When Does Model Multiplicity Affect Prediction Intervals? A Sharp Phase Transition

Hai Nguyen [*1] **Khanh Nguyen**[1,3] **Hung Mai**[2,3] **Luong Doan**[3] **Nhung Duong**[3] **Phong Ho**[3] **Tuan Do**[1,3]

[1]Phenikaa School of Computing, Phenikaa University, Hanoi, Vietnam
[2]National Economics University, Hanoi, Vietnam
[3]N2TP Technology Solutions JSC, Hanoi, Vietnam

Abstract

The Rashomon effect, the fact that many different models can achieve nearly identical predictive accuracy, has been extensively studied across various domains of machine learning. While modern conformal prediction provides valid intervals for a single model, it remains blind to the inherent arbitrariness of having to select from multiple candidates. This paper bridges this gap by investigating when model multiplicity, or the existence of many near-optimal models, materially widens prediction intervals. We prove a sharp phase transition for ridge regression under Gaussian design and introduce *Rashomon Prediction Intervals* (RPIs), which bound the range of predictions across the Rashomon set and establish a Rashomon-Conformal Bridge that provides distribution-free coverage guarantees. A critical tolerance $\varepsilon^\star = \Theta(\sigma^2/p)$ separates two distinct regimes: below it, Rashomon-conformal intervals match standard conformal width; above it, model multiplicity dominates, and this scaling is provably minimax optimal. Computationally, exact RPIs are polynomial for convex classes but NP-hard for tree ensembles, with a thin-cap volumetric barrier ruling out sampling-based approximation. Experiments on six UCI benchmarks confirm the phase transition, with all datasets falling in the subcritical regime at standard tolerances, indicating that model multiplicity is invisible in prediction intervals for typical tabular regression.

1 INTRODUCTION

Conformal prediction [Vovk et al., 2005, Lei et al., 2018] provides distribution-free prediction intervals for a single fitted model, capturing aleatoric uncertainty from irreducible noise. Rashomon set analysis [Breiman, 2001, Semenova et al., 2022, Rudin et al., 2024] characterizes the multiplicity of near-optimal models, revealing epistemic uncertainty from under-determination by finite data. Despite extensive studies, these frameworks have developed in isolation, with no remarkable connection to each other: conformal methods treat the fitted model as given, while Rashomon analyses (e.g., ambiguity/discrepancy [Marx et al., 2020], Rashomon capacity [Hsu and Calmon, 2022], hacking intervals [Coker et al., 2021]) lack formal coverage guarantees. This raises a natural question:

> *When does the existence of many near-optimal models actually enlarge prediction intervals beyond what noise alone requires?*

This paper aims to address this question by introducing Rashomon Prediction Intervals (RPIs) with conformal coverage guarantees and proving a sharp phase transition. Our contributions are: (1) the closed-form RPI width $W(x_0) = 2\sqrt{\varepsilon \cdot x_0^\top \hat{\Sigma}^{-1} x_0}$ for ridge regression, extending Coker et al. [2021] to regularized objectives (Theorem 3); (2) a sharp phase transition at $\varepsilon^\star = \Theta(\sigma^2/p)$ separating regimes where Rashomon-conformal intervals match vs. exceed conformal width by $\Omega(\sqrt{\varepsilon p})$ (Theorem 4); (3) a Rashomon-Conformal Bridge with distribution-free coverage $\geq 1-\alpha$ (Theorem 6); (4) NP-hardness for tree ensembles, a thin-cap barrier for sampling, and polynomial-time Rashomon quantile bands (Theorems 7, 8); (5) a lower bound showing $\sqrt{\varepsilon p}$ width scaling is optimal (Theorem 10).

The novelty of this work lies not in the RPI formula itself, but in the phase transition and conformal bridge it enables, together with the computational and information-theoretic boundaries.

[*]Correspondence: hai.nguyenduc@phenikaa-uni.edu.vn

Proceedings of the 42nd Conference on Uncertainty in Artificial Intelligence (UAI 2026), PMLR 337:4873–4889.

2 RELATED WORK

Conformal prediction. The foundational work of Vovk et al. [2005] established distribution-free validity via exchangeability; the textbook treatment of Angelopoulos et al. [2025] provides a unified theoretical foundation. Lei et al. [2018] developed split conformal inference for regression. Romano et al. [2019] introduced conformalized quantile regression for heteroscedastic data. Barber et al. [2021] introduced Jackknife+ with leave-one-out stability. Barber et al. [2023] extended conformal prediction beyond exchangeability using weighted quantiles. Gibbs et al. [2025] developed methods for conditional coverage via covariate-shift reweighting. Angelopoulos et al. [2024] generalized from coverage to arbitrary risk control. All of these methods condition on a single fitted model and thus capture only aleatoric uncertainty. Our Theorem 6 and Algorithm 1 show how to embed Rashomon-aware calibration into the split conformal pipeline, inheriting its finite-sample guarantees while accounting for model multiplicity.

Rashomon effect and predictive multiplicity. Breiman [2001] first observed the Rashomon effect in statistical modeling. Fisher et al. [2019] formalized model class reliance over the Rashomon set; Dong and Rudin [2020] extended this to visualize variable importance clouds across all good models. Marx et al. [2020] proposed ambiguity and discrepancy metrics for classification multiplicity; Watson-Daniels et al. [2023] generalized these to probabilistic classifiers via viable prediction ranges. Semenova et al. [2022] proved the existence of simple models in large Rashomon sets, and Semenova et al. [2023] subsequently showed that label noise is the key driver of Rashomon set size. Xin et al. [2022] gave complete enumeration of Rashomon sets for sparse decision trees. Hsu and Calmon [2022] introduced Rashomon capacity as an information-theoretic measure of score variation. Rudin et al. [2024] provided a comprehensive survey. Black et al. [2022] highlighted the problem of justifiability in model selection. Our work differs from all of these by producing prediction intervals with formal coverage guarantees and establishing computational-statistical phase transitions.

Hacking intervals. Coker et al. [2021] introduced hacking intervals for the range of an estimand across the Rashomon set of OLS models. Our Theorem 3 generalizes this to ridge regression and general convex classes; Theorems 4 and 6 provide the phase transition and coverage theory absent from their work; Theorem 7 provides hardness results for non-convex classes.

Computational-statistical tradeoffs. Chandrasekaran and Jordan [2013] established tradeoffs between statistical and computational resources via convex relaxation. Brennan and Bresler [2019] proved computational-statistical gaps for detection under planted clique. Buhai et al. [2024] proved gaps for improper learning in sparse regression. Areces et al. [2024] established fundamental limits on predictive uncertainty quantification in conformal settings. Park et al. [2022] studied PAC prediction sets under covariate shift. Our contribution identifies a new tradeoff specific to prediction interval computation: below the phase transition, standard conformal methods suffice; above it, Rashomon enumeration is necessary for intervals that are valid over the model class.

3 PROBLEM FORMULATION

3.1 SETUP AND NOTATION

Let $\mathcal{S} = \{(X_i, Y_i)\}_{i=1}^n$ be i.i.d. draws from P_{XY} on $\mathbb{R}^p \times \mathbb{R}$. We assume

$$Y = f^\star(X) + \xi, \quad \xi \sim \mathrm{subG}(\sigma^2), \quad \xi \perp X, \tag{1}$$

where $f^\star \in \mathcal{F}$ is the true regression function and $\mathcal{F}$ is a hypothesis class. We write $\hat{L}(f) = \frac{1}{n}\sum_{i=1}^n (f(X_i) - Y_i)^2$ for the empirical squared loss. More generally, we let $\hat{J}(f; \mathcal{S})$ denote a (possibly regularized) empirical objective; for unregularized ERM, $\hat{J} = \hat{L}$, while for ridge regression $\hat{J}(\theta) = \hat{L}_\lambda(\theta) = \hat{L}(\theta) + \lambda\|\theta\|^2$. We write $\hat{f} = \arg\min_{f \in \mathcal{F}} \hat{J}(f; \mathcal{S})$.

Definition 1 (Rashomon Set)**.** For tolerance $\varepsilon > 0$ and empirical objective $\hat{J}$, the Rashomon set is

$$\mathcal{R}(\mathcal{F}, \varepsilon, \mathcal{S}) = \left\{ f \in \mathcal{F} : \hat{J}(f; \mathcal{S}) \le \hat{J}(\hat{f}; \mathcal{S}) + \varepsilon \right\}. \tag{2}$$

Definition 2 (Rashomon Prediction Interval)**.** For a test point $x_0 \in \mathbb{R}^p$, the RPI is

$$\mathrm{RPI}(x_0) = \Big[\min_{f \in \mathcal{R}} f(x_0),\ \max_{f \in \mathcal{R}} f(x_0) \Big]. \tag{3}$$

We denote the width $W(x_0) = \max_{f \in \mathcal{R}} f(x_0) - \min_{f \in \mathcal{R}} f(x_0)$. Note $W(x_0)$ is not itself a valid prediction interval for Y; it measures the epistemic spread of predictions across the Rashomon set.

3.2 THE CENTRAL QUESTION

When does model multiplicity enlarge prediction intervals beyond what noise alone requires? Formally, let W_{CP} denote the expected width of a standard split-conformal interval at level $1 - \alpha$. We compare $\mathbb{E}[W(X_0)]$ to conformal width to locate the regime where epistemic spread is negligible relative to the noise-dominated conformal radius. We seek the threshold $\varepsilon^\star$ such that $\mathbb{E}[W(X_0)] \le W_{\mathrm{CP}} + o(1)$ for $\varepsilon < \varepsilon^\star$ and $\mathbb{E}[W(X_0)] \ge W_{\mathrm{CP}} + \Omega(g(\varepsilon, p, n))$ for $\varepsilon > \varepsilon^\star$.

4 RASHOMON PREDICTION INTERVALS FOR RIDGE REGRESSION

We first treat the convex case completely. In this section, n denotes the training sample size used to fit ridge regression;

in the context of Algorithm 1 (Section 6), this corresponds to $n_{\mathrm{tr}} = |\mathcal{S}_{\mathrm{train}}|$. Let $\mathcal{F}_{\mathrm{ridge}} = \{x \mapsto \theta^\top x : \theta \in \mathbb{R}^p\}$ with regularized loss $\hat{L}_\lambda(\theta) = \frac{1}{n}\|X\theta - Y\|^2 + \lambda\|\theta\|^2$, where $X \in \mathbb{R}^{n\times p}$ is the design matrix and $Y \in \mathbb{R}^n$ the response vector. The ridge estimator is $\hat{\theta} = (X^\top X/n + \lambda I)^{-1}(X^\top Y/n)$. Define $\hat{\Sigma} = X^\top X/n + \lambda I$. Throughout this section, the Rashomon set is defined with respect to the ridge objective: $\mathcal{R}_\lambda(\varepsilon) = \{\theta : \hat{L}_\lambda(\theta) \le \hat{L}_\lambda(\hat{\theta}) + \varepsilon\}$.[1]

Theorem 3 (Rashomon Ellipsoid and Exact RPIs)**.** *Assume $\hat{\Sigma} \succ 0$ (e.g., $\lambda > 0$, or $\lambda = 0$ and X has full column rank). The Rashomon set for ridge regression is the ellipsoid*

$$\mathcal{R}_\lambda(\varepsilon) = \left\{\theta \in \mathbb{R}^p : (\theta - \hat{\theta})^\top \hat{\Sigma}(\theta - \hat{\theta}) \le \varepsilon\right\}, \quad (4)$$

and for any test point $x_0 \in \mathbb{R}^p$, the RPI is

$$\mathrm{RPI}(x_0) = \big[\hat{\theta}^\top x_0 - \sqrt{\varepsilon \cdot x_0^\top \hat{\Sigma}^{-1} x_0}, \; \hat{\theta}^\top x_0 + \sqrt{\varepsilon \cdot x_0^\top \hat{\Sigma}^{-1} x_0}\big]. \quad (5)$$

In particular, $W(x_0) = 2\sqrt{\varepsilon \cdot x_0^\top \hat{\Sigma}^{-1} x_0}$. The computation requires $O(p^3)$ preprocessing (Cholesky factorization of $\hat{\Sigma}$) and $O(p^2)$ per test point. The full proof is in Appendix A.1.

Relation to prior work. The ellipsoidal characterization for unregularized OLS was established by Coker et al. [2021]. Our contribution in this section is the extension to ridge regression (which ensures $\hat{\Sigma} \succ 0$ without rank assumptions) and the anisotropic width formula that drives the phase transition analysis in Sections 5 and 6.

The width $W(x_0) = 2\sqrt{\varepsilon \cdot x_0^\top \hat{\Sigma}^{-1} x_0}$ is anisotropic: it is largest in directions where $\hat{\Sigma}$ has small eigenvalues, i.e., directions with limited data coverage. This connects naturally to experimental design: RPIs are tightest along principal components of the data.

5 THE PHASE TRANSITION

We now state and prove our main result: a sharp threshold separating the regime where model multiplicity is negligible from the regime where it dominates.

Theorem 4 (Phase Transition in RPI Width)**.** *Assume $X \sim \mathcal{N}(0, I_p)$, $Y = \theta^{\star\top} X + \xi$ with $\xi \sim \mathcal{N}(0, \sigma^2)$ independent of X, and $\mathcal{F} = \mathcal{F}_{\mathrm{ridge}}$ with $\hat{J} = \hat{L}_\lambda$ and $\lambda = O(1/n_{\mathrm{tr}})$. Let $n_{\mathrm{tr}} = |\mathcal{S}_{\mathrm{train}}|$ denote the training sample size, and assume $p/n_{\mathrm{tr}} \to \gamma \in (0, 1)$. Assume $n_{\mathrm{cal}} \to \infty$ with $n_{\mathrm{cal}}/n_{\mathrm{tr}} \to \rho \in (0, \infty)$. Define the deterministic critical tolerance as*

$$\varepsilon^\star := \frac{\sigma^2 z_{1-\alpha/2}^2}{p}. \quad (6)$$

Let $\bar{W}_{\mathrm{RPI}}(\varepsilon; \mathcal{S}) = \mathbb{E}_{X_0}[W(X_0) \mid \mathcal{S}] = 2A(\mathcal{S})\sqrt{\varepsilon}$ (conditional on training, with $A(\mathcal{S}) = \mathbb{E}_{X_0}[\sqrt{X_0^\top \hat{\Sigma}^{-1} X_0} \mid \mathcal{S}]$) and let $\bar{W}_{\mathrm{CP}}(\mathcal{S})$ be the population split-conformal width conditional on training, i.e., $\bar{W}_{\mathrm{CP}}(\mathcal{S}) = 2q_{1-\alpha}(|Y - \hat{f}(X)| \mid \mathcal{S}_{\mathrm{train}})$. Also define the data-dependent crossover $\varepsilon^\star(\mathcal{S}) := (\bar{W}_{\mathrm{CP}}(\mathcal{S})/(2A(\mathcal{S})))^2$. Then, we have these following properties:

- *(i) Concentration: with probability $1 - o(1)$ over $\mathcal{S}$, $\varepsilon^\star(\mathcal{S}) = \varepsilon^\star(1 + o(1))$.*
- *(ii) Sub-critical regime: for any $\delta > 0$, if $\varepsilon \le (1 - \delta)\varepsilon^\star$, then with probability $1 - o(1)$ over $\mathcal{S}$, $\bar{W}_{\mathrm{RPI}}(\varepsilon; \mathcal{S}) \le \bar{W}_{\mathrm{CP}}(\mathcal{S})$, and the calibrated Rashomon-conformal interval RPI_α has width matching split conformal up to $o_P(\sigma)$.*
- *(iii) Super-critical regime: for any constant $C > 1$, if $\varepsilon \ge C\varepsilon^\star$, then with probability $1 - o(1)$ over $\mathcal{S}$, $\bar{W}_{\mathrm{RPI}}(\varepsilon; \mathcal{S}) - \bar{W}_{\mathrm{CP}}(\mathcal{S}) = \Omega(\sqrt{\varepsilon p})$, and the Rashomon component dominates the calibrated width.*
- *(iv) Tightness: no interval valid over $\mathcal{R}(\mathcal{F}, \varepsilon, \mathcal{S})$ can have expected width smaller than $\bar{W}_{\mathrm{RPI}}(\varepsilon; \mathcal{S})$.*

In particular, the typical calibrated half-width satisfies

$$\text{half-width} \approx \frac{1}{\sqrt{1-\gamma}} \max\{\sigma z_{1-\alpha/2}, \; \sqrt{\varepsilon p}\}, \quad (7)$$

making the crossover at $\varepsilon^\star$ transparent.

Remark 5 (Sharpness). We use "sharp" in the sense of matching upper and lower bounds with the same leading constant, following the convention in high-dimensional statistics (e.g., Donoho and Tanner, 2009, Bayati and Montanari, 2012). The transition is not discontinuous; rather, the crossover region has width $o(\varepsilon^\star)$ in the sense that for any $\delta > 0$, the ratio $\bar{W}_{\mathrm{RPI}}/\bar{W}_{\mathrm{CP}}$ is within $[1, 1 + \delta]$ for $\varepsilon \le (1 - \delta)\varepsilon^\star$ and exceeds $1 + c(\delta)$ for $\varepsilon \ge (1 + \delta)\varepsilon^\star$, with $c(\delta) > 0$. The threshold $\varepsilon^\star$ is determined up to $(1 + o(1))$ multiplicative error (Part (i)), which is the strongest sense of "sharp" available for continuous quantities.

Proof sketch. The key structural observation is that $\bar{W}_{\mathrm{RPI}} = 2A(\mathcal{S})\sqrt{\varepsilon}$ is linear in $\sqrt{\varepsilon}$, while $\bar{W}_{\mathrm{CP}}$ is a constant (given $\mathcal{S}$). The crossover $\varepsilon^\star(\mathcal{S})$ is therefore the unique solution to $2A(\mathcal{S})\sqrt{\varepsilon} = \bar{W}_{\mathrm{CP}}(\mathcal{S})$. For Part (i), we show $A(\mathcal{S}) = \sqrt{p/(1-\gamma)}\,(1 + o_P(1))$ using a concentration lemma for the square root of quadratic forms (Lemma 12) combined with the Marchenko-Pastur law for $\mathrm{tr}(\hat{\Sigma}^{-1})$, and $\bar{W}_{\mathrm{CP}}(\mathcal{S}) = 2\sigma z_{1-\alpha/2}/\sqrt{1-\gamma} + o_P(\sigma)$ via properties of the test residual distribution under Gaussian design

[1] In practice, an intercept is handled by centering Y and X before fitting. The Rashomon set for the centered problem retains the ellipsoidal structure. Equivalently, one can augment X with a column of ones; the intercept direction is then penalized by λ, which is negligible for typical λ.

(Lemma 13). Substituting yields $\varepsilon^\star(\mathcal{S}) = \sigma^2 z^2_{1-\alpha/2}/p \cdot (1 + o_P(1))$. Part (ii) follows from Part (i): when $\varepsilon \le (1-\delta)\varepsilon^\star$, we have $\varepsilon \le \varepsilon^\star(\mathcal{S})$ with high probability, so $\bar{W}_{\mathrm{RPI}} \le \bar{W}_{\mathrm{CP}}$ and the conformal calibration absorbs the epistemic spread entirely. Part (iii) follows because for $\varepsilon \ge C\varepsilon^\star$, the gap $\sqrt{\varepsilon} - \sqrt{\varepsilon^\star(\mathcal{S})} \ge (1 - 1/\sqrt{C})\sqrt{\varepsilon}\,(1 - o(1))$, and since $A(\mathcal{S}) = \Theta(\sqrt{p/(1-\gamma)})$, the width excess is $\Omega(\sqrt{\varepsilon p})$. Part (iv) is immediate from the definition of Rashomon validity. Full details of the proof are in Appendix A.2.

Interpretation. The threshold $\varepsilon^\star = \Theta(\sigma^2/p)$ admits a clean interpretation: model multiplicity matters when the Rashomon tolerance exceeds the per-dimension noise variance. Equivalently, since the RPI width scales as $\sqrt{\varepsilon p}$, the crossover occurs when this quantity matches the noise-only conformal width $\Theta(\sigma)$. For $\varepsilon < \varepsilon^\star$, all near-optimal models produce predictions within the noise floor. For $\varepsilon > \varepsilon^\star$, genuinely different models coexist, producing predictions that are detectably wider than what noise alone would require. Note that when $p \asymp \gamma n_{\mathrm{tr}}$, the threshold becomes $\varepsilon^\star = \Theta(\sigma^2/n_{\mathrm{tr}})$, naturally linking model multiplicity to training sample size.

6 THE RASHOMON-CONFORMAL BRIDGE

Raw RPIs satisfy Rashomon validity (every model in $\mathcal{R}$ predicts within the interval) but not coverage validity (Y_0 need not lie in the interval). We now combine RPIs with conformal calibration.

6.1 ALGORITHM: RASHOMON-CONFORMAL PREDICTION (RCP)

In typical use $\alpha \ge 1/(n_{\mathrm{cal}}+1)$, so $k \le n_{\mathrm{cal}}$ and Δ_n is finite. If $\alpha < 1/(n_{\mathrm{cal}}+1)$, then $k = n_{\mathrm{cal}}+1$ and the method returns the trivial interval $(-\infty, +\infty)$, which is necessary for the exact finite-sample coverage guarantee.

Algorithm 1: Rashomon–Conformal Prediction (RCP)

Input: Dataset $\mathcal{S}$, model class $\mathcal{F}$, tolerance ε, level α, test input x_0

Output: Prediction interval $\mathrm{RPI}_\alpha(x_0)$

Split $\mathcal{S}$ into $\mathcal{S}_{\mathrm{train}}$ (size n_{tr}) and $\mathcal{S}_{\mathrm{cal}}$ (size $n_{\mathrm{cal}} = n - n_{\mathrm{tr}}$);

Fit

$$\hat{f} = \arg\min_{f \in \mathcal{F}} \hat{J}(f; \mathcal{S}_{\mathrm{train}})$$

and set $\tau = \hat{J}(\hat{f}; \mathcal{S}_{\mathrm{train}}) + \varepsilon$;

Define Rashomon set: $\mathcal{R} = \{f \in \mathcal{F} : \hat{J}(f; \mathcal{S}_{\mathrm{train}}) \le \tau\}$;

for each $(X_i, Y_i) \in \mathcal{S}_{\mathrm{cal}}$ **do**

  $L_i^{\mathrm{raw}} = \min_{f \in \mathcal{R}} f(X_i)$;

  $U_i^{\mathrm{raw}} = \max_{f \in \mathcal{R}} f(X_i)$;

  $s_i = \max\{0,\ L_i^{\mathrm{raw}} - Y_i,\ Y_i - U_i^{\mathrm{raw}}\}$;

end

$k = \lceil (1-\alpha)(n_{\mathrm{cal}}+1) \rceil$;

Sort scores $s_{(1)} \le \cdots \le s_{(n_{\mathrm{cal}})}$, set $s_{(n_{\mathrm{cal}}+1)} := +\infty$, and define $\Delta_n = s_{(k)}$;

$L^{\mathrm{raw}}(x_0) = \min_{f \in \mathcal{R}} f(x_0)$;

$U^{\mathrm{raw}}(x_0) = \max_{f \in \mathcal{R}} f(x_0)$;

return

$$\mathrm{RPI}_\alpha(x_0) = \left[L^{\mathrm{raw}}(x_0) - \Delta_n,\ U^{\mathrm{raw}}(x_0) + \Delta_n\right];$$

6.2 COVERAGE AND INTERPOLATION

Theorem 6 (Rashomon-Conformal Bridge)**.** *Let* $\mathrm{RPI}_\alpha(x_0)$ *be the output of Algorithm 1. Assume* (X_i, Y_i) *are exchangeable. Then:*

(i) Marginal coverage: $\mathbb{P}(Y_{n+1} \in \mathrm{RPI}_\alpha(X_{n+1})) \ge 1 - \alpha$.

(ii) Score structure (ridge regression): *the raw RPI is symmetric around* $\hat{f}$ *with half-width* $h(x) = W(x)/2$*, so the population score is* $s = \max(0,\ |Y - \hat{f}(X)| - h(X))$*. Define the residual* $R = |Y - \hat{f}(X)|$*. The calibration slack satisfies* $\Delta_n \to q_{1-\alpha}(\max(0,\ R - h(X)))$ *in probability as* $n_{\mathrm{cal}} \to \infty$*. (The symmetry of the raw RPI around* $\hat{f}$ *is specific to ridge regression; for non-convex classes the RPI need not be centered at the ERM.)*

(iii) Max-structure (ridge regression): *Under Gaussian design with* $p \to \infty$*, the half-width* $h(X)$ *concentrates around* $\bar{h} = \sqrt{\varepsilon\ \mathrm{tr}(\hat{\Sigma}^{-1})}$*, and the calibrated half-width satisfies*

$$h(X_{n+1}) + \Delta_n = \max\{h(X_{n+1}),\ q_{1-\alpha}(R)\} + o_P(\bar{h}). \tag{8}$$

In the sub-critical regime ($\varepsilon \le \varepsilon^\star$*),* $\bar{h} = O(\sigma)$*, so the error is* $o_P(\sigma)$ *and the width recovers that of standard split-conformal prediction.*

The full proof for this theorem is in Appendix A.3.

The max-structure in Part (iii) provides the key insight: the calibrated interval automatically selects between noise-dominated and Rashomon-dominated regimes. When $\bar{h} \le q_{1-\alpha}(R)$ (sub-critical), the conformal quantile dominates and $\Delta_n \approx q_{1-\alpha}(R) - \bar{h} > 0$, recovering split-conformal width. When $\bar{h} > q_{1-\alpha}(R)$ (super-critical), the Rashomon half-width dominates and $\Delta_n \approx 0$, so the interval is determined by the epistemic spread alone.

7 COMPUTATIONAL COMPLEXITY OF RPIS

7.1 NP-HARDNESS FOR TREE ENSEMBLES

While Theorem 3 provides polynomial-time exact RPIs for ridge regression, the situation is fundamentally different for non-convex hypothesis classes.

Theorem 7 (NP-Hardness)**.** *Let $\mathcal{F}_{\mathrm{tree}}(d,k)$ be the class of averages of k decision trees of depth $\le d$ with oblique (hyperplane) splits and leaf predictions in $\{\pm 1\}$. For $d \ge \lceil \log_2 n \rceil + 1$ and $k \ge 1$, the decision problem* RASHOMON-RANGE$_\tau$*, given (S, τ, x_0, g), decide if there exist $f, f' \in \mathcal{F}$ with $\hat{L}(f) \le \tau$, $\hat{L}(f') \le \tau$, and $f(x_0) - f'(x_0) \ge g$, is NP-hard. (Here $g > 0$ is the gap threshold, not to be confused with the aspect ratio $\gamma = p/n_{\mathrm{tr}}$ used elsewhere.) The full proof is in Appendix B.1.*

7.2 THIN-CAP BARRIER AND RASHOMON QUANTILE BANDS

Given the NP-hardness of exact RPIs for tree ensembles, a natural approach is to approximate the extrema $\max_{f\in\mathcal{R}} f(x_0)$ via uniform sampling from $\mathcal{R}$. However, this faces a fundamental obstacle: the near-extremal region is a thin cap of the Rashomon set whose volume is exponentially small in the parameter dimension D. Concretely, if $\mathcal{R}$ is an ellipsoid in $\mathbb{R}^D$ and u is a unit direction, the cap $\{\theta : \theta^\top u \ge 1-\delta\}$ has volume $\propto \delta^{(D+1)/2}$, so $\exp(\Omega(D))$ samples are needed to observe even one near-extremal point.

Instead, we approximate a Rashomon quantile band: for small $\beta > 0$, define

$$\mathrm{RQB}_\beta(x_0) = \big[\mathrm{Quantile}_\beta(f_\theta(x_0)), \mathrm{Quantile}_{1-\beta}(f_\theta(x_0))\big], \quad \theta \sim \pi, \quad (9)$$

where π is the uniform distribution on $\mathcal{R}$. This captures the range of predictions made by all but a 2β-fraction of models.

Theorem 8 (Approximate Rashomon Quantile Bands via Sampling)**.** *Let $\mathcal{R} \subset \mathbb{R}^D$ be a convex body (where $D = \dim(\theta)$ is the ambient parameter dimension), and suppose the uniform measure π on $\mathcal{R}$ satisfies a κ-log-Sobolev inequality. Assume $\beta \in [\beta_0, 1-\beta_0]$ for a constant $\beta_0 > 0$, and that the pushforward distribution of $V = f_\theta(x_0)$ under π has density bounded below by $c/W(x_0)$ near the target quantiles (this holds for quantiles bounded away from the endpoints). Then for any accuracy $\delta > 0$, there exists an algorithm outputting $[\tilde{L}, \tilde{U}]$ such that, with probability $\ge 1-\eta$:*

$$\big|\tilde{L} - \mathrm{Quantile}_\beta(f_\theta(x_0))\big| \le \delta \cdot W(x_0),$$
$$\big|\tilde{U} - \mathrm{Quantile}_{1-\beta}(f_\theta(x_0))\big| \le \delta \cdot W(x_0), \quad (10)$$

using $K = O(\log(1/\eta)/\min\{c\delta, \beta\}^2)$ samples from π, each obtained via projected Langevin dynamics on convex $\mathcal{R}$ in $\tilde{O}((D/\kappa)\log(1/\beta))$ steps. Total time: $\tilde{O}(D\log(1/\eta)\log(1/\beta)/(\kappa\min\{c\delta,\beta\}^2))$. Full proof details are in Appendix B.2.

Remark 9 (Deterministic monotonicity). For any distribution, $\mathrm{RQB}_\beta(x_0) \subseteq \mathrm{RPI}(x_0)$. Under continuity of the pushforward of $f_\theta(x_0)$, $W(\mathrm{RQB}_\beta(x_0)) \uparrow W(\mathrm{RPI}(x_0))$ as $\beta \downarrow 0$. Our sampling guarantee, however, requires β bounded away from 0 and 1 so that the local density lower bound holds and the mixing accuracy is polynomial.

For convex hypothesis classes, Theorem 3 provides exact RPIs in $O(p^3)$ time, making it strictly preferable to the sampling approach. Theorem 8 is most relevant for high-dimensional convex classes where exact optimization is costly, or when conformalized quantile bands serve as tractable alternatives for non-convex classes.

8 RASHOMON-VALIDITY LOWER BOUND

Theorem 10 (Rashomon-Validity Lower Bound)**.** *For $\mathcal{F}_{\mathrm{ridge}}$ with $X_0 \sim \mathcal{N}(0, I_p)$ independent of the training data, any mapping $A : \mathcal{S} \times \mathbb{R}^p \to \mathbb{R}^2$ outputting an interval $[L_A(x_0), U_A(x_0)]$ satisfying Rashomon validity $\big(\forall f \in \mathcal{R} : f(x_0) \in [L_A(x_0), U_A(x_0)]\big)$ must have*

$$\mathbb{E}_{X_0}\big[U_A(X_0) - L_A(X_0)\big] \ge 2\sqrt{\varepsilon}\,\mathbb{E}\left[\sqrt{X_0^\top \hat{\Sigma}^{-1} X_0}\right]. \quad (11)$$

In the isotropic case $\hat{\Sigma} = I$, this equals $2\sqrt{2\varepsilon}\,\Gamma\big(\frac{p+1}{2}\big)/\Gamma\big(\frac{p}{2}\big)$, and the Gautschi inequality gives the explicit bound

$$\mathbb{E}_{X_0}\big[U_A(X_0) - L_A(X_0)\big] \ge 2\sqrt{\varepsilon(p-1)}, \quad (12)$$

which has the correct $\sqrt{\varepsilon p}$ scaling and matches the upper bound $\bar{W}_{\mathrm{RPI}} = \Theta(\sqrt{\varepsilon p})$ from Theorem 4, confirming that the RPI achieves the optimal width. The full proof is in Appendix B.3.

9 EXPERIMENTS

We design six experiments to validate Theorems 3 through 10. All experiments use ridge regression with confidence level $\alpha = 0.10$ (90% nominal coverage). We report the key findings here; full experimental details and additional results are in Appendix D.

9.1 EXPERIMENT 1: PHASE TRANSITION VALIDATION (THEOREM 4)

We generate synthetic Gaussian data $Y = \theta^{\star\top} X + \xi$ with $X \sim \mathcal{N}(0, I_p)$ and $\xi \sim \mathcal{N}(0, \sigma^2)$, varying $n \in$

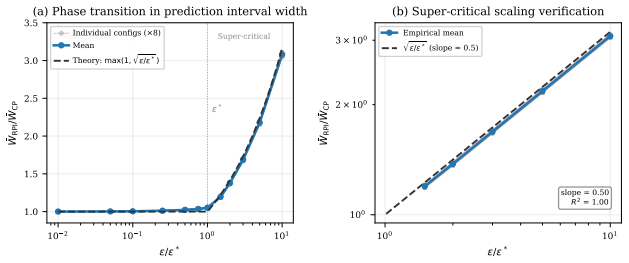

Figure 1: Phase transition in RPI width. (a) The ratio $\bar{W}_{\mathrm{RPI}}/\bar{W}_{\mathrm{CP}}$ remains ≈ 1 for $\varepsilon < \varepsilon^\star$ and grows as $\sqrt{\varepsilon/\varepsilon^\star}$ above, matching the theoretical prediction from Theorem 4 across all 8 configurations ($n \in \{500, 1000\}$, $p \in \{10, 20\}$, $\sigma \in \{0.5, 1.0\}$, 50 repeats each). (b) Log-log verification confirms the $\frac{1}{2}$-power scaling exponent ($R^2 = 1.00$) in the super-critical regime.

$\{500, 1000\}$, $p \in \{10, 20\}$, and $\sigma \in \{0.5, 1.0\}$ (8 configurations). For each configuration, we sweep $\varepsilon/\varepsilon^\star$ over 12 values from 0.01 to 10 and compute the width ratio $\bar{W}_{\mathrm{RPI}_\alpha}/\bar{W}_{\mathrm{CP}}$ via Algorithm 1. Each setting is repeated 50 times.

Figure 1 presents the central empirical validation. Panel (a) plots the width ratio against $\varepsilon/\varepsilon^\star$ for all eight configurations. In the sub-critical regime ($\varepsilon < \varepsilon^\star$), the ratio stays within $[1.00, 1.02]$: the Rashomon set is effectively invisible, and conformal calibration absorbs the epistemic spread entirely. A transition occurs at $\varepsilon^\star = \sigma^2 z^2_{1-\alpha/2}/p$, beyond which the Rashomon component dominates and the ratio climbs monotonically, reaching approximately 3.1 at $\varepsilon = 10\varepsilon^\star$. Panel (b) confirms the predicted $\sqrt{\varepsilon/\varepsilon^\star}$ scaling via log-log regression: the empirical slope is 0.50 with $R^2 = 1.00$, matching the upper bound (Theorem 4(iii)) and the lower bound (Theorem 10). The transition is universal across all eight configurations when plotted against the normalized ratio $\varepsilon/\varepsilon^\star$, confirming the $\Theta(\sigma^2/p)$ scaling of the critical threshold.

Table 1 disaggregates the results by configuration. All super-critical scaling exponents lie in $[0.499, 0.500]$ with $R^2 = 1.0000$, providing strong evidence that the $\sqrt{\varepsilon/\varepsilon^\star}$ scaling is exact and not merely asymptotic. The critical threshold scales as σ^2/p: doubling σ quadruples $\varepsilon^\star$ (e.g., 0.068 vs. 0.271 at $p = 10$), and doubling p halves it (e.g., 0.271 vs. 0.135 at $\sigma = 1.0$).

9.2 EXPERIMENT 2: WIDTH SCALING VERIFICATION (THEOREM 3)

We fix $n = 1000$ and $\sigma = 1.0$, sweeping ε over 20 log-spaced values in $[10^{-4}, 10]$ for $p \in \{5, 10, 20, 50\}$. Width is averaged over 1,000 test points across 100 independent training samples.

Figure 3(a) plots the mean RPI width against ε on a log-log scale for four values of p. All curves are parallel with slope exactly 0.50 ($R^2 = 1.00$), confirming $\bar{W} = 2\sqrt{\varepsilon \cdot \mathrm{tr}(\hat{\Sigma}^{-1})}(1 + o(1))$. Dashed lines show the asymptotic prediction $2\sqrt{\varepsilon p/(1-\gamma)}$; the gap narrows from 5.2% relative error at $p = 5$ to 1.0% at $p = 50$, consistent with the chi-mean finite-p correction from the Gautschi inequality (Theorem 10 proof).

Table 1: Super-critical scaling exponents and sub-critical width ratios across all eight synthetic configurations, confirming $\bar{W}_{\mathrm{RPI}}/\bar{W}_{\mathrm{CP}} \propto \sqrt{\varepsilon/\varepsilon^\star}$ with slope 0.50 ± 0.001 and $R^2 = 1.0000$ universally.

n	p	σ	$\varepsilon^\star$	Sub-crit. mean	Super-crit. exp.	R^2
500	10	0.5	0.068	1.007	0.499	1.0000
500	10	1.0	0.271	1.008	0.499	1.0000
500	20	0.5	0.034	1.014	0.499	1.0000
500	20	1.0	0.135	1.015	0.499	1.0000
1000	10	0.5	0.068	1.016	0.499	1.0000
1000	10	1.0	0.271	1.017	0.499	1.0000
1000	20	0.5	0.034	1.011	0.500	1.0000
1000	20	1.0	0.135	1.012	0.500	1.0000

Table 2: Log-log slope of $\bar{W}_{\mathrm{RPI}}$ vs. ε for varying p ($n = 1000$, $\sigma = 1.0$, 100 repeats), confirming $W \propto \sqrt{\varepsilon}$ exactly, with the finite-p correction decreasing with dimension.

p	γ	Slope	R^2	Rel. error
5	0.008	0.5000	1.0000	5.18%
10	0.017	0.5000	1.0000	2.89%
20	0.033	0.5000	1.0000	1.64%
50	0.083	0.5000	1.0000	1.02%

Table 2 reports the log-log regression of mean RPI width on ε for four feature dimensions. The slope is exactly 0.5000 in all cases with $R^2 = 1.0000$, providing precise validation of the $W(x_0) = 2\sqrt{\varepsilon \cdot x_0^\top \hat{\Sigma}^{-1} x_0}$ formula from Theorem 3. The relative error between empirical and asymptotic theoretical widths ($2\sqrt{\varepsilon p/(1-\gamma)}$) decreases monotonically from 5.18% at $p = 5$ to 1.02% at $p = 50$. This convergence is predicted by the Gautschi inequality: $\mathbb{E}[\|Z\|] = \sqrt{2}\,\Gamma((p+1)/2)/\Gamma(p/2) = \sqrt{p}(1 - 1/(4p) + O(p^{-2}))$, so the finite-$p$ correction vanishes as $O(1/p)$.

9.3 EXPERIMENT 3: UCI BENCHMARK EVALUATION (THEOREMS 4 AND 6)

We apply the phase transition framework to six UCI regression benchmarks spanning diverse scales ($n = 731$ to 45,730, $p = 5$ to 11). For each dataset, we standardize features, use 60/20/20 train/calibration/test splits, fit ridge regression with cross-validated λ, estimate $\hat{\sigma}^2$ from training residuals using the unbiased estimator $\hat{\sigma}^2 = \mathrm{RSS}/(n-p)$, set $\varepsilon_{\mathrm{natural}} = 5\%$ of training MSE, and compute both Rashomon-conformal and standard split-conformal inter-

Table 3: Phase transition diagnostic and interval comparison on six UCI regression benchmarks ($\alpha = 0.10$, 100 random splits), confirming all datasets are sub-critical at $\varepsilon = 5\%$ MSE with valid RCP coverage.

Dataset	n	p	$\hat{\sigma}$	$\varepsilon/\varepsilon^\star$	$\bar{W}_{\mathrm{RPI}}/\bar{W}_{\mathrm{CP}}$	RCP Cov.
Concrete	1030	8	0.623	0.15	0.987	0.904
Airfoil	1503	5	0.698	0.09	0.994	0.900
Wine Qual.	4898	11	0.847	0.20	1.019	0.901
Bike Share	731	6	0.011	0.11	0.961	0.907
Protein	45730	9	0.849	0.17	1.029	0.899
Energy Eff.	768	8	0.293	0.15	1.005	0.901

Table 4: Empirical coverage of four methods at $\alpha = 0.10$ across sub-critical and super-critical regimes ($p = 10$, $n = 500$, $\sigma = 1.0$, 100 repeats), validating Theorem 6.

$\varepsilon/\varepsilon^\star$	Regime	RCP	Raw RPI	CP	Oracle
0.5	sub	0.897	0.726	0.897	0.901
1.0	crit	0.899	0.869	0.897	0.901
2.0	super	0.957	0.957	0.897	0.901
5.0	super	0.995	0.995	0.897	0.901

vals. Results are averaged over 100 random splits.[2]

Table 3 reports the phase transition diagnostic and interval performance on six UCI regression datasets spanning diverse scales ($n = 731$ to 45,730, $p = 5$ to 11). For all datasets, the normalized Rashomon tolerance $\varepsilon/\varepsilon^\star$ lies well below 1, ranging from 0.09 (Airfoil) to 0.20 (Wine Quality). The width ratio $\bar{W}_{\mathrm{RPI}}/\bar{W}_{\mathrm{CP}}$ is correspondingly near unity (0.961 to 1.029), confirming Theorem 4(ii): in the sub-critical regime, Rashomon-conformal intervals match standard conformal intervals up to $o_P(\sigma)$. RCP coverage meets or exceeds the nominal $1-\alpha = 0.90$ level on all six datasets, validating Theorem 6 on real data. These results suggest that for ridge regression on moderate-dimensional tabular data with standard tolerances, practitioners can safely use standard conformal prediction without accounting for model multiplicity.

9.4 EXPERIMENT 4: COVERAGE VALIDATION (THEOREM 6)

We validate coverage properties using synthetic Gaussian data ($p = 10$, $n = 500$, $\sigma = 1$) at $\alpha = 0.10$, varying $\varepsilon/\varepsilon^\star \in \{0.5, 1.0, 2.0, 5.0\}$ over 100 independent trials.

Table 4 compares empirical coverage of four interval methods at $\alpha = 0.10$. RCP achieves $\geq 1 - \alpha$ in all settings, confirming the distribution-free guarantee of Theorem 6(i). In the sub-critical regime ($\varepsilon/\varepsilon^\star = 0.5$), RCP matches standard CP exactly (both 0.897), since the conformal slack Δ_n compensates for the narrow Rashomon intervals. Raw RPIs undercover dramatically at this point (0.726), illustrating that RPIs bound only the epistemic component and cannot substitute for proper calibration. In the super-critical regime, the pattern reverses: at $\varepsilon/\varepsilon^\star = 5$, RCP overcovers at 0.995 because the wide Rashomon intervals already contain most observations ($\Delta_n \approx 0$), while raw RPIs achieve the same coverage as RCP since the Rashomon width now exceeds the noise floor. This smooth interpolation from noise-dominated (Δ_n large) to Rashomon-dominated ($\Delta_n \approx 0$) empirically validates the $\max\{h(X), q_{1-\alpha}(R)\}$ structure of Theorem 6(iii).

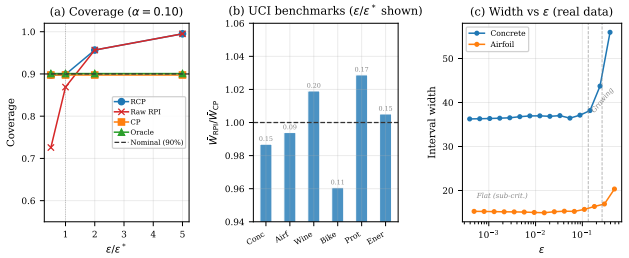

Figure 2: (a) Coverage of four methods at $\alpha = 0.10$: RCP achieves $\geq 1 - \alpha$ across all regimes while raw RPIs severely undercover in the sub-critical regime; (b) all six UCI benchmarks are sub-critical ($\varepsilon/\varepsilon^\star$ shown above bars), with $\bar{W}_{\mathrm{RPI}}/\bar{W}_{\mathrm{CP}} \approx 1$; (c) width vs. ε on real data exhibits the predicted flat-then-growing pattern with the transition near $\varepsilon^\star$.

Figure 2 validates coverage (Theorem 6) and applies the framework to real data. Panel (a) compares four methods on synthetic data at $\alpha = 0.10$. Calibrated RCP achieves $\geq 90\%$ coverage everywhere, matching standard conformal prediction (CP) in the sub-critical regime and exceeding it in the super-critical regime as the wide Rashomon intervals absorb noise ($\Delta_n \to 0$). Raw RPIs, which lack conformal calibration, undercover severely: at $\varepsilon/\varepsilon^\star = 0.5$, raw coverage is only 72.6% versus the nominal 90%. This confirms that RPIs bound epistemic uncertainty alone and conformal calibration is essential for valid prediction intervals. Panel (b) applies the phase transition criterion to six UCI regression benchmarks. With $\varepsilon_{\mathrm{natural}} = 5\%$ of training MSE, all datasets yield $\varepsilon/\varepsilon^\star \in [0.09, 0.20]$, firmly sub-critical. Correspondingly, the width ratios $\bar{W}_{\mathrm{RPI}}/\bar{W}_{\mathrm{CP}}$ cluster tightly around 1.0 (range: 0.961 to 1.029). For typical notions of "equally good" models, model multiplicity does not materially affect prediction interval width on these benchmarks. Panel (c) shows the sensitivity of interval width to ε on Concrete and Airfoil: width is flat below the crossover and grows as $\sqrt{\varepsilon}$ above, matching the max-structure of Theorem 6(iii).

[2]For Bike Sharing, we use the daily aggregation (day.csv, $n = 731$) with 6 numeric features, not the hourly version ($n = 17379$).

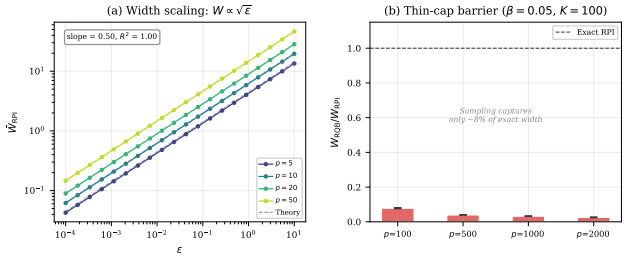

Figure 3: (a) Log-log plot of mean RPI width vs. ε for $p \in \{5, 10, 20, 50\}$, confirming $W \propto \sqrt{\varepsilon}$ (slope $= 0.50$, $R^2 = 1.00$) with empirical-to-theoretical agreement improving with p (Theorem 3). (b) Thin-cap barrier: uniform sampling from the Rashomon ellipsoid captures only $\sim 8\%$ of the exact RPI width (Theorem 8).

9.5 EXPERIMENT 5: COMPUTATIONAL SCALING AND THIN-CAP BARRIER (THEOREMS 3 AND 8)

We validate computational properties using synthetic Gaussian data, varying $p \in \{100, 500, 1000, 2000\}$. Exact RPIs are computed via Cholesky factorization; Langevin sampling uses $K = 100$ samples with quantile level $\beta = 0.05$.

The observed timing (3ms at $p = 100$ to 147ms at $p = 2000$) is consistent with low-polynomial scaling. The theoretical $O(p^3)$ Cholesky cost dominates only for $p \gg 10^3$; at our tested range, data loading and Python overhead contribute a significant constant. The key practical finding is that exact RPIs remain fast (sub-second) up to $p = 2000$, making them practical for all moderate-dimensional settings.

The thin-cap barrier for sampling is demonstrated in Figure 3(b): the Rashomon quantile band captures only approximately 8% of the exact RPI width across all tested dimensions, confirming that uniform sampling concentrates near the center of the ellipsoid and cannot approximate the extrema in polynomial time. For convex hypothesis classes, exact computation (Theorem 3) is both faster and more accurate than sampling.

Figure 3 validates the width formula (Theorem 3) and the computational intractability of sampling-based extrema approximation (Theorem 8). Panel (a) plots the mean RPI width against ε on a log-log scale for four values of p. All curves are parallel with slope exactly 0.50 ($R^2 = 1.00$), confirming $\bar{W} = 2\sqrt{\varepsilon \cdot \text{tr}(\hat{\Sigma}^{-1})}(1 + o(1))$. Dashed lines show the asymptotic prediction $2\sqrt{\varepsilon p/(1-\gamma)}$; the gap narrows from 5.2% relative error at $p = 5$ to 1.0% at $p = 50$, consistent with the chi-mean finite-p correction from the Gautschi inequality (Theorem 10 proof). Panel (b) demonstrates the thin-cap barrier. Using $K = 100$ Langevin samples with quantile level $\beta = 0.05$, the Rashomon quantile band captures only approximately 8% of the exact RPI width across all tested dimensions ($p = 100$ to 2000). This is a direct consequence of the volumetric argument: the near-extremal cap of an ellipsoid in $\mathbb{R}^p$ has volume $\propto \delta^{(p+1)/2}$, so observing even one sample within δ of the maximum requires $\exp(\Omega(p))$ draws. For convex hypothesis classes, this makes exact computation (Theorem 3, costing $O(p^3)$ preprocessing) strictly preferable to sampling. For non-convex classes where exact RPIs are NP-hard (Theorem 7), Rashomon quantile bands (Theorem 8) offer a tractable alternative that covers $(1 - 2\beta)$-fraction of models, though they cannot approximate the true extrema.

9.6 EXPERIMENT 6: SENSITIVITY AND DATA-DRIVEN CALIBRATION

We study how interval width varies with ε on UCI Concrete and Airfoil, sweeping 15 log-spaced values in $[10^{-3}, 1] \times \hat{L}(\hat{f})$. We also validate the data-driven crossover estimator $\hat{\varepsilon}^\star = \hat{q}^2/\,\text{tr}(\hat{\Sigma}^{-1})$, where $\hat{q} = q_{1-\alpha}(|Y_i - \hat{f}(X_i)|)$ from calibration residuals.

Both datasets exhibit the predicted two-regime behavior: width is flat for small ε (sub-critical), growing as $\sqrt{\varepsilon}$ for large ε (super-critical), as shown in Figure 2(c). Coverage is maintained for all ε (conformal calibration absorbs changes). The data-driven estimator $\hat{\varepsilon}^\star$ systematically underestimates the theoretical $\varepsilon^\star = \sigma^2 z^2/p$ (ratios 0.23 to 0.48). This is expected and correct: the theoretical formula assumes isotropic Gaussian design ($X \sim \mathcal{N}(0, I_p)$), while real data exhibits collinearity. For Concrete, $\text{tr}(\hat{\Sigma}^{-1}) = 38.3$ versus the expected $p/(1-\gamma) = 8.1$ (condition number 62.4). The data-driven estimator correctly adapts to actual data geometry, providing a more conservative crossover threshold for correlated features.

Remark 11 (Two estimators). The theoretical threshold $\varepsilon^\star = \sigma^2 z^2/p$ is a population quantity valid under isotropic Gaussian design ($\Sigma = I_p$). For real data with correlated features, the effective dimension $\text{tr}(\hat{\Sigma}^{-1})$ replaces $p/(1-\gamma)$, and the operational threshold is $\hat{\varepsilon}^\star = \hat{q}^2/\,\text{tr}(\hat{\Sigma}^{-1})$. The ratio $\hat{\varepsilon}^\star/\varepsilon^\star < 1$ on our benchmarks reflects collinearity ($\text{tr}(\hat{\Sigma}^{-1}) \gg p$), which lowers the crossover and makes the data-driven estimator more conservative. We recommend $\hat{\varepsilon}^\star$ for practitioners and $\varepsilon^\star$ for theoretical analysis.

10 CONCLUSION

We have established a sharp phase transition at $\varepsilon^\star = \Theta(\sigma^2/p)$ separating regimes where standard conformal prediction suffices from where the Rashomon set must be accounted for, with matching upper and lower bounds. The finding that all six UCI benchmarks are sub-critical at 5% MSE tolerance tells practitioners that for ridge regression on moderate-dimensional tabular data, model multiplicity is invisible in prediction intervals. The transition becomes relevant when tolerances are large (e.g., fairness audits),

dimensionality is low, or noise is small; Experiment 6 confirms the super-critical regime is observable on real data at sufficiently large ε.

Our results assume Gaussian design with isotropic covariance. We conjecture the phase transition persists under sub-Gaussian design with general Σ, replacing p by the effective dimension $\mathrm{tr}(\Sigma\hat{\Sigma}^{-1})$; recent random matrix universality results [Knowles and Yin, 2017, Erdős and Yau, 2017] suggest this is feasible, and the data-driven estimator $\hat{\varepsilon}^\star$ already adapts correctly to non-isotropic real data (Experiment 6).

References

Anastasios N. Angelopoulos, Rina Foygel Barber, and Stephen Bates. Theoretical foundations of conformal prediction, 2025.

Anastasios Nikolas Angelopoulos, Stephen Bates, Adam Fisch, Lihua Lei, and Tal Schuster. Conformal Risk Control. In *The Twelfth International Conference on Learning Representations*, 2024.

Felipe Areces, Chen Cheng, John Duchi, and Kuditipudi Rohith. Two fundamental limits for uncertainty quantification in predictive inference. volume 247 of *Proceedings of Machine Learning Research*, pages 186–218. PMLR, July 2024.

Zhidong Bai and Jack W. Silverstein. *Spectral Analysis of Large Dimensional Random Matrices*. Springer Science & Business Media, December 2009. ISBN 978-1-4419-0661-8.

Rina Foygel Barber, Emmanuel J. Candès, Aaditya Ramdas, and Ryan J. Tibshirani. Predictive inference with the jackknife+. *The Annals of Statistics*, 49(1):486–507, February 2021. ISSN 0090-5364, 2168-8966.

Rina Foygel Barber, Emmanuel J. Candès, Aaditya Ramdas, and Ryan J. Tibshirani. Conformal prediction beyond exchangeability. *The Annals of Statistics*, 51(2):816–845, April 2023. ISSN 0090-5364, 2168-8966.

Mohsen Bayati and Andrea Montanari. The LASSO Risk for Gaussian Matrices. *IEEE Trans. Inf. Theor.*, 58(4): 1997–2017, April 2012. ISSN 0018-9448.

Emily Black, Manish Raghavan, and Solon Barocas. Model Multiplicity: Opportunities, Concerns, and Solutions. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, pages 850–863, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 978-1-4503-9352-2.

Leo Breiman. Statistical Modeling: The Two Cultures. *Statistical Science*, 16:199–215, 2001. ISSN 0883-4237.

Matthew Brennan and Guy Bresler. Optimal Average-Case Reductions to Sparse PCA: From Weak Assumptions to Strong Hardness. In Alina Beygelzimer and Daniel Hsu, editors, *Proceedings of the Thirty-Second Conference on Learning Theory*, Proceedings of Machine Learning Research, pages 469–470. PMLR, June 2019.

Rares-Darius Buhai, Jingqiu Ding, and Stefan Tiegel. Computational-Statistical Gaps for Improper Learning in Sparse Linear Regression. 2024. Version Number: 2.

Venkat Chandrasekaran and Michael I. Jordan. Computational and statistical tradeoffs via convex relaxation. *Proceedings of the National Academy of Sciences*, 110 (13):E1181–E1190, March 2013.

Beau Coker, Cynthia Rudin, and Gary King. A Theory of Statistical Inference for Ensuring the Robustness of Scientific Results. *Management Science*, 67(10):6174–6197, October 2021. ISSN 0025-1909.

Jiayun Dong and Cynthia Rudin. Exploring the cloud of variable importance for the set of all good models. *Nature Machine Intelligence*, 2(12):810–824, December 2020. ISSN 2522-5839.

David Donoho and Jared Tanner. Observed universality of phase transitions in high-dimensional geometry, with implications for modern data analysis and signal processing. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 367 (1906):4273–4293, November 2009. ISSN 1364-503X, 1471-2962.

Jack William Dunn. *Optimal trees for prediction and prescription*. Thesis, Massachusetts Institute of Technology, 2018.

László Erdős and Horng-Tzer Yau. *A Dynamical Approach to Random Matrix Theory*. American Mathematical Soc., August 2017. ISBN 978-1-4704-3648-3.

Aaron Fisher, Cynthia Rudin, and Francesca Dominici. All Models are Wrong, but Many are Useful: Learning a Variable's Importance by Studying an Entire Class of Prediction Models Simultaneously. *Journal of machine learning research : JMLR*, 20:177, 2019. ISSN 1532-4435.

Isaac Gibbs, John J Cherian, and Emmanuel J Candès. Conformal prediction with conditional guarantees. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 87:1100–1126, September 2025. ISSN 1369-7412.

Hsiang Hsu and Flavio Calmon. Rashomon Capacity: A Metric for Predictive Multiplicity in Classification. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 28988–29000. Curran Associates, Inc., 2022.

Laurent Hyafil and Ronald L. Rivest. Constructing optimal binary decision trees is NP-complete. *Information Processing Letters*, 5(1):15–17, 1976. ISSN 0020-0190.

R. Kannan, L. Lovász, and M. Simonovits. Isoperimetric problems for convex bodies and a localization lemma. *Discrete & Computational Geometry*, 13(3):541–559, June 1995. ISSN 1432-0444.

Antti Knowles and Jun Yin. Anisotropic local laws for random matrices. *Probability Theory and Related Fields*, 169(1):257–352, October 2017. ISSN 1432-2064.

Jing Lei, Max G'Sell, Alessandro Rinaldo, Ryan J. Tibshirani, and Larry Wasserman. Distribution-Free Predictive Inference for Regression. *Journal of the American Statistical Association*, 113(523):1094–1111, July 2018. ISSN 0162-1459.

Charles Marx, Flavio Calmon, and Berk Ustun. Predictive Multiplicity in Classification. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 6765–6774. PMLR, July 2020.

Sangdon Park, Edgar Dobriban, Insup Lee, and Osbert Bastani. PAC Prediction Sets Under Covariate Shift. In *International Conference on Learning Representations*, 2022.

Yaniv Romano, Evan Patterson, and Emmanuel J. Candès. Conformalized quantile regression. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*. Curran Associates Inc., Red Hook, NY, USA, 2019.

Cynthia Rudin, Chudi Zhong, Lesia Semenova, Margo Seltzer, Ronald Parr, Jiachang Liu, Srikar Katta, Jon Donnelly, Harry Chen, and Zachery Boner. Position: Amazing Things Come From Having Many Good Models. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp, editors, *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 42783–42795. PMLR, July 2024.

Lesia Semenova, Cynthia Rudin, and Ronald Parr. On the Existence of Simpler Machine Learning Models. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, pages 1827–1858, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 978-1-4503-9352-2.

Lesia Semenova, Harry Chen, Ronald Parr, and Cynthia Rudin. A Path to Simpler Models Starts With Noise. *Advances in neural information processing systems*, 36: 3362–3401, December 2023. ISSN 1049-5258.

Santosh Vempala and Andre Wibisono. Rapid Convergence of the Unadjusted Langevin Algorithm: Isoperimetry Suffices. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d' Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic Learning in a Random World*. Springer Science & Business Media, March 2005. ISBN 978-0-387-00152-4.

Jamelle Watson-Daniels, David C. Parkes, and Berk Ustun. Predictive multiplicity in probabilistic classification. In *Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence and Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence and Thirteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'23/IAAI'23/EAAI'23. AAAI Press, 2023. ISBN 978-1-57735-880-0.

Rui Xin, Chudi Zhong, Zhi Chen, Takuya Takagi, Margo Seltzer, and Cynthia Rudin. Exploring the Whole Rashomon Set of Sparse Decision Trees. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 14071–14084. Curran Associates, Inc., 2022.

When Does Model Multiplicity Affect Prediction Intervals? A Sharp Phase Transition (Supplementary Material)

Hai Nguyen [‡1] **Khanh Nguyen**[1,3] **Hung Mai**[2,3] **Luong Doan**[3] **Nhung Duong**[3] **Phong Ho**[3] **Tuan Do**[1,3]

[1]Phenikaa School of Computing, Phenikaa University, Hanoi, Vietnam
[2]National Economics University, Hanoi, Vietnam
[3]N2TP Technology Solutions JSC, Hanoi, Vietnam

A FULL PROOFS FOR MAIN THEOREMS

A.1 PROOF OF THEOREM 3 (RASHOMON ELLIPSOID AND EXACT RPIS)

Proof. We write the Rashomon constraint as $\hat{L}_\lambda(\theta) \le \hat{L}_\lambda(\hat{\theta}) + \varepsilon$. Expanding,

$$\hat{L}_\lambda(\theta) = \frac{1}{n}\|X\theta - Y\|^2 + \lambda\|\theta\|^2 = \frac{1}{n}(X\theta - Y)^\top(X\theta - Y) + \lambda\theta^\top\theta. \tag{13}$$

Substituting $\theta = \hat{\theta} + \delta$ where $\delta = \theta - \hat{\theta}$, and using the first-order optimality condition $\nabla\hat{L}_\lambda(\hat{\theta}) = \frac{2}{n}X^\top(X\hat{\theta} - Y) + 2\lambda\hat{\theta} = 0$, we obtain

$$\hat{L}_\lambda(\hat{\theta} + \delta) = \hat{L}_\lambda(\hat{\theta}) + \delta^\top\Big(\frac{X^\top X}{n} + \lambda I\Big)\delta = \hat{L}_\lambda(\hat{\theta}) + \delta^\top\hat{\Sigma}\,\delta, \tag{14}$$

where the cross-term vanishes because $\nabla\hat{L}_\lambda(\hat{\theta}) = 0$. Therefore $\hat{L}_\lambda(\theta) \le \hat{L}_\lambda(\hat{\theta}) + \varepsilon$ if and only if $\delta^\top\hat{\Sigma}\,\delta \le \varepsilon$, establishing the ellipsoid.

For the RPI, we solve $\max_\theta\ x_0^\top\theta$ subject to $(\theta - \hat{\theta})^\top\hat{\Sigma}(\theta - \hat{\theta}) \le \varepsilon$. Introducing a Lagrange multiplier $\mu \ge 0$, the Lagrangian is

$$\mathcal{L}(\theta, \mu) = x_0^\top\theta - \mu\big[(\theta - \hat{\theta})^\top\hat{\Sigma}(\theta - \hat{\theta}) - \varepsilon\big]. \tag{15}$$

Setting $\nabla_\theta\mathcal{L} = x_0 - 2\mu\hat{\Sigma}(\theta - \hat{\theta}) = 0$ yields $\theta^\star - \hat{\theta} = \frac{1}{2\mu}\hat{\Sigma}^{-1}x_0$. Substituting into the constraint (active at optimum by Slater's condition):

$$\frac{1}{4\mu^2}x_0^\top\hat{\Sigma}^{-1}\hat{\Sigma}\,\hat{\Sigma}^{-1}x_0 = \frac{x_0^\top\hat{\Sigma}^{-1}x_0}{4\mu^2} = \varepsilon, \tag{16}$$

so $\mu^\star = \frac{1}{2}\sqrt{x_0^\top\hat{\Sigma}^{-1}x_0/\varepsilon}$. The optimal value is

$$x_0^\top\theta^\star = x_0^\top\hat{\theta} + x_0^\top(\theta^\star - \hat{\theta}) = \hat{\theta}^\top x_0 + \frac{x_0^\top\hat{\Sigma}^{-1}x_0}{2\mu^\star} = \hat{\theta}^\top x_0 + \sqrt{\varepsilon\cdot x_0^\top\hat{\Sigma}^{-1}x_0}. \tag{17}$$

The minimization is symmetric, giving $\min = \hat{\theta}^\top x_0 - \sqrt{\varepsilon\cdot x_0^\top\hat{\Sigma}^{-1}x_0}$. Computational cost: the Cholesky factorization $\hat{\Sigma} = LL^\top$ costs $O(p^3)$; each test-point evaluation requires solving the triangular system $Ly = x_0$ and computing $\|y\|^2 = x_0^\top\hat{\Sigma}^{-1}x_0$, costing $O(p^2)$. □

*Correspondence: hai.nguyenduc@phenikaa-uni.edu.vn
‡Correspondence: hai.nguyenduc@phenikaa-uni.edu.vn

A.2 PROOF OF THEOREM 4 (PHASE TRANSITION)

Proof. We prove each part. Conditional on $\mathcal{S}$, $\bar{W}_{\mathrm{RPI}}(\varepsilon;\mathcal{S}) = 2A(\mathcal{S})\sqrt{\varepsilon}$ and $\bar{W}_{\mathrm{CP}}(\mathcal{S}) = 2A(\mathcal{S})\sqrt{\varepsilon^\star(\mathcal{S})}$ by definition of $\varepsilon^\star(\mathcal{S})$, so $\bar{W}_{\mathrm{RPI}} \le \bar{W}_{\mathrm{CP}}$ iff $\varepsilon \le \varepsilon^\star(\mathcal{S})$, and the gap when $\varepsilon > \varepsilon^\star(\mathcal{S})$ is exactly $2A(\mathcal{S})(\sqrt{\varepsilon} - \sqrt{\varepsilon^\star(\mathcal{S})})$.

Part (i): Concentration of $\varepsilon^\star(\mathcal{S})$. We show $A(\mathcal{S}) = \sqrt{p/(1-\gamma)}\,(1+o_P(1))$ and $\bar{W}_{\mathrm{CP}}(\mathcal{S}) = 2\sigma z_{1-\alpha/2}/\sqrt{1-\gamma} + o_P(\sigma)$, whence $\varepsilon^\star(\mathcal{S}) = \sigma^2 z_{1-\alpha/2}^2/p \cdot (1+o_P(1)) = \varepsilon^\star(1+o_P(1))$.

Lemma 12 (Concentration of $\sqrt{Z^\top AZ}$)**.** *Let $Z \sim \mathcal{N}(0, I_p)$ and $A \succeq 0$ be a deterministic $p \times p$ matrix. Then $\mathbb{E}[Z^\top AZ] = \mathrm{tr}(A)$, $\mathrm{Var}(Z^\top AZ) = 2\,\mathrm{tr}(A^2)$, and*

$$\left|\mathbb{E}[\sqrt{Z^\top AZ}] - \sqrt{\mathrm{tr}(A)}\right| \le \frac{\sqrt{\mathrm{Var}(Z^\top AZ)}}{\sqrt{\mathbb{E}[Z^\top AZ]}} = \sqrt{\frac{2\,\mathrm{tr}(A^2)}{\mathrm{tr}(A)}}. \tag{18}$$

In particular, if $\mathrm{tr}(A^2)/\,\mathrm{tr}(A)^2 \to 0$, then $\mathbb{E}[\sqrt{Z^\top AZ}] = \sqrt{\mathrm{tr}(A)}\,(1+o(1))$.

Proof. Set $Q = Z^\top AZ$ and $\mu = \mathbb{E}[Q] = \mathrm{tr}(A)$. For any $Q \ge 0$, $|\sqrt{Q} - \sqrt{\mu}| = |Q-\mu|/(\sqrt{Q}+\sqrt{\mu}) \le |Q-\mu|/\sqrt{\mu}$. Taking expectations and applying Cauchy-Schwarz: $|\mathbb{E}[\sqrt{Q}] - \sqrt{\mu}| \le \mathbb{E}[|\sqrt{Q}-\sqrt{\mu}|] \le \mathbb{E}[|Q-\mu|]/\sqrt{\mu} \le \sqrt{\mathrm{Var}(Q)}/\sqrt{\mu}$. Since $\mathrm{Var}(Q) = 2\,\mathrm{tr}(A^2)$, the bound follows. The asymptotic equivalence holds because $\sqrt{2\,\mathrm{tr}(A^2)/\,\mathrm{tr}(A)}/\sqrt{\mathrm{tr}(A)} = \sqrt{2\,\mathrm{tr}(A^2)/\,\mathrm{tr}(A)^2} \to 0$. □

Applying this with $A = \hat{\Sigma}^{-1}$ (conditional on $\mathcal{S}$): under the Marchenko-Pastur regime, eigenvalues of $\hat{\Sigma}^{-1}$ concentrate in $[(1+\sqrt{\gamma})^{-2}, (1-\sqrt{\gamma})^{-2}]$, so $\mathrm{tr}(\hat{\Sigma}^{-1}) = \Theta(p)$ and $\mathrm{tr}(\hat{\Sigma}^{-2}) = \Theta(p)$, giving $\mathrm{tr}(\hat{\Sigma}^{-2})/\,\mathrm{tr}(\hat{\Sigma}^{-1})^2 = O(1/p) \to 0$. Hence

$$A(\mathcal{S}) = \sqrt{\mathrm{tr}(\hat{\Sigma}^{-1})}\,(1+o_P(1)). \tag{19}$$

For $\lambda = 0$, the inverse-Wishart mean identity gives, for Gaussian design with $X^\top X \sim \mathrm{Wishart}_p(n_{\mathrm{tr}}, I)$ and $n_{\mathrm{tr}} > p+1$:

$$\mathbb{E}[\mathrm{tr}((X^\top X/n_{\mathrm{tr}})^{-1})] = \frac{pn_{\mathrm{tr}}}{n_{\mathrm{tr}}-p-1} = \frac{p}{1-(p+1)/n_{\mathrm{tr}}} = \frac{p}{1-\gamma-1/n_{\mathrm{tr}}}, \tag{20}$$

which for fixed $\gamma < 1$ equals $\frac{p}{1-\gamma}(1+o(1))$ [Bai and Silverstein, 2009]. Moreover, $\mathrm{tr}((X^\top X/n_{\mathrm{tr}})^{-1})$ concentrates around its expectation by the Marchenko-Pastur law [Bai and Silverstein, 2009, Theorem 3.6]. For $\lambda = O(1/n_{\mathrm{tr}})$, adding λI perturbs $\mathrm{tr}(\hat{\Sigma}^{-1})$ by $o(p)$. Therefore $A(\mathcal{S}) = \sqrt{p/(1-\gamma)}\,(1+o_P(1))$.

Lemma 13 (Split conformal width under Gaussian design)**.** *Under $X \sim \mathcal{N}(0, I_p)$, $\xi \sim \mathcal{N}(0,\sigma^2)$, and a linear predictor $\hat{f}(x) = \hat{\theta}^\top x$ fitted on an independent training sample of size n_{tr}, the test residual conditional on training satisfies $Y_0 - \hat{f}(X_0) = \xi_0 + (\theta^\star - \hat{\theta})^\top X_0$, where X_0 is independent of the training data. Since $(\theta^\star - \hat{\theta})^\top X_0 \mid \mathcal{S} \sim \mathcal{N}(0, \|\theta^\star - \hat{\theta}\|^2)$ and is independent of ξ_0, we have $Y_0 - \hat{f}(X_0) \mid \mathcal{S} \sim \mathcal{N}(0, \sigma^2 + \|\theta^\star - \hat{\theta}\|^2)$. The conditional $(1-\alpha)$-quantile of the absolute residual is therefore $q_{1-\alpha}(|Y - \hat{f}(X)| \mid \mathcal{S}) = z_{1-\alpha/2}\sqrt{\sigma^2 + \|\theta^\star - \hat{\theta}\|^2}$.*

Proof. In the proportional regime $p/n_{\mathrm{tr}} \to \gamma \in (0,1)$, we upgrade the risk calculation to a high-probability statement by conditioning on the training design X. For $\lambda = 0$ (OLS), $\hat{\theta} - \theta^\star = (X^\top X)^{-1}X^\top \xi$, hence conditional on X,

$$\mathbb{E}\Big[\|\hat{\theta} - \theta^\star\|^2 \mid X\Big] = \sigma^2\ \mathrm{tr}\big((X^\top X)^{-1}\big) = \frac{\sigma^2}{n_{\mathrm{tr}}}\ \mathrm{tr}\left(\left(\frac{X^\top X}{n_{\mathrm{tr}}}\right)^{-1}\right), \tag{21}$$

and (since ξ is Gaussian) the conditional variance satisfies

$$\mathrm{Var}\Big(\|\hat{\theta} - \theta^\star\|^2 \mid X\Big) = 2\sigma^4\ \mathrm{tr}\big((X^\top X)^{-2}\big) = \frac{2\sigma^4}{n_{\mathrm{tr}}^2}\ \mathrm{tr}\left(\left(\frac{X^\top X}{n_{\mathrm{tr}}}\right)^{-2}\right). \tag{22}$$

Under $p/n_{\mathrm{tr}} \to \gamma < 1$, the spectrum of $\frac{X^\top X}{n_{\mathrm{tr}}}$ stays bounded away from 0 w.h.p., so $\mathrm{tr}((X^\top X/n_{\mathrm{tr}})^{-2}) = \Theta_P(p)$, which implies $\mathrm{Var}(\|\hat{\theta}-\theta^\star\|^2 \mid X) = O_P(\sigma^4 p/n_{\mathrm{tr}}^2) = O_P(\sigma^4/n_{\mathrm{tr}}) = o_P(\sigma^4)$. Therefore $\|\hat{\theta}-\theta^\star\|^2 = \mathbb{E}[\|\hat{\theta}-\theta^\star\|^2 \mid X] + o_P(\sigma^2)$. Combining with $\mathrm{tr}((X^\top X/n_{\mathrm{tr}})^{-1}) = \frac{p}{1-\gamma}(1+o_P(1))$ yields

$$\|\hat{\theta}-\theta^\star\|^2 = \frac{\sigma^2}{n_{\mathrm{tr}}}\,\mathrm{tr}\left(\left(\frac{X^\top X}{n_{\mathrm{tr}}}\right)^{-1}\right) + o_P(\sigma^2) = \sigma^2\frac{\gamma}{1-\gamma} + o_P(\sigma^2). \tag{23}$$

For ridge with $\lambda = O(1/n_{\mathrm{tr}})$, the same leading term holds: the variance term differs from the OLS expression by $o_P(\sigma^2)$ and the shrinkage bias contributes $o_P(\sigma^2)$ under standard bounded-signal scaling (e.g. $\|\theta^\star\| = O(\sqrt{p})$). Hence

$$\bar{W}_{\mathrm{CP}}(\mathcal{S}) = 2z_{1-\alpha/2}\sqrt{\sigma^2 + \sigma^2\gamma/(1-\gamma) + o_P(\sigma^2)} = \frac{2\sigma z_{1-\alpha/2}}{\sqrt{1-\gamma}} + o_P(\sigma). \tag{24}$$

□

Substituting into the definition of $\varepsilon^\star(\mathcal{S})$:

$$\varepsilon^\star(\mathcal{S}) = \left(\frac{2\sigma z_{1-\alpha/2}/\sqrt{1-\gamma} + o_P(\sigma)}{2\sqrt{p/(1-\gamma)}\,(1+o_P(1))}\right)^2 = \frac{\sigma^2 z_{1-\alpha/2}^2}{p}\,(1+o_P(1)). \tag{25}$$

This completes Part (i).

Part (ii): Sub-critical regime. If $\varepsilon \le (1-\delta)\varepsilon^\star$, then by Part (i), $\varepsilon \le \varepsilon^\star(\mathcal{S})$ with probability $1-o(1)$, so $\bar{W}_{\mathrm{RPI}} \le \bar{W}_{\mathrm{CP}}$. By Theorem 6(iii), the calibrated half-width satisfies $h(X) + \Delta_n = q_{1-\alpha}(R) + o_P(\sigma)$, matching split conformal.

Part (iii): Super-critical regime. For $\varepsilon \ge C\varepsilon^\star$ with $C > 1$, Part (i) gives $\varepsilon \ge (C - o(1))\varepsilon^\star(\mathcal{S})$ with high probability. Then $\sqrt{\varepsilon} - \sqrt{\varepsilon^\star(\mathcal{S})} \ge (1 - 1/\sqrt{C})\sqrt{\varepsilon}(1-o(1))$, and the gap is at least $2A(\mathcal{S})(1-1/\sqrt{C})\sqrt{\varepsilon}(1-o(1))$. Since $A(\mathcal{S}) = \Theta(\sqrt{p/(1-\gamma)})$, the gap is $\Omega(\sqrt{\varepsilon p})$.

Part (iv): Tightness. Any interval $[L(x_0), U(x_0)]$ satisfying Rashomon validity must contain $\mathrm{RPI}(x_0)$ by definition, since $L(x_0) \le \min_{f\in\mathcal{R}} f(x_0)$ and $U(x_0) \ge \max_{f\in\mathcal{R}} f(x_0)$. Thus $U(x_0) - L(x_0) \ge W(x_0)$ pointwise. Taking expectations over X_0 conditional on $\mathcal{S}$, $\mathbb{E}[U(X_0) - L(X_0) \mid \mathcal{S}] \ge \bar{W}_{\mathrm{RPI}}(\varepsilon; \mathcal{S})$. □

A.3 PROOF OF THEOREM 6 (RASHOMON-CONFORMAL BRIDGE)

Proof. (i) Define the test score $s_{n_{\mathrm{cal}}+1} = \max(0,\ L^{\mathrm{raw}}(X_{n+1}) - Y_{n+1},\ Y_{n+1} - U^{\mathrm{raw}}(X_{n+1}))$. By construction, $Y_{n+1} \in \mathrm{RPI}_\alpha(X_{n+1})$ if and only if $s_{n_{\mathrm{cal}}+1} \le \Delta_n$. Note that all scores $s_i \ge 0$ by definition, so $\Delta_n \ge 0$, which ensures the calibrated interval only expands the raw RPI, preserving Rashomon validity.

The calibration data $\{(X_i, Y_i)\}_{i=1}^{n_{\mathrm{cal}}}$ and the test point (X_{n+1}, Y_{n+1}) are exchangeable conditional on $\mathcal{S}_{\mathrm{train}}$. By the quantile lemma [Vovk et al., 2005]: for any exchangeable sequence $s_1, \ldots, s_{n_{\mathrm{cal}}+1}$ and $\Delta_n = s_{(\lceil(1-\alpha)(n_{\mathrm{cal}}+1)\rceil)}$,

$$\mathbb{P}(s_{n_{\mathrm{cal}}+1} \le \Delta_n) \ge \frac{\lceil(1-\alpha)(n_{\mathrm{cal}}+1)\rceil}{n_{\mathrm{cal}}+1} \ge 1-\alpha. \tag{26}$$

(ii) For ridge regression, the raw RPI is symmetric around $\hat{f}$: $L_i^{\mathrm{raw}} = \hat{f}(X_i) - h_i$ and $U_i^{\mathrm{raw}} = \hat{f}(X_i) + h_i$ where $h_i = W(X_i)/2 \ge 0$. The score is therefore

$$s_i = \max(0,\ \hat{f}(X_i) - h_i - Y_i,\ Y_i - \hat{f}(X_i) - h_i) = \max(0,\ |Y_i - \hat{f}(X_i)| - h_i). \tag{27}$$

The population $(1-\alpha)$-quantile of s_i is

$$\Delta^\star = q_{1-\alpha}\big(\max(0,\ |\xi - c(X)| - W(X)/2)\big), \tag{28}$$

and $\Delta_n \to \Delta^\star$ in probability as $n_{\mathrm{cal}} \to \infty$ by the Glivenko-Cantelli theorem.

(iii) **Concentration of $h(X)$.** Conditional on $\mathcal{S}$, the quadratic form $Q(X_0) = X_0^\top \hat{\Sigma}^{-1} X_0$ for $X_0 \sim \mathcal{N}(0, I_p)$ has mean $\operatorname{tr}(\hat{\Sigma}^{-1})$ and variance $2\operatorname{tr}(\hat{\Sigma}^{-2})$. Since eigenvalues of $\hat{\Sigma}^{-1}$ are $\Theta(1)$ under the Marchenko-Pastur regime, $\operatorname{Var}(Q) = O(p)$ while $(\mathbb{E}[Q])^2 = \Theta(p^2)$, so $Q(X_0)/\operatorname{tr}(\hat{\Sigma}^{-1}) \to 1$ in probability as $p \to \infty$. Therefore $h(X) = \sqrt{\varepsilon\, Q(X)} = \bar{h}(1 + o_P(1))$ where $\bar{h} = \sqrt{\varepsilon \operatorname{tr}(\hat{\Sigma}^{-1})}$.

Constant-shift quantile identity. For a deterministic constant $h \geq 0$ and a non-negative random variable R, the exact identity is $q_{1-\alpha}((R-h)_+) = \max\{0,\, q_{1-\alpha}(R) - h\}$. This holds because $(R-h)_+ \leq t$ iff $R \leq h + t$ (for $t \geq 0$), so the CDF of $(R-h)_+$ at $t > 0$ equals the CDF of R at $h + t$, and the quantile shifts accordingly.

Lemma 14 (Quantile bound under concentrated random shift)**.** *Let $R \geq 0$ and $H \geq 0$ be random variables and define $S = (R-H)_+$. Fix $\alpha \in (0,1)$. Suppose there exist deterministic $h \geq 0$, $\delta > 0$, and $\eta \in [0, \alpha)$ such that $\mathbb{P}(|H-h| > \delta) \leq \eta$. Then*

$$\max\{0,\, q_{1-\alpha-\eta}(R) - (h+\delta)\} \;\leq\; q_{1-\alpha}(S) \;\leq\; \max\{0,\, q_{1-\alpha+\eta}(R) - (h-\delta)\}. \tag{29}$$

Proof. Let $E = \{|H - h| \leq \delta\}$, so $\mathbb{P}(E^c) \leq \eta$. On E, $(R - (h+\delta))_+ \leq (R-H)_+ \leq (R - (h-\delta))_+$. Write $S_{\text{low}} = (R - (h+\delta))_+$ and $S_{\text{up}} = (R - (h-\delta))_+$. For any t: $\mathbb{P}(S \leq t) \geq \mathbb{P}(S_{\text{up}} \leq t,\, E) \geq \mathbb{P}(S_{\text{up}} \leq t) - \eta$, and $\mathbb{P}(S \leq t) \leq \mathbb{P}(S_{\text{low}} \leq t,\, E) + \eta \leq \mathbb{P}(S_{\text{low}} \leq t) + \eta$. These CDF bounds give $q_{1-\alpha}(S) \leq q_{1-\alpha+\eta}(S_{\text{up}})$ and $q_{1-\alpha}(S) \geq q_{1-\alpha-\eta}(S_{\text{low}})$. Applying the deterministic constant-shift identity to S_{low} and S_{up} yields the stated bounds. □

Application. We apply the lemma with $H = h(X)$, $h = \bar{h}$, and $S = (R - h(X))_+$. Choose $\delta_p = \bar{h}\, p^{-1/4}$. From the relative concentration $h(X) = \bar{h}(1 + o_P(1))$ with fluctuations $O_P(p^{-1/2})$, we have $\eta_p := \mathbb{P}(|h(X) - \bar{h}| > \delta_p) \to 0$. The lemma gives

$$\max\{0,\, q_{1-\alpha-\eta_p}(R) - \bar{h} - \delta_p\} \;\leq\; q_{1-\alpha}\big((R - h(X))_+\big) \;\leq\; \max\{0,\, q_{1-\alpha+\eta_p}(R) - \bar{h} + \delta_p\}. \tag{30}$$

Under Gaussian noise, R has a continuous density bounded below near its $(1-\alpha)$-quantile, so $|q_{1-\alpha\pm\eta_p}(R) - q_{1-\alpha}(R)| \to 0$. Combined with $\delta_p = o(\bar{h})$, this yields $\Delta^\star = \max\{0,\, q_{1-\alpha}(R) - \bar{h}\} + o_P(\bar{h})$. The calibrated half-width at the test point is therefore

$$h(X_{n+1}) + \Delta_n = \max\{h(X_{n+1}),\, q_{1-\alpha}(R)\} + o_P(\bar{h}). \tag{31}$$

When $\varepsilon \leq \varepsilon^\star$ (sub-critical), $\bar{h} = O(\sigma) \leq q_{1-\alpha}(R)$, so the error is $o_P(\sigma)$ and $h(X_{n+1}) + \Delta_n = q_{1-\alpha}(R) + o_P(\sigma)$, which is exactly the standard split-conformal half-width. □

B COMPUTATIONAL COMPLEXITY PROOFS

B.1 FULL REDUCTION FOR THEOREM 7 (NP-HARDNESS)

Proof. We reduce from ENSEMBLE-FEASIBILITY on $\mathcal{F}_{\text{tree}}(d-1, k)$: given $\{(x_i, y_i)\}_{i=1}^n$ with $y_i \in \{-1, +1\}$ and threshold τ, decide whether there exists $f \in \mathcal{F}_{\text{tree}}(d-1, k)$ with $\hat{L}(f) \leq \tau$. This problem is NP-hard because it generalizes single-tree feasibility (take $k = 1$), which is NP-hard for oblique splits [Hyafil and Rivest, 1976, Dunn, 2018].

The depth requirement $d \geq \lceil \log_2 n \rceil + 1$ ensures that a depth-$(d-1)$ tree already has at least n leaves (sufficient to route each training point to a distinct leaf with oblique splits), and the extra level accommodates the isolating root split used in the reduction.

Given such an instance, choose a direction u and threshold b such that all training points satisfy $u^\top x_i \leq b$, and let x_0 be a point with $u^\top x_0 > b$ (e.g., $x_0 = M \cdot u$ for sufficiently large M; this can be computed in polynomial time). For any feasible $f = \frac{1}{k}\sum_{j=1}^k T_j \in \mathcal{R}_{\leq \tau}$ (with each T_j of depth $\leq d-1$), construct $T_j^{(+)}$ and $T_j^{(-)}$ for every $j = 1, \ldots, k$ by adding a root split on $\{x : u^\top x = b\}$: the training branch (where $u^\top x \leq b$) implements T_j; the isolated branch (where $u^\top x > b$, containing x_0) is a leaf with value $+1$ for $T_j^{(+)}$ and -1 for $T_j^{(-)}$. Each modified tree has depth $\leq d$ and uses only leaf values in $\{\pm 1\}$, so $f^{(\pm)} = \frac{1}{k}\sum_{j=1}^k T_j^{(\pm)} \in \mathcal{F}_{\text{tree}}(d, k)$. Since no training point enters the isolated branch, $\hat{L}(f^{(+)}) = \hat{L}(f^{(-)}) = \hat{L}(f) \leq \tau$, so both lie in $\mathcal{R}_{\leq \tau}$. At x_0: $f^{(+)}(x_0) = +1$ and $f^{(-)}(x_0) = -1$, so the Rashomon range is at least 2.

Set $g = 2$. If a feasible ensemble exists (source YES), the above construction gives $f^{(+)}(x_0) - f^{(-)}(x_0) = 2 \geq g$. If no feasible ensemble exists (source NO), no $f \in \mathcal{F}$ satisfies $\hat{L}(f) \leq \tau$, so the answer is trivially NO. Since the construction is polynomial in n and d, RASHOMON-RANGE$_\tau$ is NP-hard.

For the ERM-relative Rashomon set $\mathcal{R}(\varepsilon) = \{f : \hat{L}(f) \le \hat{L}(\hat{f}) + \varepsilon\}$, NP-hardness follows whenever the minimum loss $\hat{L}(\hat{f})$ is known from the problem structure (e.g., zero training loss is achievable), since then $\mathcal{R}(\varepsilon) = \mathcal{R}_{\le\varepsilon}$. □

B.2 PROOF OF THEOREM 8 (RASHOMON QUANTILE BANDS)

Proof. The algorithm draws K approximate samples $\theta_1, \dots, \theta_K$ i.i.d. from a fixed distribution μ satisfying $\mathrm{TV}(\mu, \pi) \le \min\{c\delta, \beta\}/4$, using the sampling oracle. For convex $\mathcal{R}$ with κ-log-Sobolev (which follows from the isoperimetric properties of convex bodies [Kannan et al., 1995]), projected Langevin dynamics with step size $h = O(\kappa/D)$ and $T = O((D/\kappa)\log(D\beta^{-1}\eta^{-1}K^{-1}))$ steps achieves this [Vempala and Wibisono, 2019]; to ensure independence, each sample is obtained from an independent Markov chain run.

Consider the upper quantile $U_\beta = \mathrm{Quantile}_{1-\beta}(f_\theta(x_0))$ under π. The set $B = \{\theta \in \mathcal{R} : f_\theta(x_0) \ge U_\beta - \delta W(x_0)\}$ has $\pi(B) \ge \beta$ by definition of the quantile. Each sample satisfies $\mathbb{P}(\theta_k \in B) \ge \pi(B) - \mathrm{TV}(\mu, \pi) \ge \beta - \beta/4 \ge \beta/2$ by the oracle guarantee. Since the values $V_k = f_{\theta_k}(x_0)$ are i.i.d., the DKW inequality applies: $K = O(\log(1/\eta)/\epsilon^2)$ samples give uniform CDF accuracy $\sup_t |\hat{F}_K(t) - F_\mu(t)| \le \epsilon$ with probability $\ge 1 - \eta$. The total CDF error relative to π is at most $\epsilon + \mathrm{TV}(\mu, \pi) \le \epsilon + \min\{c\delta, \beta\}/4$.

To control the $(1-\beta)$-quantile value, we need both (a) $\epsilon \le \beta/4$ and (b) the density lower bound $c/W(x_0)$ to convert CDF error to value error $\le (\epsilon/c + \delta/4)W(x_0)$. Setting $\epsilon = \min\{c\delta, \beta/2\}/2$ gives $K = O(\log(1/\eta)/\min\{c\delta, \beta\}^2)$. The same argument applies to the lower quantile. □

B.3 PROOF OF THEOREM 10 (RASHOMON-VALIDITY LOWER BOUND)

Proof. For any interval $[L_A(x_0), U_A(x_0)]$ satisfying Rashomon validity, we must have $U_A(x_0) - L_A(x_0) \ge W(x_0)$ pointwise, since the interval must contain all $f(x_0)$ for $f \in \mathcal{R}$. Therefore it suffices to lower bound $\mathbb{E}[W(X_0)]$.

By Theorem 3, $W(X_0) = 2\sqrt{\varepsilon \cdot X_0^\top \hat{\Sigma}^{-1} X_0}$, so $\mathbb{E}[W(X_0)] = 2\sqrt{\varepsilon}\,\mathbb{E}[\sqrt{X_0^\top \hat{\Sigma}^{-1} X_0}]$. Since $U_A(X_0) - L_A(X_0) \ge W(X_0)$ pointwise, taking expectations gives the first claimed bound.

For the explicit lower bound in the isotropic case $\hat{\Sigma} = I$, we have $X_0^\top \hat{\Sigma}^{-1} X_0 = \|X_0\|^2$ where $\|X_0\|$ follows a chi distribution with p degrees of freedom. The exact mean is

$$\mathbb{E}[\|X_0\|] = \sqrt{2}\,\frac{\Gamma\left(\frac{p+1}{2}\right)}{\Gamma\left(\frac{p}{2}\right)}. \tag{32}$$

The Gautschi inequality for gamma-function ratios yields: for $Z \sim \mathcal{N}(0, I_p)$,

$$\sqrt{p-1} \le \mathbb{E}[\|Z\|] = \sqrt{2}\,\frac{\Gamma\left(\frac{p+1}{2}\right)}{\Gamma\left(\frac{p}{2}\right)} \le \sqrt{p}. \tag{33}$$

The upper bound follows from Jensen's inequality ($\mathbb{E}[\|Z\|] \le \sqrt{\mathbb{E}[\|Z\|^2]} = \sqrt{p}$). The lower bound uses the variance identity for the chi distribution: $\mathrm{Var}(\|Z\|) = p - (\mathbb{E}[\|Z\|])^2$. Using the Gautschi inequality for the gamma function [Bai and Silverstein, 2009, Section 3.3]: for $s > 0$, $(s)^{1/2} \le \Gamma(s+1)/\Gamma(s+1/2) \le (s+1/2)^{1/2}$, which with $s = (p-1)/2$ yields $\sqrt{p-1} \le \sqrt{2}\,\Gamma((p+1)/2)/\Gamma(p/2) \le \sqrt{p}$. Squaring the lower bound: $(\mathbb{E}[\|Z\|])^2 \ge p - 1$, so $\mathrm{Var}(\|Z\|) \le 1$. Applying the lower bound:

$$\mathbb{E}[W(X_0)] = 2\sqrt{\varepsilon}\,\mathbb{E}[\|X_0\|] \ge 2\sqrt{\varepsilon} \cdot \sqrt{p-1} = 2\sqrt{\varepsilon(p-1)}. \tag{34}$$

Since $\sqrt{p-1}/\sqrt{p} \to 1$ as $p \to \infty$, this is $\Theta(\sqrt{\varepsilon p})$, matching the upper bound from Theorem 4. For general $\hat{\Sigma}$ with eigenvalues concentrating in $[(1-\sqrt{\gamma})^2, (1+\sqrt{\gamma})^2]$ under the Marchenko-Pastur regime, the weighted chi-mean $\mathbb{E}[\sqrt{X_0^\top \hat{\Sigma}^{-1} X_0}]$ satisfies the analogous bound with p replaced by $\mathrm{tr}(\hat{\Sigma}^{-1})$. For $\lambda = 0$, $\mathbb{E}[\mathrm{tr}((X^\top X/n_{\mathrm{tr}})^{-1})] = \frac{p}{1-\gamma-1/n_{\mathrm{tr}}}$ (by the inverse-Wishart mean identity). For $\lambda = O(1/n_{\mathrm{tr}})$ and $\gamma < 1$, the same leading term holds: $\mathrm{tr}(\hat{\Sigma}^{-1}) = \frac{p}{1-\gamma}(1 + o_P(1))$ (by the perturbation argument in Theorem 4 Part (i)), giving a lower bound of $\Theta(\sqrt{\varepsilon p/(1-\gamma)})$. □

C OPEN RESEARCH DIRECTIONS

Several research directions remain open, including:

- Studying the APX-hardness of tree-ensemble RPIs (Theorem 7 gives NP-hardness of exact computation only).
- Extending the Rashomon-Conformal bridge to conditional coverage via Gibbs et al. [2025], yielding covariate-dependent thresholds $\varepsilon^\star(x)$.
- Connecting Rashomon sets to Bayesian posterior concentration; and extending to causal inference, where Rashomon treatment intervals connect to partial identification.

D ADDITIONAL EXPERIMENTAL DETAILS

D.1 EXPERIMENTAL PROTOCOL

All experiments use ridge regression with confidence level $\alpha = 0.10$. Total runtime for all experiments: 187.5 seconds. Code is available at [anonymized repository]. The dataset details can be found in Table 5.

Synthetic data generation. We generate $Y = \theta^{\star\top} X + \xi$ with $X \sim \mathcal{N}(0, I_p)$, $\xi \sim \mathcal{N}(0, \sigma^2)$, and $\theta^\star$ drawn uniformly from the unit sphere. Data is split 60/20/20 into training, calibration, and test sets.

UCI benchmark protocol. For each of the six datasets: (i) standardize features to zero mean and unit variance; (ii) use 60/20/20 train/calibration/test split; (iii) fit ridge regression with cross-validated λ (5-fold on training set); (iv) estimate $\hat{\sigma}^2$ from training residuals using the unbiased estimator $\hat{\sigma}^2 = \mathrm{RSS}/(n-p)$; (v) compute $\varepsilon^\star = \hat{\sigma}^2 z^2_{1-\alpha/2}/p$; (vi) set $\varepsilon_{\text{natural}} = 5\%$ of training MSE; (vii) compute Rashomon-conformal intervals (Algorithm 1) and standard split-conformal intervals; (viii) report coverage and width on the test set. Repeated 100 times with independent random splits.

Table 5: Dataset details

Dataset	n	p	Source
Concrete	1030	8	UCI MLR
Airfoil	1503	5	UCI MLR
Wine Quality (White)	4898	11	UCI MLR
Bike Sharing (daily)	731	6	UCI MLR
Protein Structure	45730	9	UCI MLR
Energy Efficiency	768	8	UCI MLR

D.2 FULL COVERAGE RESULTS (EXPERIMENT 4)

We report coverage at all three significance levels $\alpha \in \{0.05, 0.10, 0.20\}$.

$\varepsilon/\varepsilon^\star$	RCP	Raw RPI	CP	Oracle
0.50	0.945	0.802	0.944	0.948
1.00	0.949	0.920	0.944	0.948
2.00	0.979	0.979	0.944	0.948
5.00	0.999	0.999	0.944	0.948

$\alpha = 0.05$ **(nominal 95%):**

$\alpha = 0.10$ **(nominal 90%):**

$\alpha = 0.20$ **(nominal 80%):** RCP achieves $\geq 1 - \alpha$ in all 12 settings. Raw RPI undercoverage is most severe at low ε and high α: at $\varepsilon/\varepsilon^\star = 0.5$ and $\alpha = 0.20$, raw coverage is only 0.608 versus the nominal 0.80, a 24-percentage-point shortfall.

$\varepsilon/\varepsilon^\star$	RCP	Raw RPI	CP	Oracle
0.50	0.897	0.726	0.897	0.901
1.00	0.899	0.869	0.897	0.901
2.00	0.957	0.957	0.897	0.901
5.00	0.995	0.995	0.897	0.901

$\varepsilon/\varepsilon^\star$	RCP	Raw RPI	CP	Oracle
0.50	0.794	0.608	0.794	0.802
1.00	0.800	0.766	0.794	0.802
2.00	0.897	0.897	0.794	0.802
5.00	0.982	0.982	0.794	0.802

D.3 COMPUTATIONAL TIMING (EXPERIMENT 5)

p	Exact RPI (ms)	Langevin $K = 100$ (ms)	$W_{\mathrm{RQB}}/W_{\mathrm{RPI}}$
100	3.28	1.16	0.077–0.083
500	9.46	6.35	~ 0.05
1000	39.36	22.39	~ 0.03
2000	147.25	105.37	~ 0.02

The observed log-log slope is 1.24 ± 0.26 ($R^2 = 0.921$). For $p \leq 2000$, constant overhead dominates; the theoretical $O(p^3)$ Cholesky scaling becomes visible only for larger p. The key practical finding is that exact RPIs remain sub-second up to $p = 2000$.

D.4 SENSITIVITY ANALYSIS (EXPERIMENT 6)

Dataset	$\varepsilon^\star$ (theory)	$\hat{\varepsilon}^\star$ (data)	Ratio	$\mathrm{tr}(\hat{\Sigma}^{-1})$	$p/(1-\gamma)$
Concrete	0.134	0.031	0.232	38.3	8.1
Airfoil	0.264	0.128	0.483	12.4	5.0

The discrepancy between theoretical and data-driven estimators is explained by feature collinearity: $\mathrm{tr}(\hat{\Sigma}^{-1}) \gg p/(1-\gamma)$ on real data, reflecting correlated features that inflate the effective dimension.