Predicting Benign Overfitting via Spectral Geometry

Divit Rawal

UC Berkeley

Motivation

Classical intuition: More parameters $\Rightarrow$ more overfitting.
Modern ML:

- Deep / overparameterized models ($p \gg N$) often interpolate and still generalize.
- Some interpolating solutions are *benign*, others *catastrophic*.
- Width or sample size alone do not predict which.

Question. Can we predict benign vs non-benign interpolation using only the unlabeled feature matrix X?
Thesis. Generalization is controlled by a simple spectral quantity of the feature covariance, not by p or N in isolation.

Setup and Certificate

Model. Fixed features $\phi(x) \in \mathbb{R}^p$ and realizable linear regression

$$y = \phi(x)^\top \theta^\star + \varepsilon, \quad \mathbb{E}[\varepsilon \mid x] = 0,\ \mathbb{E}[\varepsilon^2 \mid x] \le \sigma^2,\ \|\theta^\star\|_2 \le B.$$

With N samples, $X \in \mathbb{R}^{N \times p}$, empirical covariance $\widehat{\Sigma} = \frac{1}{N} X^\top X$, population $\Sigma = \mathbb{E}[\phi(x)\phi(x)^\top]$.
Ridge predictor.

$$\widehat{\theta}_\lambda = (\widehat{\Sigma} + \lambda I)^{-1} \tfrac{1}{N} X^\top y.$$

Effective dimension.

$$d_\lambda(\widehat{\Sigma}) \doteq \mathrm{Tr}(\widehat{\Sigma}(\widehat{\Sigma} + \lambda I)^{-1}) = \sum_j \frac{\widehat{\mu}_j}{\widehat{\mu}_j + \lambda}.$$

Directions with $\widehat{\mu}_j \gg \lambda$ contribute ≈ 1 (active DoF); directions with $\widehat{\mu}_j \ll \lambda$ contribute ≈ 0 (frozen).

Spectral Risk Certificate

$$\boxed{\widehat{\mathcal{R}}_\lambda \doteq \underbrace{\lambda B^2}_{\text{worst-case bias}} + \underbrace{\frac{\sigma^2}{N} d_\lambda(\widehat{\Sigma})}_{\text{variance from geometry}}}$$

Computable from X alone (unlabeled geometry).

Spectral Intuition

Let $\alpha = p/N$ and consider small λ:

$$\frac{d_\lambda(\widehat{\Sigma})}{N} \approx \frac{\min(p, N)}{N} = \min(\alpha, 1).$$

- $\alpha < 1$: $d_\lambda/N \uparrow \alpha$ as $\lambda \to 0$.
- $\alpha > 1$: $d_\lambda/N \uparrow 1$ and **saturates** (only N samples).
- $d_\lambda/N \approx 1$: effective DoF $\approx$ sample size — the interpolation threshold.

Classical formulas suggest

$$\mathrm{Var} \sim \frac{d_\lambda/N}{1 - d_\lambda/N},$$

which blows up at $d_\lambda/N = 1$.
Prediction: In (d_λ/N, risk) coordinates:

- Nearly linear scaling of risk with d_λ/N away from 1.
- A vertical "spike" near $d_\lambda/N = 1$ (double-descent peak).

Bias–Variance Decomposition & Main Theorem

Excess empirical prediction error

$$\mathcal{E}_{\mathrm{emp}}(\theta; X) = \tfrac{1}{N}\|X(\theta - \theta^\star)\|_2^2 = (\theta - \theta^\star)^\top \widehat{\Sigma} (\theta - \theta^\star).$$

For ridge $\widehat{\theta}_\lambda$, let $\Delta_\lambda = \widehat{\theta}_\lambda - \theta^\star$. Conditioned on X,

$$\mathbb{E}[\mathcal{E}_{\mathrm{emp}}(\widehat{\theta}_\lambda; X) \mid X] = \mathrm{Bias}^2_\lambda(X) + \mathrm{Var}_\lambda(X).$$

Theorem 1 (Fixed-design spectral risk bound).
Assume $\|\theta^\star\|_2 \le B$ and $\mathbb{E}[\varepsilon\varepsilon^\top \mid X] \preceq \sigma^2 I_N$. Then for all $\lambda > 0$,

$$\mathbb{E}[\mathcal{E}_{\mathrm{emp}}(\widehat{\theta}_\lambda; X) \mid X] \le \lambda B^2 + \frac{\sigma^2}{N} d_\lambda(\widehat{\Sigma}).$$

Proof sketch:

- Express Δ_λ in eigenbasis of $\widehat{\Sigma}$.
- Show $\mathrm{Bias}^2_\lambda(X) = \lambda^2 \theta^{\star\top} (\widehat{\Sigma} + \lambda I)^{-1} \widehat{\Sigma} (\widehat{\Sigma} + \lambda I)^{-1} \theta^\star \le \lambda B^2$.
- Show $\mathrm{Var}_\lambda(X) \le (\sigma^2/N)\mathrm{Tr}[\widehat{\Sigma}^2(\widehat{\Sigma} + \lambda I)^{-2}] \le (\sigma^2/N) d_\lambda(\widehat{\Sigma})$.

Interpretation. Conservative: if $\widehat{\mathcal{R}}_\lambda$ is small, then risk is small even in the worst orientation of $\theta^\star$. When the variance term dominates, risk $\approx (\sigma^2/N) d_\lambda$, giving the linear trend in the collapse plot.

Certificate Stability in Random Design

Assume sub-Gaussian features with parameter κ.
Lemma (Lipschitz stability). If $\|\widehat{\Sigma} - \Sigma\|_{\mathrm{op}} \le \delta$, then

$$|d_\lambda(\widehat{\Sigma}) - d_\lambda(\Sigma)| \le \frac{\delta}{\lambda} \mathrm{rank}(\Sigma + \widehat{\Sigma}).$$

Lemma (Covariance concentration). w.p. $\ge 1 - \eta$,

$$\|\widehat{\Sigma} - \Sigma\|_{\mathrm{op}} \le C_\kappa \left(\sqrt{\frac{\mathrm{Tr}(\Sigma)}{N}} + \sqrt{\frac{\log(1/\eta)}{N}} \right).$$

Corollary (Effective-dimension concentration). Combining the two,

$$d_\lambda(\widehat{\Sigma}) \approx d_\lambda(\Sigma) \quad \text{with high probability.}$$

Spectral certificate. The empirical quantity

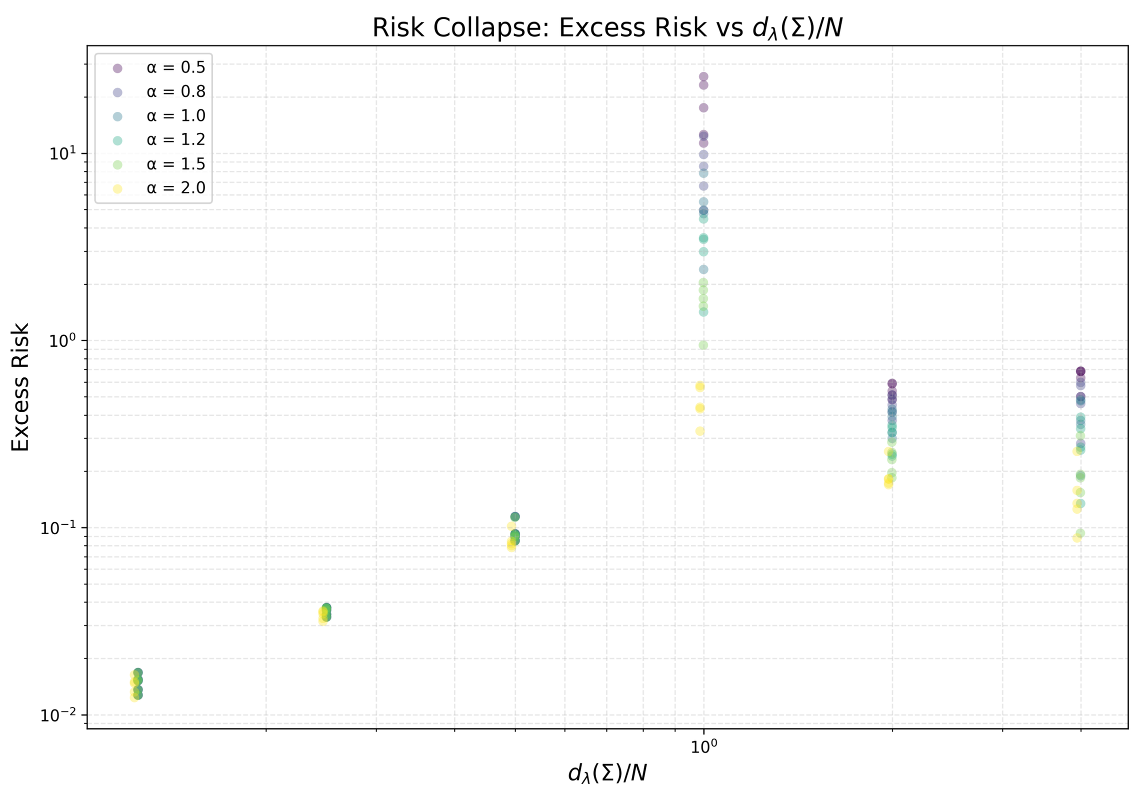
$$\widehat{\mathcal{R}}_\lambda = \lambda B^2 + \frac{\sigma^2}{N} d_\lambda(\widehat{\Sigma})$$

concentrates around the population bound $\mathcal{R}^{\mathrm{pop}}_\lambda = \lambda B^2 + \frac{\sigma^2}{N} d_\lambda(\Sigma)$.

- Works in $p \gg N$ (no assumption on $\lambda_{\min}(\widehat{\Sigma})$).
- If the certificate is small, both empirical and population risks are small: and we have a scalar certificate of benignness.

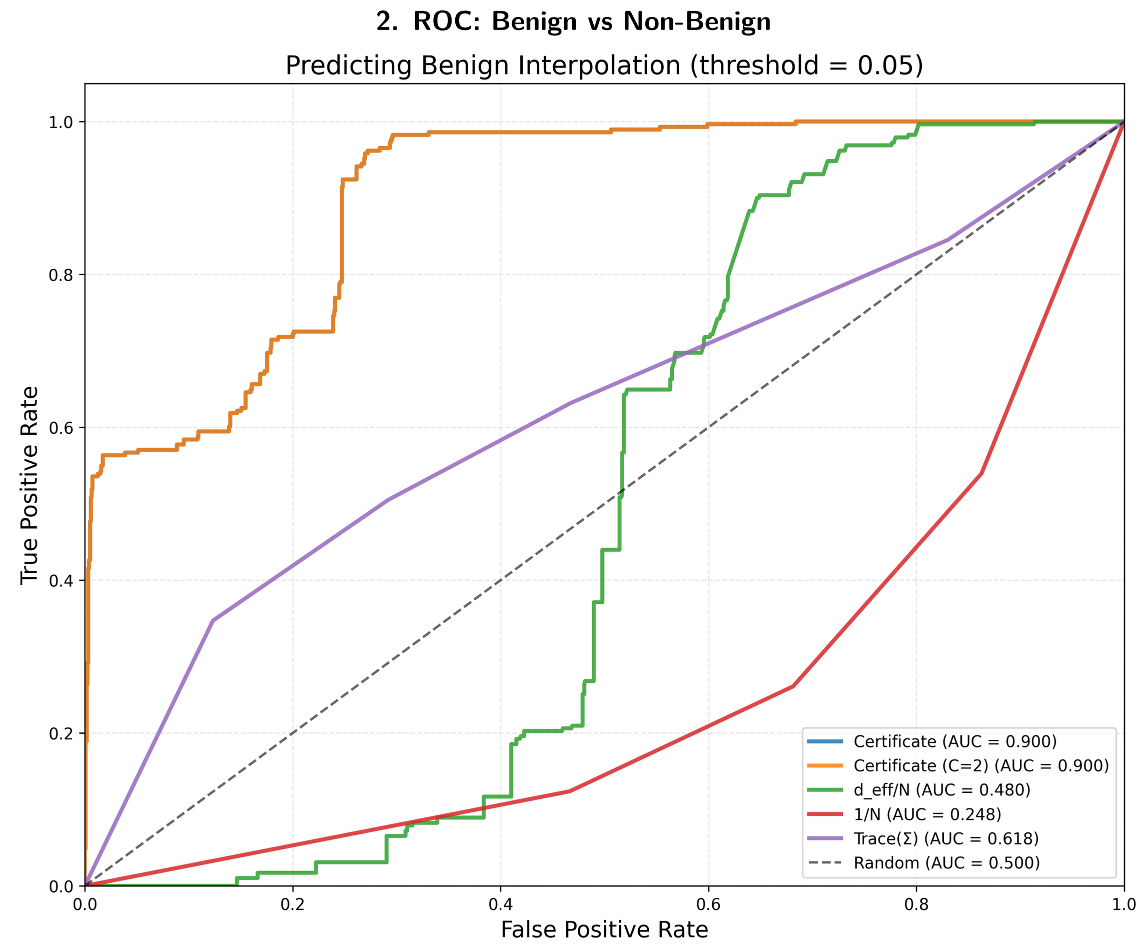
Empirical Validation: Geometry Predicts Benign Interpolation

1. Risk vs Effective Dimension



Excess risk vs. $d_\lambda(\Sigma)/N$ across aspect ratios $\alpha = p/N$ and regularization λ. Away from $d_\lambda/N \approx 1$, all points lie on an almost linear trend predicted by $(\sigma^2/N)d_\lambda$. The tall column at $d_\lambda/N \approx 1$ is the predicted interpolation "singularity".

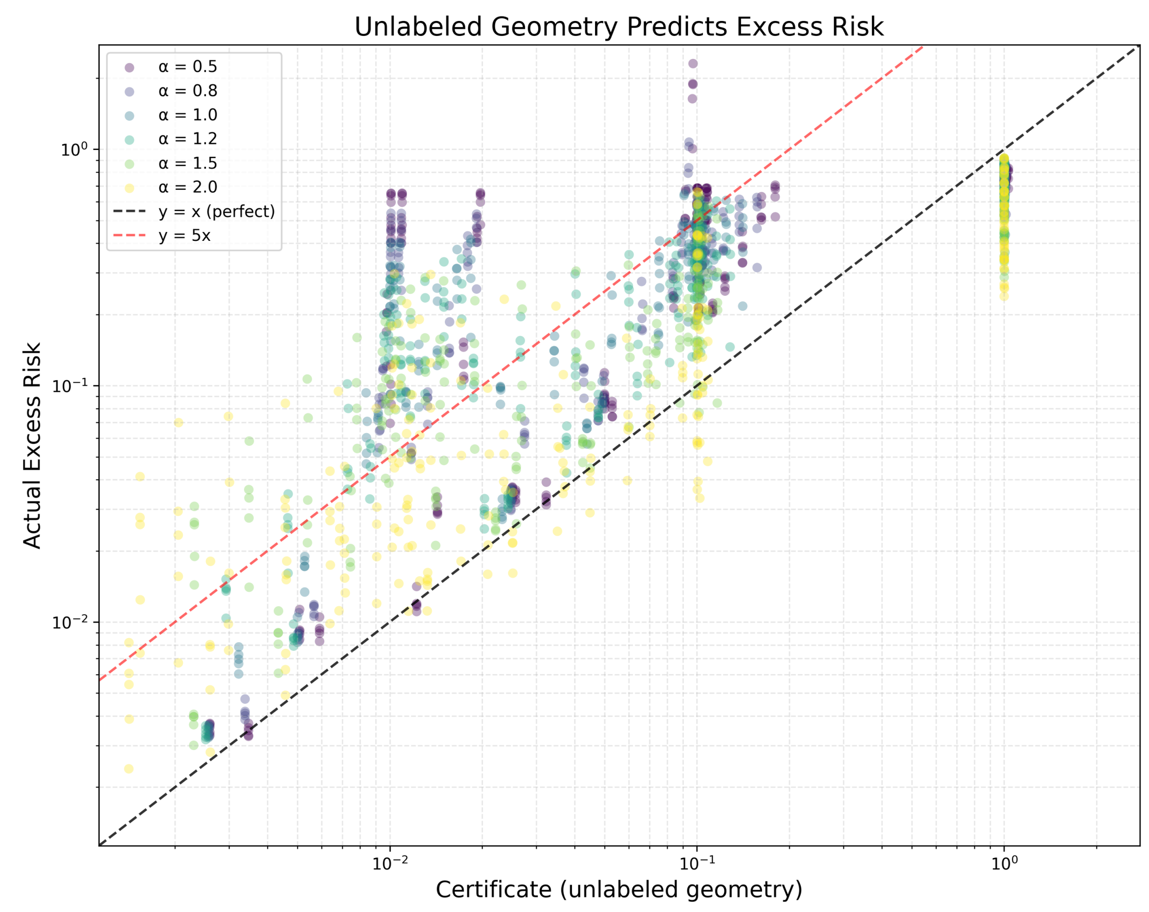
2. ROC: Benign vs Non-Benign



We label a model as benign if its test excess risk is below 0.05. The spectral certificate $\widehat{\mathcal{R}}_\lambda$ (blue/orange, AUC ≈ 0.90) strongly outperforms:

- effective dimension alone d_λ/N (AUC ≈ 0.48)
- classical $1/N$ scaling (AUC ≈ 0.25)
- total variance $\mathrm{Tr}(\Sigma)$ (AUC ≈ 0.62).

3. Certificate vs Actual Risk



Each point is one trained model (various α and λ). X-axis: $\widehat{\mathcal{R}}_\lambda$ (unlabeled geometry). Y-axis: test excess risk. Most points lie between $y = x$ (black) and $y = 5x$ (red): the certificate upper-bounds risk within a small constant factor over three orders of magnitude. $r^2 \approx .83$.

Summary, Contributions, and Outlook

Main takeaways.

- A single scalar

$$\widehat{\mathcal{R}}_\lambda = \lambda B^2 + \frac{\sigma^2}{N} d_\lambda(\widehat{\Sigma})$$

computed from unlabeled features serves as a spectral certificate of benign generalization.

- Risk curves across widths and regularization collapse when parameterized by effective degrees of freedom d_λ/N, with a universal spike at $d_\lambda/N \approx 1$.
- Width and sample size matter only through the spectrum of the learned representation.

Contributions.

1. A finite-sample bias–variance bound for ridge depending only on $d_\lambda(\widehat{\Sigma})$, valid for $p \gg N$.
2. Concentration results showing that the empirical certificate tracks the ideal population bound without relying on $\lambda_{\min}(\widehat{\Sigma})$.
3. Empirical evidence that the certificate almost linearly parameterizes risk and achieves AUC ≈ 0.9 for predicting benign vs non-benign interpolation.

Future directions.

- Apply spectral certificates to full deep nets (beyond last-layer linearization).
- Study robustness to label noise, distribution shift, and heavy-tailed features.
- Unsupervised choice of λ from spectral geometry alone.
- Connections to NTK, kernel methods, and information-theoretic capacity measures.

Stat mech view.

- $\sigma^2/N \approx$ temperature / noise power.
- d_λ counts active degrees of freedom.
- Benign interpolation arises when both energy (λB^2) and thermal fluctuations $(\sigma^2/N)d_\lambda$ are small.