal{Y}$, where, $\mathcal{X} \in \mathbb{R}^m$, $\mathcal{Y} \in \{-1, 1\}$, classes are balanced $P(Y = -1) = P(Y = 1)$ and generated by Gaussian distributions $P(X|Y = -1) = \mathcal{N}(\bar{0}, \Sigma)$, $P(X|Y = 1) = \mathcal{N}(\mu, \Sigma)$, where Σ is a diagonal matrix with non-zero elements. Let the hypothesis space $\mathcal{F}$ be the set of linear models, $f = \omega^T x + b$, where $\omega \in \mathbb{R}^m$. $\omega \neq 0$ and $b \in \mathbb{R}$. We add margin noise by moving the means of the Gaussians towards each other by a factor of k, where $0 < k < 1$, meaning that the mean of the positive class becomes $\mu_k = k \cdot \mu$. For a fixed $f \in \mathcal{F}$, if $L_\mu(f) < 0.5$, we get that the variance of losses increases with more noise,*

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

Proof. Without loss of generality, we will show that the variance of the losses increases for data generated from two Gaussian distributions $P(X|Y = -1) = \mathcal{N}(\bar{0}, I)$ and $P(X|Y = 1) = \mathcal{N}(\mu, I)$ (where I is the identity matrix) when we move them towards each other. More specifically, since normalization by variance $\left(\frac{1}{\Sigma_{i,i}}\right)$ is a bijective linear transformation, by Lemma 13 we can work

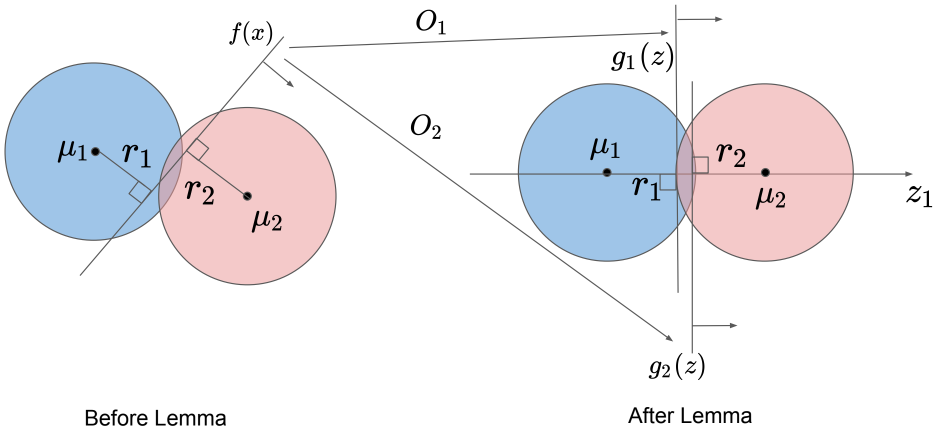


Figure 4: An illustration of how Lemma 14 rotates each of the Gaussians $\mathcal{N}(\mu_1, I), \mathcal{N}(\mu_2, I)$ and the decision boundary $f(x)$ in order to compute loss as CDF of the signed distances (r_1, r_2) from means (μ_1, μ_2) to the rotated boundaries $(g_1(z), g_2(z))$. Note that we apply Lemma 14 separately to each Gaussian, thus there are two rotation operators O_1, and O_2.

with $P(X|Y=-1) = \mathcal{N}(\bar{0}, I)$ and $P(X|Y=1) = \mathcal{N}(\mu, I)$ instead of $P(X|Y=-1) = \mathcal{N}(\bar{0}, \Sigma)$ and $P(X|Y=1) = \mathcal{N}(\mu, \Sigma)$.

Let $r_1 = \frac{b}{\|\omega\|}$ and $r_2 = \frac{\omega^T\mu + b}{\|\omega\|}$ be the signed distances from the centers of the two Gaussians to the decision boundary. Then, from Lemma 14 (see illustration in Figure 4), the loss can be computed using the CDFs based on the signed distance:

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

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

Recall that we induce noise by moving the Gaussians towards each other by decreasing k. Now we will show that loss is monotonically decreasing with respect to increasing values of k, or equivalently that $\frac{\partial}{\partial k} L_{\mu_k}(f) < 0$:

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

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

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

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

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

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

Note, that we can generalize Theorem 15 to the case when Σ is any positive-definite matrix that is not necessarily diagonal (covariance matrices are always positive semi-definite, and we now additionally assume that Σ does not have zero eigenvalues). Since Σ is real and symmetric, by the spectral theorem, there exists an orthogonal matrix $Q \in \mathbb{R}^{m\times m}$ such that $D = Q\Sigma Q^T$ where D is diagonal and contains eigenvalues of Σ. The diagonal elements of D must be real and positive since Σ is positive-definite. Then, consider the data distribution $(Q\mathcal{X}) \times \mathcal{Y}$. From the properties of the Gaussian distribution, $Q\mathcal{X}$ is Gaussian with mean $Q\mu$ and covariance matrix $Q\mathcal{X}Q^T = D$. Thus, we can generalize the results of Theorem 15 to apply to positive-definite non-diagonal matrices Σ.

For a fixed model, we additionally verify the results of Theorem 15 empirically, by generating Gaussian distributions and introducing margin noise by moving the Gaussians closer together (see Figure 5(b).) The variance of losses increases with additive and uniform random attribute noise as well, as we show empirically in Figure 5(c)-(d).

While the results of Theorems 12, 15 are for a given and fixed model f, they hold for the f that achieves the maximum variance in the Rashomon set as well, meaning that Corollary 3 extends to Theorems 12, 15.

B Bernstein's and Hoeffding's inequalities

In this section, we compare Bernstein's and Hoeffding's inequalities and show that, under certain assumptions on variance, Bernstein's inequality is tighter than Hoeffding's inequality. We provide Bernstein's inequality in Lemma 16 and Hoeffding's inequality in Lemma 17.

Lemma 16 (Bernstein's inequality for loss class)**.** *Consider a hypothesis space $\mathcal{F}$. For a fixed $f \in \mathcal{F}$, let loss l be bounded by $C > 0$ such that $|l(f,z)| \le C$ for every $z \in \mathcal{Z}$. For any $\varepsilon > 0$,*

$$P\left(L(f) - \hat{L}(f) > \varepsilon\right) \le e^{\frac{-n\varepsilon^2}{2\sigma_f^2 + 2C\varepsilon/3}}, \tag{4}$$

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

Lemma 17 (Hoeffding's inequality for loss class)**.** *Consider a hypothesis space $\mathcal{F}$. For a fixed $f \in \mathcal{F}$, let loss l be bounded by $a, b \ge 0$ such that $a \le l(f,z) \le b$ for every $z \in \mathcal{Z}$. For any $\varepsilon > 0$,*

$$P\left(L(f) - \hat{L}(f) > \varepsilon\right) \le e^{\frac{-2n\varepsilon^2}{(b-a)^2}}, \tag{5}$$

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

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

Theorem 18 (Bernstein's inequality is stronger than Hoeffding's for lower variance)**.** *For a fixed $f \in \mathcal{F}$, let loss $l \in [a,b]$ so that $a \le l(f,z) \le b$ for every $z \in \mathcal{Z}$. Then, Bernstein's inequality is stronger than Hoeffding's inequality for all $\varepsilon \in (0, b-a)$ if $\sigma_f^2 \le \frac{(b-a)^2}{12}$ or if $\left|L(f) - \frac{a+b}{2}\right| > \frac{b-a}{\sqrt{6}}$ where $\sigma_f^2 = \mathrm{Var}_{z\sim\mathcal{D}}\, l(f,z)$.*

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

Proof. According to Hoeffding's inequality (5), we have that

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

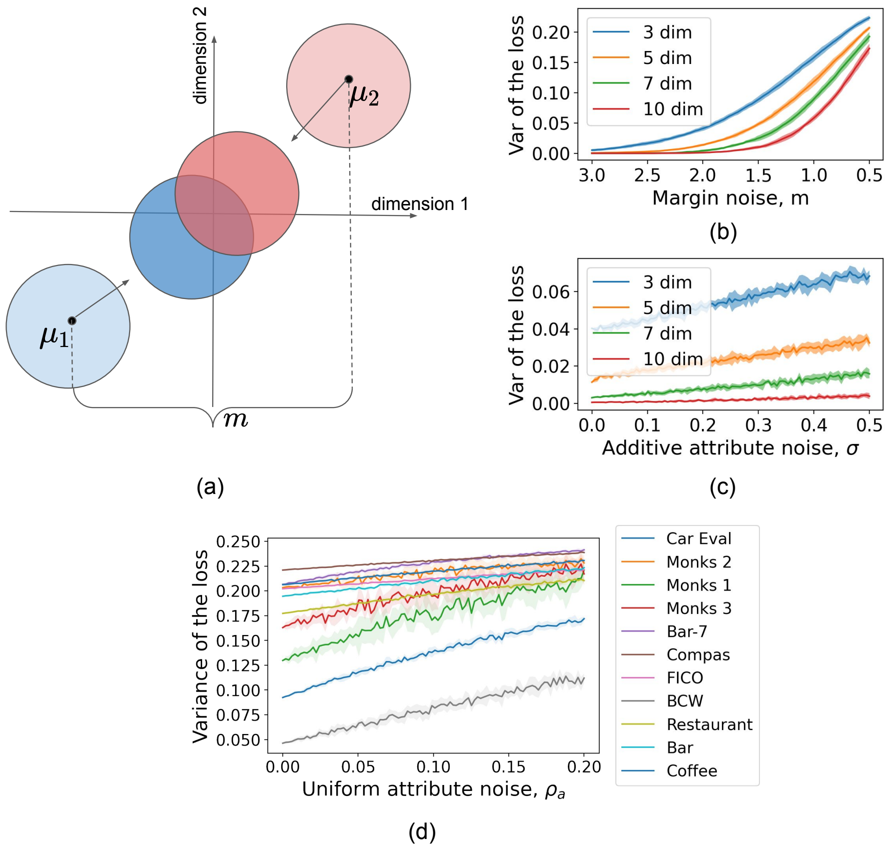


Figure 5: The variance of losses increases with margin (b) and additive attribute (c, d) noise. For (b) and (c) we generated data from Gaussians in 3, 5, 7, and 10 dimensions. For margin noise (b), as illustrated in (a), the negative class is generated from $\mathcal{N}(\bar{\mu_1}, I)$ and positive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

Recall that Bernstein's inequality [4] states

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

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

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

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

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

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

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

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

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

C Proof for Theorem 4

For a discrete hypothesis space, Theorem 4 bounds generalization error with a term that depends on the size of the true Rashomon set and maximum variance of the loss. Recall Theorem 4.

Theorem 4 (Variance-based "generalization bound")**.** *Consider dataset S, 0-1 loss l , and finite hypothesis space $\mathcal{F}$. With probability at least $1-\delta$, we have that for every $f \in R_{set}(\mathcal{F},\gamma)$:*

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

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

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

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

According to the union bound:

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

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

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

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

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

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

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

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

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

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

□

D Proof for Proposition 5

We recall and provide proof for Proposition 5 below.

Proposition 5 (ERM can be close to the true Rashomon set)**.** *Assume that through the cross-validation process, we can assess ξ such that $L(\hat{f}) - \hat{L}(\hat{f}) \leq \xi$ with high probability (at least $1-\epsilon_\xi$) with respect to the random draw of data. Then, for any $\epsilon > 0$, with probability at least $1 - e^{-2n\epsilon^2} - \epsilon_\xi$ with respect to the random draw of training data, when $\xi + \epsilon \leq \gamma$, then $\hat{f} \in R_{set}(\mathcal{F},\gamma)$.*

Proof. For a fixed $f \in \mathcal{F}$ for 0-1 loss by Hoeffding's inequality (5):

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

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

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

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

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

□

E Proof for Proposition 6

For the hypothesis space of decision trees, the number of possible decision trees in the hypothesis space grows exponentially fast with the depth of the tree and the number of features. In Proposition 6 we show that the Rashomon set growth rate is smaller than the growth rate of the hypothesis space, and thus this leads to larger Rashomon ratios for simpler hypothesis spaces. We recall Proposition 6 below.

Proposition 6 (Rashomon ratio is larger for decision trees of smaller depth)**.** *For a dataset $S = X \times Y$ with binary feature matrix $X \in \{0,1\}^{n\times m}$, consider a hypothesis space $\mathcal{F}_d$ of fully grown trees of depth d. Let the number of dimensions $m < 2^{2^d}$. Assume: (Leaves are correct) all leaves in all trees in the Rashomon set have at least $\lceil \theta n \rceil$ more correctly classified points than incorrectly classified points; (Bad features) there is a set of $m_{bad} \geq d$ "bad" features such that the empirical risk minimizer of models using only the bad features is not in the Rashomon set. Then $\hat{R}_{ratio}(\mathcal{F}_{d+1},\theta) < \hat{R}_{ratio}(\mathcal{F}_d,\theta)$.*

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

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

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

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

Second, since the empirical risk minimizer of models using only the bad features m_{bad} is not in the Rashomon set, every tree that has only features from the set m_{bad} is not in the Rashomon set. Therefore, trees in the Rashomon set must have at least one "good" feature at some node, where good means that the feature is not in m_{bad}. The cardinality of the set of good features is $\bar{m} = m - |m_{bad}|$, then the cardinality of the Rashomon set is:

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

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

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

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

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

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

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

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

□

F Proof for Theorem 7

Recall Theorem 7:

Theorem 7 (Rashomon ratio increases with noise for ridge regression)**.** *Consider dataset $S = X \times Y$, X is a non-zero matrix, and a hypothesis space of linear models $\mathcal{F} = \{f = \omega^T x, \omega \in \mathbb{R}^m, \omega^T \omega \leq \hat{L}_{\max}/C\}$. Let ϵ_i, such that $\epsilon_i \sim \mathcal{N}(\bar{0}, \lambda I)$ ($\lambda > 0$, I is identity matrix), be i.i.d. noise vectors added to every sample: $x_i' = x_i + \epsilon_i$. Consider options $\lambda_1 > 0$ and $\lambda_2 > 0$ that control how much noise we add to the dataset. For ridge regression, if $\lambda_1 < \lambda_2$, then the Rashomon ratios obey $\hat{R}_{ratio_{\lambda_1}}(\mathcal{F}, \theta)) < \hat{R}_{ratio_{\lambda_2}}(\mathcal{F}, \theta))$.*

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

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

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

Therefore, adding attribute noise to the training data becomes equivalent to ℓ_2-regularization, and the new regularization parameter is $C + \lambda$. According to Theorem 10 in Semenova et al. [39], the Rashomon volume can be computed as:

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

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

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

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

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

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

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

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

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

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

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

□

Compared to the Rashomon ratio, the relationship between the regularization parameter and the Rashomon volume is inverted: the stronger the regularization, the smaller the Rashomon volume. This means that adding more noise leads to stronger regularization and smaller Rashomon volume. In some ways, this is consistent with what we saw in Figure 1(b), where CART preferred shorter trees in the presence of noise.

Next we show that the variance of losses (recall notation $\sigma^2(f, \mathcal{D}) = \mathrm{Var}_{z\sim\mathcal{D}}\, l(f, z))$ increases for the least squares loss function under additive attribute noise, as in Step 1 of the path in Section 4:

Theorem 19 (Variance of least squares loss increases with noise). *Consider dataset $S = X \times Y$, where $X \in \mathbb{R}^{n\times m}$, $Y \in \mathbb{R}^n$, $z_i = (x_i, y_i)$. Let $\epsilon_i = \{\epsilon_{ij}\}_{j=1}^m$, such that $\epsilon_{ij} \sim \mathcal{N}(0, \sigma_\mathcal{N}^2)$, be i.i.d. noise vectors added to every sample: $x_i' = x_i + \epsilon_i$. Consider $\sigma_{\mathcal{N}_1}^2 > 0$, $\sigma_{\mathcal{N}_2}^2 > 0$ that control how much noise is added to the dataset. For the least squares loss $l(z_i) = r_i^2 = (w^T x_i - y_i)^2$ and a fixed model $f(x) = \omega^T x$, where $\omega \in \mathbb{R}^m$, $\omega \neq \bar{0}$, the variance of losses increases with more noise: if $\sigma_{\mathcal{N}_1}^2 < \sigma_{\mathcal{N}_2}^2$, then: $\sigma^2(f, S_{\sigma_{\mathcal{N}_1}}) < \sigma^2(f, S_{\sigma_{\mathcal{N}_2}})$.*

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

$$\begin{aligned}Var_{z,\bar{\epsilon}}\left[\left((x+\epsilon)^T\omega - y\right)^2\right] = Var_{z,\bar{\epsilon}}\left[\left((x^T\omega - y) + \epsilon^T\omega\right)^2\right] = Var_{z,\bar{\epsilon}}\left[\left(r + \epsilon^T\omega\right)^2\right] \\ = \mathbb{E}_{z,\bar{\epsilon}}\left[\left(r + \epsilon^T\omega\right)^4\right] - \left(\mathbb{E}_{z,\bar{\epsilon}}\left[\left(r + \epsilon^T\omega\right)^2\right]\right)^2.\end{aligned} \tag{6}$$

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

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

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

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

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

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

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

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

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

Let's focus on the first term of the variance equation (6):

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

Now, we focus of the second term of the variance equation (6):

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

Therefore, for the variance we get that:

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

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

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

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

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

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

□

As before, Corollary 3 is easily extendable to the results of Theorem 19, meaning that the maximum variance of losses, $\sigma^2 = \sup_{f\in\mathbb{R}_{set}(\mathcal{F},\gamma)} Var_{z\sim\mathcal{D}}l(f,z)$, will also increase for the least squares loss under increasing additive attribute noise.

Next, we show that when the maximum variance σ^2 increases, the generalization bound becomes worse for the least squares loss and the continuous hypothesis space. Cucker and Smale [11] proved the generalization bound based on Bernstein's inequality for the least squares loss (Theorem B). We state and provide proof of the theorem for the true Rashomon set. To derive the generalization bound, we use the covering number over the true Rashomon set. Recall that for the functional space $\mathcal{F}$ and any $\epsilon > 0$, the ℓ_∞ *covering number* $N(\mathcal{F}, \epsilon)$ of $\mathcal{F}$ is the minimum number of balls of radius ϵ, such that they can cover $\mathcal{F}$, meaning that there exist $h_1, ..., h_{N(\mathcal{F},\epsilon)} \in \mathcal{F}$, such that for every $f \in \mathcal{F}$ there is $k \leq N(\mathcal{F}, \epsilon)$ such that $\|f - h_k\|_\infty = \max_{x\in\mathcal{X}} |f(x) - h_k(x)| \leq \epsilon$. Now we focus on the theorem.

Theorem 20 (Variance-based "generalization bound" for least squares loss)**.** *Consider data distribution $\mathcal{D}$ over $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$, dataset $S = \{z_i\}_{i=1}^n \sim \mathcal{D}$, hypothesis space $\mathcal{F}$, and the least squares loss $l(f, z) = (f(x) - y)^2$. Let the loss be bounded by $C^2 > 0$ such that $l(f, z) \leq C^2$ for every $z \in \mathcal{Z}$. For any $\epsilon > 0$:*

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

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

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

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

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

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

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

Therefore, if $L(f)-\hat{L}(f)>\epsilon$, we have $L(h_k)-\hat{L}(h_k)\ge(L(f)-\hat{L}(f))-\frac{\epsilon}{2}>\epsilon-\frac{\epsilon}{2}=\frac{\epsilon}{2}$. This holds for every $f\in B_k$, and thus for $\arg\sup_{f\in B_k}$ as well:

$$
P\left(\sup_{f\in B_k} L(f)-\hat{L}(f)>\varepsilon\right)\le P\left(L(h_k)-\hat{L}(h_k)>\frac{\varepsilon}{2}\right). \tag{7}
$$

Since the exponential function is monotonic, $e^{-\frac{1}{\left(\sigma_{h_k}^2\right)}}\le e^{-\frac{1}{\sigma^2}}$. Based on the definition of the covering number, according to the union bound and (7) we have that:

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

Therefore we obtained the desired bound. □

Since $e^{-1/x}$ monotonically increases for $x>0$, in Theorem 20, as the maximum variance of losses increases, the bound on the right side increases as well, and thus the generalization bound becomes worse.

G Pattern Diversity and Other Metrics of the Rashomon set

In this appendix, we discuss similarities and differences between pattern diversity and pattern Rashomon ratio as well as expected pairwise disagreement (as in [5]).

Pattern Rashomon ratio. Pattern Rashomon ratio measures how expressive the Rashomon set is compared to the whole hypothesis space. Interestingly, for the hypothesis space of linear models, for different datasets with the same number of samples and attributes, as long as no $m-1$ points are collinear, the denominator of the pattern Rashomon ratio is the same and equal to $2\sum_{i=0}^{m}\binom{n-1}{i}$ [10]. If we focus only on the numerator of the pattern Rashomon ratio, it is the number of distinct predictions, whereas the pattern diversity is the average Hamming distance between distinct predictions.

Intuitively, the more distinct prediction we have, the more different they are from each other, and the higher pattern diversity we should expect. However, it is not always the case, and there exists datasets such that we can have a large number of patterns with very small Hamming distance and a small number of patterns with larger Hamming distance.

Similarly to pattern diversity, we can upper-bound the number of patterns in the pattern Rashomon set by a bound that depends on the empirical risk of the empirical risk minimizer and the Rashomon parameter θ. We discuss this bound in Lemma 21.

Lemma 21. *Given the dataset S of size n, the pattern Rashomon set $\pi(\mathcal{F}, \theta)$, the empirical risk of the empirical risk minimizer $\hat{L}(\hat{f})$, and the Rashomon parameter θ, the cardinality of the pattern Rashomon set obeys:*

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

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

Expected pairwise disagreement (as in [5]). Following [6] and [30], empirical expected pairwise disagreement $I(\hat{R}_{set}(\mathcal{F}, \theta))$ over the Rashomon set can be defined as $I(\hat{R}_{set}(\mathcal{F}, \theta)) = \mathbb{E}_{f_1, f_2 \sim \hat{R}_{set}(\mathcal{F}, \theta)} \hat{\mathbb{E}}_{z \sim S} \mathbb{1}_{[f_1(x) \neq f_2(x)]}$. Expected pairwise disagreement measures the average disagreement between every two hypotheses from the Rashomon set, while pattern diversity measures the average disagreement between two patterns from the Rashomon set. The expected pairwise disagreement is equivalent to pattern diversity when every pattern is achievable with the same probability by models from the Rashomon set. However, these metrics can be very different and we can have a small expected pairwise disagreement and larger pattern diversity as we show next.

Proposition 22 (Same pattern diversity but different expected pairwise disagreement)**.** *Consider finite Rashomon set $\hat{R}_{set}(\mathcal{F}, \theta)$ of size $d \geq 2$. Let $\pi(\mathcal{F}, \theta)$ be the pattern set of size Π, $2 \leq \Pi \leq d$. Assume that every pattern except p_1 is achievable by only one hypothesis in the Rashomon set, and thus p_1 is achievable by $d - \Pi + 1$ hypotheses. Let d^* be the current value of d, then as $d \to \infty$ (for example, by replicating hypotheses that realize p_1 an infinite number of times), expected pairwise disagreement converges to zero, $I(\hat{R}_{set}(\mathcal{F}_d, \theta)) \to 0$, and pattern diversity does not change, $div(\hat{R}_{set}(\mathcal{F}_d, \theta)) = div(\hat{R}_{set}(\mathcal{F}_{d^*}, \theta))$.*

Proof. The proof proceeds in two steps.

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

Expected pairwise disagreement. Given a pattern $p \in \pi(\mathcal{F}, \theta)$, let $P_{f \sim \hat{R}_{set}(\mathcal{F}, \theta)} \left[p = p^f\right]$ be a probability with which this pattern is achieved by models from the Rashomon set. Since support for all patterns except p_1 is 1, then $P_k = P_{f \sim \hat{R}_{set}(\mathcal{F}, \theta)} \left[p_k = p^f\right] = \frac{1}{d}$ for $k = 2..\Pi$. And for p_1 we have $P_1 = P_{f \sim \hat{R}_{set}(\mathcal{F}, \theta)} \left[p_1 = p^f\right] = \frac{d - \Pi + 1}{d}$. Then expected pairwise disagreement:

$$\begin{aligned} I(\hat{R}_{set}(\mathcal{F}_d, \theta)) &= \mathbb{E}_{f_1, f_2 \sim \hat{R}_{set}(\mathcal{F}, \theta)} \hat{\mathbb{E}}_{x \sim S} \mathbb{1}_{[f_1(x) \neq f_2(x)]} \\ &= \sum_{k=1}^{\Pi} \left[P_{f \sim \hat{R}_{set}(\mathcal{F}, \theta)} \left[p_k = p^f\right] \sum_{j=1}^{\Pi} P_{f \sim \hat{R}_{set}(\mathcal{F}, \theta)} \left[p_j = p^f\right] \frac{1}{n} H(p_k, p_j) \right] \\ &= \left(P_{f \sim \hat{R}_{set}(\mathcal{F}, \theta)} \left[p_1 = p^f\right] \right)^2 \frac{1}{n} H(p_1, p_1) \end{aligned}$$

$$\begin{aligned}
&+ 2P_{f\sim \hat{R}_{set}(\mathcal{F},\theta)}\left[p_1 = p^f\right]\sum_{j=2}^{\Pi} P_{f\sim \hat{R}_{set}(\mathcal{F},\theta)}\left[p_j = p^f\right]\frac{1}{n}H(p_1,p_j)\\
&+ \sum_{k=2}^{\Pi}\left[P_{f\sim \hat{R}_{set}(\mathcal{F},\theta)}\left[p_k = p^f\right]\sum_{j=2}^{\Pi} P_{f\sim \hat{R}_{set}(\mathcal{F},\theta)}\left[p_j = p^f\right]\frac{1}{n}H(p_k,p_j)\right]\\
&= 0 + 2\frac{d-\Pi+1}{d^2}\sum_{j=2}^{\Pi}\frac{1}{n}H(p_1,p_j) + \frac{1}{d^2}\sum_{k=2}^{\Pi}\sum_{j=2}^{\Pi}\frac{1}{n}H(p_k,p_j)\\
&= \frac{1}{d}\left(2\left(1-\frac{\Pi-1}{d}\right)\sum_{j=2}^{\Pi}\frac{1}{n}H(p_1,p_j) + \frac{1}{d}\sum_{k=2}^{\Pi}\sum_{j=2}^{\Pi}\frac{1}{n}H(p_k,p_j)\right).
\end{aligned}$$

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

As we see from Proposition 22, we can change expected pairwise disagreement, for example, by adding multiple copies of the same functions f to the hypothesis space. Expected pairwise disagreement measures predictive multiplicity [6], but it has the issue we showed above that it can depend on the weighting of hypotheses in the hypothesis space. In the case we described in Proposition 22, the multiplicity is small because one subset of hypotheses (which produce the same pattern) is weighted very heavily. Thus, expected pairwise disagreement can be influenced by overparameterization or poor choice of parameter space. We further illustrate the effect of re-parameterization on pairwise disagreement on a simple one-dimensional example in Figure 6. Pattern diversity does not depend on the parameter space and is computed in the pattern space. It is not impacted by any probability distribution or weighting on the hypotheses. Moreover, we can compute the pattern diversity by enumerating all possible patterns of the given finite dataset as described in Appendix K.3. We cannot do the same for the pairwise disagreement metric without additional assumptions on the patterns' support.

H Proof for Theorem 9

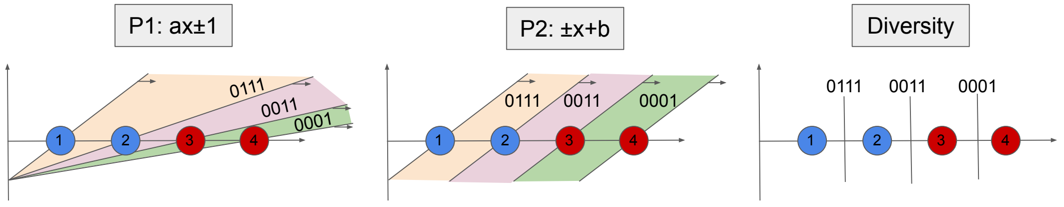
Recall that a_i is sample agreement over the pattern Rashomon set. Then we can compute the pattern diversity based on sample agreement as in Theorem 9.

Theorem 9 (Pattern diversity via sample agreement)**.** *For 0-1 loss, dataset S, and pattern Rashomon set $\pi(\mathcal{F},\theta)$, pattern diversity can be computed as $div(\hat{R}_{set}(\mathcal{F},\theta)) = \frac{2}{n}\sum_{i=1}^{n} a_i(1-a_i)$, where $a_i = \frac{1}{\Pi}\sum_{k=1}^{\Pi}\mathbb{1}_{[p_k^i = y_i]}$ is sample agreement over the pattern Rashomon set.*

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

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

$$\begin{aligned}
div(\hat{R}_{set}(\mathcal{F},\theta)) &= \frac{1}{n\Pi\Pi}\sum_{j=1}^{\Pi}\sum_{k=1}^{\Pi}\sum_{i=1}^{n}\mathbb{1}_{[p_j^i \neq p_k^i]} =\\
&= \frac{1}{n\Pi\Pi}\sum_{j=1}^{\Pi}\sum_{k=1}^{\Pi}\sum_{i=1}^{n}\left[p_j^i(1-p_k^i) + p_k^i(1-p_j^i)\right]\\
&= \frac{1}{n\Pi\Pi}\sum_{j=1}^{\Pi}\sum_{k=1}^{\Pi}\sum_{i=1}^{n}\left[p_j^i + p_k^i - 2p_k^i p_j^i\right]\\
&= \frac{1}{n\Pi}\sum_{i=1}^{n}\sum_{j=1}^{\Pi}\left[\frac{1}{\Pi}\sum_{k=1}^{\Pi}p_j^i + \frac{1}{\Pi}\sum_{k=1}^{\Pi}p_k^i - 2\frac{1}{\Pi}\sum_{k=1}^{\Pi}p_k^i p_j^i\right]
\end{aligned}$$



(a) (b) (c)

Figure 6: Illustration of how reparameterization changes pairwise disagreement metric. Consider a separable dataset of four data points with a real-valued feature in one dimension: $S = \{(1,0),(2,0),(3,1),(4,1)\}$ and a hypothesis space of linear models. Let the Rashomon parameter be $\theta = 0.25$. There are three patterns in the Rashomon set: 0111, 0011, and 0001. The pattern diversity (c) is 0.444. Consider two different parameterizations for the hypothesis space of linear models: $ax \pm 1$ and $\pm x + b$. These two parameterizations produce the same decision boundaries for the dataset S. For the parameterization $\pm x + b$ (b), each pattern is achieved with the same number of models. For the parameterization $ax \pm 1$ (a), more models will support patterns that are closer to the origin. The support of each pattern is shown in a different color. The pairwise disagreement metric is 0.321 for $ax \pm 1$ and 0.444 for $\pm x + b$. (For the parameterization $ax \pm 1$, we see that the pattern 0001 occurs when $a \in (1, \frac{1}{2})$, the pattern 0011 occurs when $a \in (\frac{1}{2}, \frac{1}{3})$, and the pattern 0111 occurs when $a \in (\frac{1}{3}, \frac{1}{4})$. Therefore, the pattern 0001 has probability $w_1 = \frac{1-\frac{1}{2}}{\frac{3}{4}} = 0.666$, the pattern 0011 has probability $w_2 = \frac{\frac{1}{2}-\frac{1}{3}}{\frac{3}{4}} = 0.222$, and the pattern 0111 has probability $w_3 = \frac{\frac{1}{3}-\frac{1}{4}}{\frac{3}{4}} = 0.111$. Recall that $H(cdot, \cdot)$ is the Hamming distance, then the pairwise disagreement metric is $w_1 w_2 H(0001, 0011) + w_1 w_3 H(0001, 0111) + w_2 w_3 H(0011, 0111) = w_1 w_2 + 2 w_1 w_3 + w_2 w_3 = 0.321$. For the parameterization $\pm x + b$, each pattern has equal probability $\frac{1}{3}$. We can then similarly calculate that the pairwise disagreement metric is 0.444). Note that if the data points are shifted together to the left, the difference in pairwise disagreement metrics for parameterizations in (a) and (b) will only grow.

$$
\begin{aligned}
&= \frac{1}{n\Pi} \sum_{i=1}^{n} \sum_{j=1}^{\Pi} \left[p_j^i + b_i - 2 b_i p_j^i\right] \\
&= \frac{1}{n} \sum_{i=1}^{n} \left[\frac{1}{\Pi} \sum_{j=1}^{\Pi} p_j^i + \frac{1}{\Pi} \sum_{j=1}^{\Pi} b_i - 2 \frac{1}{\Pi} \sum_{j=1}^{\Pi} b_i p_j^i\right] \\
&= \frac{1}{n} \sum_{i=1}^{n} \left[b_i + b_i - 2 b_i^2\right] \\
&= \frac{2}{n} \sum_{i=1}^{n} b_i (1 - b_i).
\end{aligned}
$$

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

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

$$= 2y_ib_i + 1 - y_i - b_i.$$

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

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

Therefore we get:

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

□

I Proof for Theorem 10

Before providing the proof for Theorem 10, we show that average sample agreement over hypotheses that realize patterns in the pattern Rashomon set is negatively proportional to the average loss of these hypotheses. We use this intuition to derive an upper bound for average sample agreement and then discuss the upper bound for pattern diversity.

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

Theorem 23. *Average sample agreement over the pattern Rashomon set is negatively proportional to the average loss of models in the hypothesis pattern Rashomon set* $\mathcal{H}_\pi(\mathcal{F}, \theta)$*,*

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

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

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

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

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

---

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

$$
\begin{aligned}
&= 1 - \frac{1}{n}\sum_{i=1}^{n}\frac{1}{\Pi}\sum_{k=1}^{\Pi}\mathbb{1}_{[p_k^i \neq y_i]} \\
&= 1 - \frac{1}{\Pi}\sum_{k=1}^{\Pi}\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}_{[f_k^\pi(x_i) \neq y_i]} \\
&= 1 - \frac{1}{\Pi}\sum_{k=1}^{\Pi}\hat{L}(f_k^\pi) \\
&= 1 - \hat{L}_{avg}(\mathcal{H}_\pi(\mathcal{F},\theta)).
\end{aligned}
$$

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

Given the definition of models in the Rashomon set, we can derive an upper bound on average sample agreement in Corollary 24.

Corollary 24. *For any parameter $\theta > 0$, average sample agreement is upper and lower bounded by the empirical loss of the empirical risk minimizer,*

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

Proof. Proof follows directly from Theorem 23 and the fact that for every model f from the Rashomon set, $\hat{L}(\hat{f}) \leq \hat{L}(f) \leq \hat{L}(\hat{f}) + \theta$. ☐

Finally, we provide proof for Theorem 10.

Theorem 10 (Upper bound on pattern diversity)**.** *Consider hypothesis space $\mathcal{F}$, 0-1 loss, and empirical risk minimizer $\hat{f}$. For any $\theta \geq 0$, pattern diversity can be upper bounded by*

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

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

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

Given this and from the definition of pattern diversity and Corollary 24 we get that:

$$
\begin{aligned}
div(\hat{R}_{set}(\mathcal{F},\theta)) &= \frac{2}{n}\sum_{i=1}^{n} a_i(1-a_i) = \frac{2}{n}\sum_{i=1}^{n} a_i - \frac{2}{n}\sum_{i=1}^{n} a_i^2 \\
&\leq \frac{2}{n}\sum_{i=1}^{n} a_i - \frac{2}{n^2}\left(\sum_{i=1}^{n} a_i\right)^2 = 2\left(\frac{1}{n}\sum_{i=1}^{n} a_i\right) - 2\left(\frac{1}{n}\sum_{i=1}^{n} a_i\right)^2 \\
&\leq 2(1 - \hat{L}(\hat{f})) - 2(1 - \hat{L}(\hat{f}) - \theta)^2 \\
&= 2 - 2\hat{L}(\hat{f}) - 2 + 4(\hat{L}(\hat{f}) + \theta) - 2(\hat{L}(\hat{f}) + \theta)^2 \\
&= 2(\hat{L}(\hat{f}) + \theta - (\hat{L}(\hat{f}) + \theta)^2 + \theta) \\
&= 2(\hat{L}(\hat{f}) + \theta)(1 - (\hat{L}(\hat{f}) + \theta)) + 2\theta.
\end{aligned}
$$

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

J Proof for Theorem 11

We state and prove Theorem 11 below.

Theorem 11. *Consider a hypothesis space $\mathcal{F}$, 0-1 loss, and a dataset S. Let $\rho \in (0, \frac{1}{2})$ be the probability with which each label y_i is flipped independently, and let $S_\rho \sim \Omega(S_\rho)$ denote a noisy version of S. For the Rashomon parameter $\theta \geq 0$, if $\inf_{f\in\mathcal{F}} \mathbb{E}_{S_\rho} \hat{L}_{S_\rho}(f) < \frac{1}{2} - \theta$ and $\hat{L}_S(\hat{f}_S) < \frac{1}{2}$, then adding noise to the dataset increases the upper bound on pattern diversity of the expected Rashomon set:*

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

Proof. Given the noise model, hypothesis f is in the expected Rashomon set if $\mathbb{E}_{S_\rho} \hat{L}_{S_\rho}(f) \leq \inf_{f\in\mathcal{F}} \mathbb{E}_{S_\rho} \hat{L}_{S_\rho}(f) + \theta$. Let $\bar{f} \in \mathcal{F}$ be such that $\bar{f} \in \arg\inf_{f\in F} \mathbb{E}_{S_\rho} \hat{L}_{S_\rho}(f)$. Since $\rho \in (0, \frac{1}{2})$, and $\hat{L}_S(\hat{f}_S) < \frac{1}{2}$ by assumption, from (3) and the definition of the empirical risk minimizer, we have that:

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

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

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

□

Interestingly, in the proof of Theorem 11, the empirical risk minimizer of dataset S, $\hat{f}_S$, also minimizes the expected risks over noisy datasets, meaning that $\hat{f}_S \in \arg\inf_{f\in F} \mathbb{E}_{S_\rho} \hat{L}_{S_\rho}(f)$. To see this, assume that $\hat{f}_S \notin \arg\inf_{f\in F} \mathbb{E}_{S_\rho} \hat{L}_{S_\rho}(f)$, then there is $\bar{f} \in \arg\inf_{f\in F} \mathbb{E}_{S_\rho} \hat{L}_{S_\rho}(f)$, such that:

$$\mathbb{E}_{S_\rho} \hat{L}_{S_\rho}(\bar{f}) < \mathbb{E}_{S_\rho} \hat{L}_{S_\rho}(\hat{f}_S).$$

Applying (3) to both sides of the inequality above, we get that:

$$(1-2\rho)\hat{L}_S(\bar{f}) + \rho < (1-2\rho)\hat{L}_S(\hat{f}_S) + \rho,$$

which after simplification becomes:

$$\hat{L}_S(\bar{f}) < \hat{L}_S(\hat{f}_S).$$

This is a clear contradiction, since $\hat{f}_S$ is the empirical risk minimizer on S, and thus $\hat{L}_S(\hat{f}_S) \leq \hat{L}_S(f)$ for any $f \in \mathcal{F}$, including $\bar{f}$. Therefore our assumption was incorrect, and $\hat{f}_S \in \arg\inf_{f\in F} \mathbb{E}_{S_\rho} \hat{L}_{S_\rho}(f)$.

K Setup for experiments

K.1 Datasets Description

Please see Table K.1 for the description of datasets used in the paper and all the processing steps. We normalize all real-valued features.

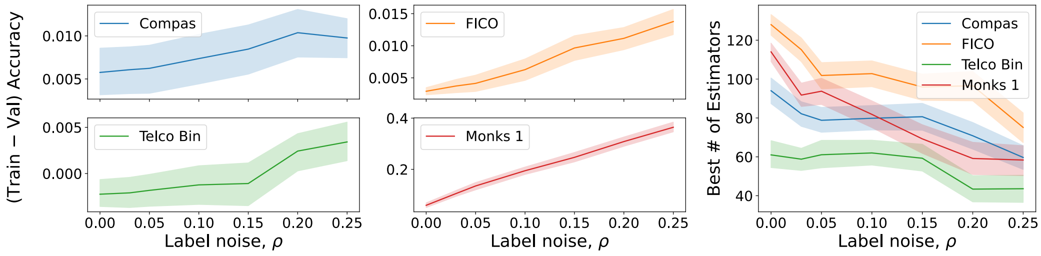
Table 1: Preprocessed datasets

Dataset	Number of Samples	Number of Features	Notes
Car Evaluation	1728	16	We use one-hot encoding for features
Breast Cancer Wisconsin	699	11	We use one-hot encoding for features
Monks 1	124	12	We use one-hot encoding for features
Monks 2	169	12	We use one-hot encoding for features
Monks 3	122	12	We use one-hot encoding for features
SPECT	267	23	We use one-hot encoding for features
COMPAS	6907	13	Processed in [52]
FICO	10459	18	Processed in [52]
Bar 7 (Coupon)	1913	15	Processed in [52]
Expensive Restaurant	1417	16	Processed in [52]
Carryout Takeaway	2280	16	Processed in [52]
Cheap Restaurant	2653	16	Processed in [52]
Coffee House	3816	16	Processed in [52]
Bar	1913	16	Processed in [52]
Telco Bin	7043	6	We use only the binary features
Iris	100	4	We consider classes Versicolour and Setosa
Wine	130	13	
Wine 4	130	4	We use PCA to create 4 features
Seeds 4	140	4	We consider classes 1 and 2 and use PCA to create 4 features
Immunotherapy 4	90	4	[21, 22]. We use one-hot encoding for feature “type”. We use PCA to create 4 features
Penguin 4	265	4	We use one-hot encoding for feature “island.” We consider classes “adelie” and “gentoo” only and use PCA to create 4 features
Digits 0-4 4	359	4	We consider digit 0 and digit 4. We use PCA to create 4 features

K.2 Illustration of Cross-Validation Process in Step 3

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

For Figure 1(a), we tuned the parameters and then added noise to see what happens, which is that performance degrades. For every train/validation split we performed 5-fold cross-validation



(a)

(b)

Figure 7: Practitioner’s validation process in the presence of noise for gradient boosted trees. For a fixed number of estimators, as we add noise, the gap between training and validation accuracy increases (Subfigure a). As we use cross-validation to select the number of estimators, the best number of estimators decreases with noise (Subfigure b).

on the training set and computed the best depth. We fixed this depth (and thus hypothesis space). Then, we start adding noise to the dataset. We considered six different noise levels, $\rho \in \{0, 0.03, 0.05, 0.10, 0.15, 0.20.0.25\}$. For every level, we performed 25 draws of S_ρ. For every noise level, noise draw, and train/validation split, we evaluated train and validation performance and reported the average.

For Figure 1(b), we tuned the parameters for each noise level. We will see that noisier datasets lead us to use more regularization. We started adding noise to the dataset and then chose the best parameter based on cross-validation. More specifically, we considered six different noise levels, $\rho \in \{0, 0.03, 0.05, 0.10, 0.15, 0.20.0.25\}$. For every level, we performed 25 draws of S_ρ. Then we performed 5-fold cross-validation on the training data to choose the best depth for CART. For every noise level, noise draw, and train/validation split, we report mean depth based on cross-validation results.

We performed a similar experiment to Figure 1 for the gradient boosting algorithm, where we varied the number of tree estimators. We observe similar behaviors, where, with more noise, the best number of estimators (according to cross-validation) decreases. We used the same level of noise and cross-validation procedure as discussed above. For the number of estimators, we considered values $d \in \{5, 10, 20, \dots, 150\}$.

K.3 Branch and Bound Method to Compute Patterns in the Rashomon set

Here we describe a two-step method that allows us to compute all patterns in the pattern Rashomon set. In the first step, we reduce the complexity of the problem, by discarding points that have low sample agreement. In the second step, we use a branch-and-bound approach in order to enumerate patterns and discard prefixes of those patterns that will not be in the Rashomon set based on the Rashomon parameter and the empirical risk of the empirical risk minimizer.

Consider a dataset $S = \{z_i\}_{i=1}^n$. For every point z_i assume that we have sample agreement a_i. If $a_i = 0$, it means that all patterns in the pattern Rashomon set assign an incorrect label to sample z_i. On the other hand, if $a_i = 1$, all patterns assign the correct label. If we exclude all z_i such that $a_i = 0$ or $a_i = 1$ from the dataset, then the number of patterns will not change in the Rashomon set. Therefore, for a given point z_k ($k = 1..n$) we will try to answer a question: is there a model in the Rashomon set such that it classifies $\bar{z}_k = (x_k, -y_k)$ correctly and still stays in the Rashomon set. If there is no such model, then sample z_k has no influence on the pattern Rashomon set. Since it is harder to optimize for 0-1 loss, we instead consider exponential loss. If the problem is separable by 0-1 loss, then exponential loss will converge to a separable solution exponentially fast (which is known from the convergence analysis of AdaBoost [4]). Then given hypothesis space of linear models $\mathcal{F} = \{w^T x\}$, for every z_k, we aim to solve following optimization problem:

$$\min \frac{1}{n} \sum_{i=0}^{n} e^{-y_i w^T x_i} \tag{8}$$

$$y_k w^T x_k \le 0, \tag{9}$$

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

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

```
procedure DISCARD POINTS(dataset S, ERM $\hat{f}$, the Rashomon parameter $\theta$)
    Initialize $S_{dp}$.
    for every $z = (x, y) \in S$ do
        Solve optimization problem (8)-(9). Let $\bar{f}$ be a solution.
        if $\hat{L}(\bar{f}) > \hat{L}(\hat{f}) + \theta$ then
            add $z$ to $S_{dp}$   (this point has a single predicted label for the entire Rashomon set).
        end if
    end for
    return $S_{dp}$.
end procedure
```

In the second step, we build a search tree over the set of patterns that are formed by samples in S_{inside}. We use breadth-first search over subsets of data. We “bound” (i.e., exclude part of the search space) when the prefix of the pattern (which is the part of the dataset we are working with) misclassifies more samples than the threshold to stay in the Rashomon set, which is $\hat{L}(\hat{f}) + \theta$. Since not all patterns can be realized by the model class. We “bound” if the prefix or pattern can not be achieved (the pattern is achievable when all points with labels matching the pattern are classified correctly by some model from the hypothesis space). In order to perform branch and bound more effectively, given an empirical risk minimizer (ERM), we sort points in the dataset based on their distance to the decision boundary of the ERM. More specifically, we split the points into four categories depending on whether the point is a true positive, false positive, true negative, or false negative. Then for every category, we compute the distances from each point to the decision boundary of the ERM and then sort points from least distance to greatest distance. Finally, we cyclically choose one point from each category until all samples have been considered. Conceptually, true positive and true negative samples that are closest to the decision boundary determine most of the patterns in the pattern Rashomon set. We add false positives and false negatives early to the order of points as they are more likely to be misclassified, allowing us to bound the prefixes sooner. We describe the branch and bound procedure in Algorithm 1. We use bit vectors to represent prefixes and patterns to speed up computations. Since we apply this approach to linear models, we use logistic regression without regularization to check the achievability of the patterns and their prefixes. However, the algorithm in general can be applied to other hypothesis spaces and losses (for example, hinge loss).

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

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

K.4 Computation of Rashomon Ratio and Pattern Rashomon Ratio for Step 4

We considered the hypothesis space of sparse decision trees of various depths and the hypothesis space of linear models with a given number of non-zero coefficients.

For decision trees (Figure 2 (a)), we varied the depth of the maximum allowed decision tree from 1 to 7. To compute the numerator of the Rashomon ratio, we used TreeFARMS [52]. We set the Rashomon parameter to 5%. To compute the denominator of the Rashomon ratio, we considered all possible sparse decision trees up to a given depth d and used the following recursive formula to compute the size of hypothesis space with m features:

$$C(d, m) = 2 + mC(d-1, m-1)^2,$$

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

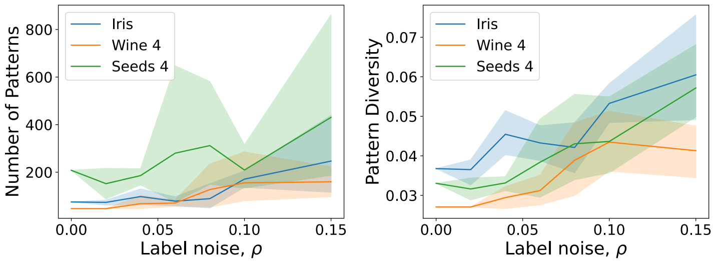
Note that for decision trees of depth exactly equal to d for every tree path, following recursive formula holds

$$C_{complete}(d, m) = mC_{complete}(d-1, m-1)^2,$$

which is equivalent to closed-form solution described in appendix E.

For the hierarchy of regularized linear models (Figure 2 (b)), we considered regularization for 1 non-zero coefficient, 2 non-zero coefficients, 3 non-zero coefficients, and 4 non-zero coefficients. To compute the numerator of the pattern Rashomon ratio, we used the approach described in Section K.3. We set the Rashomon parameter to 3%. To compute the denominator of the Rashomon ratio, we used the formula that gives the number of all possible patterns for the hypothesis space of linear models: if no $m-1$ points are coplanar, $C(n, m) = 2\sum_{i=0}^{m}\binom{n-1}{i}$ [10].

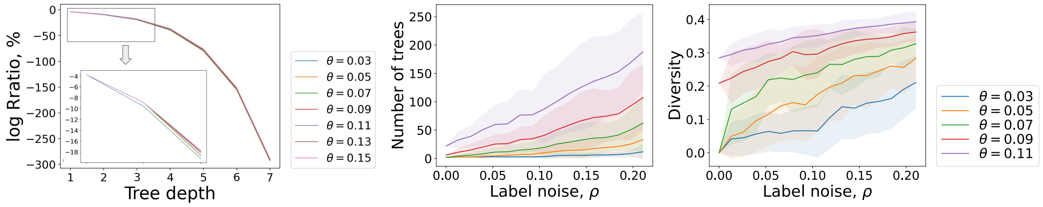
K.5 Experiments for Pattern Diversity and Label Noise for Linear Models



(a) (b)

Figure 8: Rashomon set characteristics such as the number of patterns in the Rashomon set (Subfigure a) and pattern diversity (Subfigure b) tend to increase with uniform label noise for hypothesis spaces of linear models.

For the hypothesis space of linear classifiers, we show the pattern diversity and the number of patterns in the Rashomon set for different datasets in the presence of noise in Figure 8. We considered uniform label noise, where each label is flipped independently with probability ρ. We set noise level ρ to values in $\{0, 0.02, 0.04, 0.06, 0.08, 0.10, 0.15\}$ and performed five draws of S_ρ for every noise level. We then computed the pattern Rashomon set for each draw using the method described in Appendix K.3 and finally computed the pattern diversity. Both the number of patterns and pattern diversity tend to increase with label noise.



(a) (b) (c)

Figure 9: (a) Curve of the hypothesis space complexity vs. Rashomon ratio (as in Figure 2) stays the same shape for different Rashomon parameters for the Monks 3 dataset for the hypothesis space of sparse decision trees. (b, c) Rashomon characteristics tend to increase with uniform label noise for the hypothesis space of decision trees (as in Figure 3) for different Rashomon parameters for the Monks 3 dataset. For (b) and (c), we averaged over 25 iterations.

K.6 The Choice of the Rashomon Parameter does not Influence Results

For Figures 2(a) and 3, we set the Rashomon parameter to be 5%. In Figure 9, on the example of Monks 3 dataset, we show that the results in Figures 2(a) and 3 hold for different values of the Rashomon parameter.

K.7 Computation Resources

We performed experiments on Duke University's Computer Science Department cluster. We parallelized computations for the majority of the figures. It took up to 3 hours to compute the Rashomon sets for the hypothesis space of sparse decision trees for different noise levels and draws (Figures 3 and 9), and up to 48 hours for the hypothesis space of linear models (Figures 2 and 8).