Predicting Benign Overfitting via Spectral Geometry

Divit Rawal
Department of Statistics
UC Berkeley
divit.rawal@berkeley.edu

Abstract

Highly overparameterized models frequently interpolate noisy training data while still generalizing well, a phenomenon known as *benign overfitting*. Existing theory characterizes when interpolation is benign in terms of population spectral structure, but offers limited guidance on how to predict benign behavior from finite samples. This work proposes a simple geometry-driven certificate based on a single scalar of the empirical feature covariance:

$$d_\lambda(\widehat{\Sigma}) \;=\; \mathrm{Tr}\big[\widehat{\Sigma}(\widehat{\Sigma}+\lambda I)^{-1}\big],$$

the effective dimension at ridge scale λ. We derive a finite-sample fixed-design excess-risk bound for ridge regression whose only data-dependent term is $d_\lambda(\widehat{\Sigma})$, and give a stability argument showing that $d_\lambda(\widehat{\Sigma})$ tracks its population analogue under random design under mild spectral conditions. Empirically, excess risk exhibits a near-universal dependence on $d_\lambda(\widehat{\Sigma})/N$, and a simple spectral score built from $d_\lambda(\widehat{\Sigma})$ predicts benign versus non-benign interpolation with AUC ≈ 0.9 across a wide range of aspect ratios and regularization levels. We also outline an extension of the same spectral geometry to kernel ridge regression and Neural Tangent Kernel (NTK) models, and discuss why geometry-only certificates can degrade in predictive power for trained neural networks.

1 Introduction

Classical learning theory associates overparameterization with overfitting: models that interpolate noisy data are expected to generalize poorly. In contrast, modern practice shows that highly overparameterized models (including deep neural networks) often fit training data nearly perfectly while maintaining strong test performance. This phenomenon, known as *benign overfitting*, is closely related to double-descent behavior [Belkin et al., 2019].

Recent theory explains benign overfitting in linear and kernel models via the spectral structure of the data covariance: interpolation can be benign when spectral decay sufficiently controls variance [Bartlett et al., 2020, Montanari et al., 2021]. Kernel methods and the Neural Tangent Kernel (NTK) emphasize a similar message: generalization is governed by *effective degrees of freedom* rather than raw parameter count [Jacot et al., 2018].

This paper reframes benign overfitting through the lens of certification. Rather than asking when benign overfitting can occur in principle, we ask a predictive question:

> Given a feature matrix X and ridge scale λ, can we compute a simple scalar from the unlabeled geometry that (i) upper-bounds excess risk under minimal assumptions, and (ii) predicts whether interpolation will be benign?

We show that the answer is yes: the effective dimension

$$d_\lambda(\widehat{\Sigma}) \;=\; \mathrm{Tr}\big[\widehat{\Sigma}(\widehat{\Sigma}+\lambda I)^{-1}\big]$$

Preprint. Under review.

provides the key control parameter. The claim is not that effective dimension is new, but that it yields a finite-sample, geometry-only proxy that can be computed directly from a single dataset and used to rank regimes by benignness.[1] Our contributions are:

- A finite-sample fixed-design excess-risk bound for ridge regression whose only data-dependent term is $d_\lambda(\widehat{\Sigma})$.
- A stability result showing $d_\lambda(\widehat{\Sigma})$ concentrates around its population analogue under random design, without requiring well-conditioned covariances.
- Empirical evidence that risk curves collapse when parameterized by $d_\lambda(\widehat{\Sigma})/N$, and that a spectral score derived from the bound predicts benign interpolation with AUC ≈ 0.9.
- An extension of the same spectral geometry to kernel ridge regression and NTK models, plus an empirical discussion of where the geometry-only story weakens for trained neural networks.

2 Related Work

Benign overfitting and double descent. The phenomenon of benign overfitting was formalized by Belkin et al. [2019], who demonstrated that test error can decrease beyond the interpolation threshold. Bartlett et al. [2020] provided sharp conditions for benign overfitting in linear regression, showing that covariance spectral decay governs the excess risk of minimum-norm interpolants. High-dimensional asymptotics and extensions to classification refine this picture [Montanari et al., 2021].

Implicit bias and algorithms. Benign overfitting depends on both data and algorithm. Chatterji et al. [2022] analyze how implicit bias in optimization interacts with benign overfitting in overparameterized networks. In contrast, our certificate is deliberately conservative: it is designed to hold uniformly over all targets consistent with a norm bound, and therefore need not match the performance of implicitly regularized algorithms. This conservatism is a feature (it enables certification), but also highlights a limitation when algorithmic effects dominate geometry.

Effective dimension. Effective degrees of freedom appear throughout classical statistics for linear smoothers, ridge regression, and model selection [Hastie et al., 2009]. Their behavior in adaptive and high-dimensional regimes has been examined critically [Janson, 2015, Tibshirani, 2015]. We revisit effective dimension as a *scale-dependent spectral summary* that simultaneously (i) controls a clean finite-sample bound, and (ii) empirically collapses risk curves in overparameterized regimes.

Kernels and NTKs. Kernel methods provide an inherently spectral view of generalization. The NTK framework shows that infinitely wide neural networks trained by gradient descent behave like kernel methods [Jacot et al., 2018]. Our extension of the certificate to kernel/NTK settings connects these perspectives through the same effective-dimension geometry, while our empirical discussion emphasizes that NTK-at-initialization can be a weak predictor of post-training generalization for finite networks.

3 Setup and Effective Dimension

We study supervised regression with a feature map $\phi(x) \in \mathbb{R}^p$ and the realizable model

$$y \;=\; \phi(x)^\top \theta^\star + \varepsilon, \qquad \mathbb{E}[\varepsilon \mid x] = 0, \qquad \mathbb{E}[\varepsilon^2 \mid x] \le \sigma^2, \qquad \|\theta^\star\|_2 \le B. \tag{1}$$

Given samples $(x_i, y_i)_{i=1}^N$, let $X \in \mathbb{R}^{N\times p}$ be the feature matrix with rows $\phi(x_i)^\top$, and define the empirical covariance

$$\widehat{\Sigma} \;=\; \frac{1}{N} X^\top X.$$

[1] Here, we define an order on benignness as ordering by excess test risk.

3.1 Ridge regression and two risks

The ridge estimator is then

$$\widehat{\theta}_\lambda \;=\; (\widehat{\Sigma}+\lambda I)^{-1}\frac{1}{N}X^\top y, \qquad \lambda>0.$$

We distinguish:

- **Fixed-design prediction error** (conditioning on X):

$$E_{\mathrm{fd}}(\theta;X) \;=\; \frac{1}{N}\|X(\theta-\theta^\star)\|_2^2 \;=\; (\theta-\theta^\star)^\top\widehat{\Sigma}(\theta-\theta^\star).$$

- **Population prediction error** (random design with covariance $\Sigma=\mathbb{E}[\phi(x)\phi(x)^\top]$):

$$R(\theta) \;=\; \mathbb{E}_x\big[(\phi(x)^\top(\theta-\theta^\star))^2\big] \;=\; (\theta-\theta^\star)^\top\Sigma(\theta-\theta^\star).$$

Our main theorem is fixed-design; we then argue that its geometry term is stable and therefore predictive for held-out/test behavior under random design.

3.2 Effective dimension

Definition 1 (Effective dimension)**.** *For a PSD matrix M and $\lambda>0$, define*

$$d_\lambda(M) \;=\; \mathrm{Tr}\big[M(M+\lambda I)^{-1}\big].$$

If M has eigenvalues $(\mu_i)_i$, then $d_\lambda(M)=\sum_i \mu_i/(\mu_i+\lambda)$: directions with $\mu_i \gg \lambda$ contribute nearly 1 and directions with $\mu_i \ll \lambda$ contribute nearly 0.

4 Spectral Risk Bound and Certificate

Theorem 1 (Fixed-design spectral bound)**.** *Under* (1)*, for all $\lambda>0$,*

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

Proof sketch. Write $y=X\theta^\star+\varepsilon$ with $\mathbb{E}[\varepsilon\mid X]=0$ and $\mathbb{E}[\varepsilon\varepsilon^\top\mid X]\preceq\sigma^2 I$. Using $\widehat{\theta}_\lambda=(\widehat{\Sigma}+\lambda I)^{-1}\frac{1}{N}X^\top y$ and $\widehat{\Sigma}=\frac{1}{N}X^\top X$, we obtain

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

Plugging into $E_{\mathrm{fd}}(\theta;X)=(\theta-\theta^\star)^\top\widehat{\Sigma}(\theta-\theta^\star)$ yields a bias–variance decomposition after conditioning on X (the cross term vanishes by zero-mean noise). The bias term satisfies

$$\lambda^2(\theta^\star)^\top(\widehat{\Sigma}+\lambda I)^{-1}\widehat{\Sigma}(\widehat{\Sigma}+\lambda I)^{-1}\theta^\star \;\le\; \lambda\|\theta^\star\|_2^2 \;\le\; \lambda B^2,$$

using the eigenwise inequality $\lambda^2\mu/(\mu+\lambda)^2\le\lambda$. For the variance term, diagonalize $\widehat{\Sigma}=U\mathrm{diag}(\hat{\mu}_i)U^\top$ and bound

$$\frac{1}{N}\mathbb{E}\Big[\varepsilon^\top X(\widehat{\Sigma}+\lambda I)^{-1}\widehat{\Sigma}(\widehat{\Sigma}+\lambda I)^{-1}X^\top\varepsilon\,\Big|\,X\Big] \;\le\; \frac{\sigma^2}{N}\sum_i\frac{\hat{\mu}_i^2}{(\hat{\mu}_i+\lambda)^2} \;\le\; \frac{\sigma^2}{N}\sum_i\frac{\hat{\mu}_i}{\hat{\mu}_i+\lambda}=\frac{\sigma^2}{N}d_\lambda(\widehat{\Sigma}).$$

□

A geometry-driven certificate (up to scale). Motivated by Theorem 1, define the *spectral certificate*

$$R^b_\lambda(X) \;:=\; \lambda B^2 \;+\; \frac{\sigma^2}{N}\,d_\lambda(\widehat{\Sigma}). \tag{3}$$

The certificate is data-dependent only through the unlabeled geometry $d_\lambda(\widehat{\Sigma})$. The remaining factors B and σ^2 are scale parameters; in synthetic experiments they are known, and in applications one can treat them as hyperparameters or estimate them. For ranking regimes by benignness, the geometry term typically drives most of the variation across aspect ratios and λ.

5 Random-Design Stability

We now explain why $d_\lambda(\widehat{\Sigma})$ is a stable proxy for population geometry under random design. Assume $\phi(x)$ is mean-zero sub-Gaussian with covariance $\Sigma = \mathbb{E}[\phi(x)\phi(x)^\top]$, and $x_1, \ldots, x_N$ are i.i.d.

Lemma 1 (Stability via nuclear norm)**.** *For any PSD matrices A, B and $\lambda > 0$,*

$$|d_\lambda(A) - d_\lambda(B)| \;\le\; \frac{1}{\lambda}\,\|A - B\|_*,$$

where $\|\cdot\|_*$ is the nuclear norm. Then,

$$|d_\lambda(A) - d_\lambda(B)| \;\le\; \frac{1}{\lambda}\, r_{\mathrm{eff}}(A, B)\,\|A - B\|_{\mathrm{op}},$$

where $r_{\mathrm{eff}}(A, B)$ is any quantity satisfying $\|A - B\|_* \le r_{\mathrm{eff}}(A, B)\|A - B\|_{\mathrm{op}}$ (e.g., an effective rank induced by spectral decay).

Corollary. Standard matrix concentration bounds yield $\|\widehat{\Sigma} - \Sigma\|_{\mathrm{op}} = O_{\mathbb{P}}(N^{-1/2})$ under sub-Gaussian assumptions. Lemma 1 then implies that for fixed λ, $d_\lambda(\widehat{\Sigma})$ concentrates around $d_\lambda(\Sigma)$ provided the spectrum is not too "diffuse" relative to λ (formally, $r_{\mathrm{eff}}(\widehat{\Sigma}, \Sigma)/\lambda$ is controlled). This is the same spectral regime where ridge is well-behaved and where benign overfitting theory is typically phrased in terms of spectral decay.

6 Empirical Evaluation

We evaluate the predictive power and limitations of the spectral certificate. Our experiments address three questions: (i) whether excess risk exhibits a universal dependence on effective dimension, (ii) how conservative the certificate is as an upper bound, and (iii) whether the spectral score reliably discriminates benign from non-benign interpolation.

6.1 Experimental protocol

We generate features $x \sim \mathcal{N}(0, \Sigma)$ where Σ has a power-law spectrum $\lambda_j(\Sigma) = j^{-\gamma}$. We vary the spectral decay $\gamma \in \{0.5, 1.0, 1.5\}$, the aspect ratio $\alpha = p/N \in [0.1, 10]$, and the ridge regularization λ. The target $\theta^\star$ is drawn uniformly from the sphere of radius $B = \sqrt{p}$, and label noise is $\varepsilon \sim \mathcal{N}(0, 1)$. All results represent the mean of 20 independent trials. Code for full reproducibility is available at github.com/divitr/241_proj.

6.2 Risk collapse under effective dimension

Across all settings, prediction error curves collapse when plotted as a function of $d_\lambda(\widehat{\Sigma})/N$. For $d_\lambda/N \ll 1$, error scales approximately linearly with d_λ/N (Figure 1). As d_λ/N approaches 1, error peaks, reflecting the interpolation threshold, and then decreases again as λ increases. This collapse holds across a wide range of aspect ratios and covariance spectra, supporting the claim that d_λ is the correct geometric control parameter.

6.3 Tightness and conservatism of the certificate

We next compare realized prediction error to the certificate R^b_λ. Empirically, R^b_λ upper-bounds observed error and is typically within a modest multiplicative factor (Figure 2). This conservatism is expected: Theorem 1 is a worst-case bound over all targets with $\|\theta^\star\|_2 \le B$ and does not exploit favorable spectral alignment or algorithmic implicit regularization. Despite this, the geometry term $d_\lambda(\widehat{\Sigma})$ tracks error tightly across regimes, yielding strong correlation.

6.4 Predicting benign interpolation

Finally, we treat benign interpolation as a binary classification problem: a run is labeled benign if its prediction error falls below a fixed tolerance. Using R^b_λ as a score, we compute ROC curves (Figure 3) across all trials. The certificate achieves an AUC of approximately 0.90, substantially outperforming baselines such as d_λ/N alone, $1/N$, or $\mathrm{Tr}(\widehat{\Sigma})$.

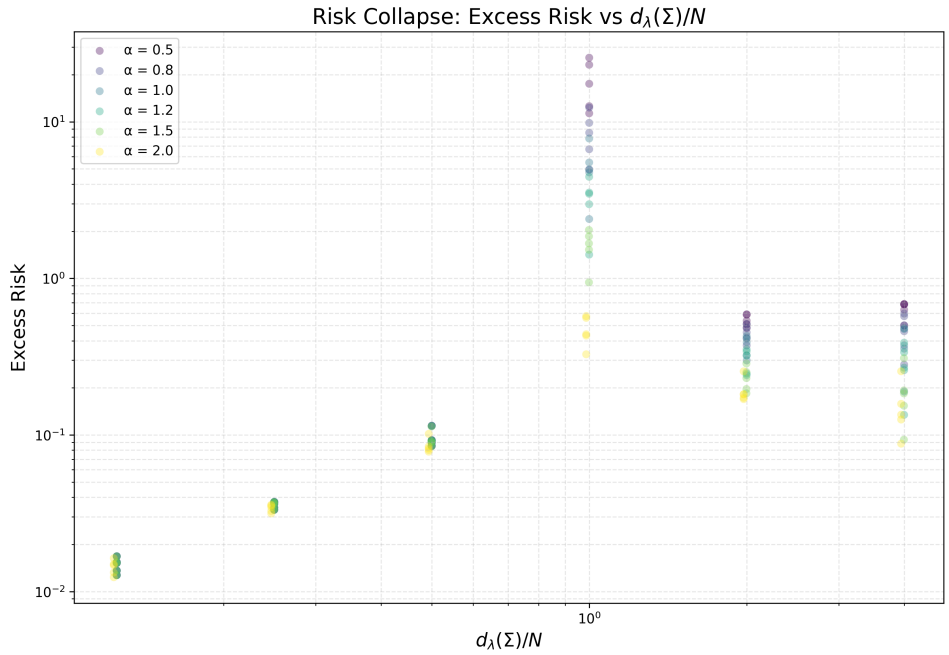

Figure 1: **Risk collapse under effective dimension.** Held-out prediction error plotted against the normalized effective dimension $d_\lambda(\widehat{\Sigma})/N$ for aspect ratios $\alpha = p/N \in \{0.5, 0.8, 1.0, 1.2, 1.5, 2.0\}$. Across widths and regularization levels, curves collapse onto a common geometry-driven profile: approximately linear growth for $d_\lambda/N \ll 1$, a pronounced peak near $d_\lambda/N \approx 1$ (the interpolation threshold), and decreasing error as λ increases. This collapse supports the view that effective dimension, rather than raw parameter count, governs generalization.

6.5 When the geometry-only certificate weakens for neural networks

To test whether the same geometric intuition extends beyond linear models, we also computed NTK-based certificates for neural networks (width $= 1024$), using the NTK at initialization as the kernel. The results are generally disappointing: while the NTK certificate can appear better calibrated as an upper bound in some regimes (roughly $\sim 2\times$ gaps in our pilot runs, versus $\sim 3\times$ for linear models), its predictive power is weak ($R^2 \approx 0.13$) and its discrimination can be worse than trivial baselines (AUC ≈ 0.71 versus ≈ 0.96 for simply using sample size N).[2]

This failure is informative and suggests that, for trained neural networks, generalization is often not determined by initialization-time kernel geometry alone:

- **Optimization dynamics can dominate geometry.** Implicit regularization from gradient descent (and architectural inductive bias) can matter more than the eigenstructure of the initialization NTK.
- **Initialization need not reflect the trained model.** The NTK at initialization can be a poor proxy for the representation learned during training, especially outside the strict lazy-training regime.
- **In extreme overparameterization, variation washes out.** When width $\gg N$, many runs interpolate nearly perfectly and differences in kernel geometry may not translate into meaningful differences in test error.

This leads us to believe that geometry-based certificates likely require algorithmic awareness: incorporating optimization and representation learning, not just the data geometry encoded by a fixed kernel.

[2]These neural-network numbers are reported from a small pilot using the same evaluation pipeline; we include them to highlight failure modes. It is also possible that for the widths we tested (restricted to 1024 due to computational limitations) that the network was still too deep in the feature learning regime, against the theoretical assumptions.

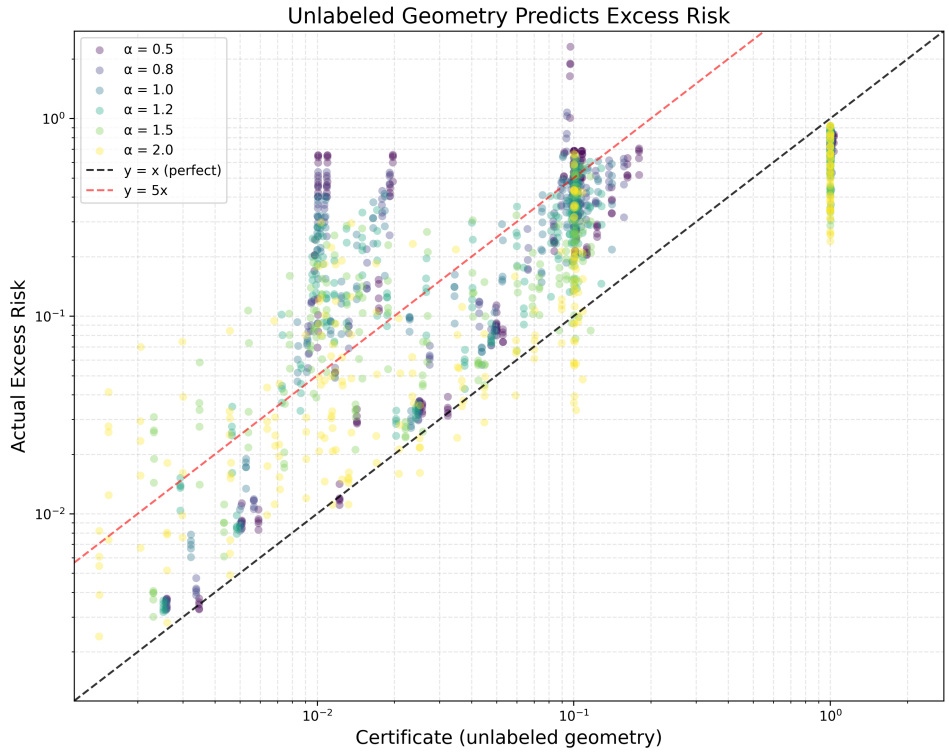

Figure 2: **Unlabeled geometry predicts excess risk.** Held-out prediction error plotted against the spectral certificate $R^b_\lambda = \lambda B^2 + (\sigma^2/N)\, d_\lambda(\widehat{\Sigma})$ across aspect ratios $\alpha = p/N \in \{0.5, 0.8, 1.0, 1.2, 1.5, 2.0\}$. The dashed black line denotes perfect prediction ($y = x$), while the dashed red line corresponds to a $5\times$ multiplicative gap. Most points lie between these lines, indicating that the certificate upper-bounds realized risk within a modest constant factor over several orders of magnitude. The certificate depends on the design only through $d_\lambda(\widehat{\Sigma})$, yet captures most of the variation in generalization performance. ($r^2 = 0.87$)

6.6 Summary

Overall, the experiments validate the theoretical picture in the kernel regime. Effective dimension governs a large fraction of the variability in generalization across aspect ratios and regularization levels. The certificate is conservative (by design), but remains highly predictive as a scalar score for benignness in the linear setting; for neural networks, naive NTK-based extensions can lose predictive power, motivating future work that integrates training dynamics.

7 Extensions to Kernels and NTKs

The same spectral geometry extends to kernel ridge regression. Given a kernel matrix $K \in \mathbb{R}^{N\times N}$ with entries $K_{ij} = k(x_i, x_j)$, kernel ridge regression solves

$$\widehat{f}_\lambda = \arg\min_{f\in\mathcal{H}_k} \frac{1}{N}\sum_{i=1}^{N}(y_i - f(x_i))^2 + \lambda\|f\|^2_{\mathcal{H}_k},$$

and the effective degrees of freedom take the form $d_\lambda(K) = \mathrm{Tr}[K(K + N\lambda I)^{-1}]$ (up to the conventional $N\lambda$ scaling). Analogously, in the infinite-width limit, neural networks trained by gradient descent correspond to kernel methods with the NTK K_{NTK} [Jacot et al., 2018], yielding an NTK-based certificate of the same form:

$$R^{b,\mathrm{NTK}}_\lambda = \lambda B^2 + \frac{\sigma^2}{N}\, d_\lambda(K_{\mathrm{NTK}}),$$

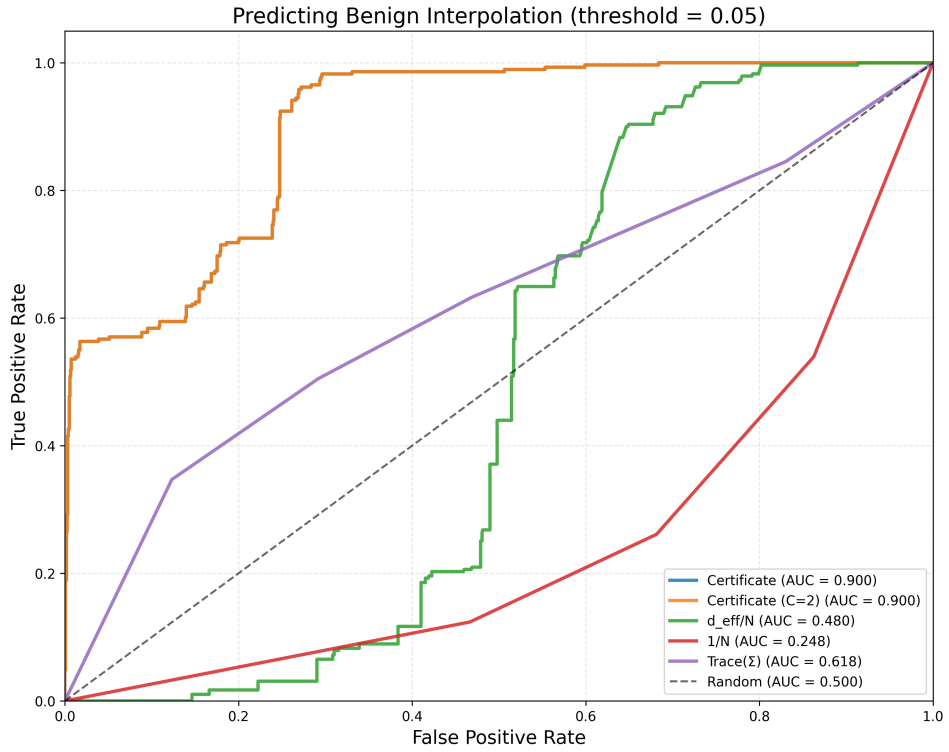

Figure 3: **Predicting benign interpolation from unlabeled geometry.** ROC curves for classifying runs as benign (prediction error < 0.05) using different scalar predictors. The spectral certificate R_λ^b achieves AUC ≈ 0.90, substantially outperforming normalized effective dimension d_λ/N (AUC ≈ 0.48), classical $1/N$ scaling (AUC ≈ 0.25), and total variance $\mathrm{Tr}(\widehat{\Sigma})$ (AUC ≈ 0.62). Despite being conservative as a bound, the certificate is highly effective at ranking regimes by benignness.

with the same interpretation: benignness is governed by effective dimension at the appropriate ridge scale. Section 6.5 emphasizes that for finite-width, feature-learning networks, using the initialization NTK alone can be a weak predictor of post-training generalization.

8 Conclusion

Effective dimension provides a simple, stable, and predictive geometric control parameter for benign overfitting in ridge regression. The resulting spectral certificate depends on the unlabeled design only through $d_\lambda(\widehat{\Sigma})$, is theoretically grounded via a finite-sample fixed-design bound, and is empirically accurate for predicting benign versus non-benign interpolation across a wide range of regimes. Extensions to kernels and NTKs suggest a unifying spectral-geometric view of benign overfitting, while the neural-network pilot highlights that bridging to real nets likely requires integrating optimization dynamics and representation learning into the certificate.

References

Peter L Bartlett, Philip M Long, Gábor Lugosi, and Alexander Tsigler. Benign overfitting in linear regression. *Proceedings of the National Academy of Sciences*, 117(48):30063–30070, 2020.

Mikhail Belkin, Daniel Hsu, Siyuan Ma, and Soumik Mandal. Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*, 116(32):15849–15854, 2019.

Niladri Shekhar Chatterji et al. The interplay between implicit bias and benign overfitting in two-layer linear networks. *Journal of Machine Learning Research*, 23(77):1–55, 2022.

Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning*. Springer, 2 edition, 2009.

Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Lucas Janson. Effective degrees of freedom: A flawed metaphor. *Statistical Science*, 30(4):629–645, 2015.

Andrea Montanari et al. The generalization error of max-margin linear classifiers: Benign overfitting and high-dimensional asymptotics. *Annals of Statistics*, 2021. Preprint arXiv:1911.01544.

Ryan J Tibshirani. Degrees of freedom and model search. *Statistical Science*, 30(4):549–573, 2015.