Near-Interpolators: Fast Norm Growth and Tempered Near-Overfitting

Anonymous Author(s)
Affiliation
Address
email

Abstract

We study linear regression when the input data population covariance matrix has eigenvalues $\lambda_i \sim i^{-\alpha}$ where $\alpha > 1$. Under a generic random matrix theory assumption, we prove that any near-interpolator, i.e., β whose training error is below the noise floor, must have its squared ℓ_2-norm growing super-linearly with the number of samples n: $\|\beta\|_2^2 = \Omega(n^\alpha)$. This implies that existing norm-based generalization bounds increase as the number of samples increases, matching the empirical observations from prior work. On the other hand, such near-interpolators when properly tuned achieve good generalization, where the test errors approach arbitrarily close to the noise floor. Our work demonstrates that existing norm-based generalization bounds are vacuous for explaining the generalization capability of *any* near-interpolators. Moreover, we show that the trade-off between train and test accuracy is better when the norm growth exponential is smaller.

1 Introduction

Learning algorithms that near-perfectly interpolate the training data such as deep neural networks have been surprisingly effective in practice despite conventional statistical wisdom suggesting otherwise [Zhang et al., 2021]. Near-interpolators arise frequently in modern machine learning, e.g., via early stopping rules [Ji et al., 2021, Kuzborskij and Szepesvári, 2022]. Therefore, understanding the fundamental trade-offs between near-interpolation and generalization is crucial.

Power law spectra assumptions arise commonly in popular settings such as in neural tangent kernels computed from practical networks. For instance, power law spectra of the neural tangent kernel matrix has been observed empirically in the MNIST, CIFAR-100 and CALTECH 101 datasets [Velikanov and Yarotsky, 2021, Wei et al., 2022, Murray et al., 2022]. Power law spectra assumptions provide a setting amenable to analysis while resembling real datasets, which has been used by previously Mallinar et al. [2022] to show that *perfect*-interpolators exhibit the so-called *tempered overfitting* phenomenon.

In this work, we analyze *near*-interpolators under power law spectra assumptions. Our result shows that such near-interpolators have norms increase super-linearly in the number of samples and exhibit tempered *near*-overfitting. Consequently, current norm-based generalization bounds are not applicable to explained this tempered near-overfitting behavior, and that tighter bounds are needed in the power law spectra assumption.

1.1 Our contributions

Super-linear growth of the squared norm. We show that when the data population covariance matrix has a power law spectra $\lambda_i = i^{-\alpha}$ with exponent $\alpha > 1$, *near*-interpolators have squared norm $\Omega(n^\alpha)$.

Submitted to the Mathematics of Modern Machine Learning Workshop at NeurIPS 2023. Do not distribute.

In this setting, our work answers the question raised in the the "Discussion" section of Koehler et al. [2021] regarding the growth of the norm for near-interpolators.

Tempered near-overfitting of near-interpolators. Tempered overfitting, coined by Mallinar et al. [2022], refers to the situation when estimators perfectly interpolate the training data and achieve test error $c\sigma^2$ for some $c \in (1, \infty)$, i.e., proportional to the Bayes optimal error/noise floor σ^2. Under the power law spectra $\lambda_i = i^{-\alpha}$ condition where $\alpha > 1$, they show that the proportionality constant $c = \alpha$. Under this same setting, we show that the *near*-interpolators achieve tempered *near*-overfitting. More precisely, properly tuned ridge regression achieve proportionality constant c down to the benign regime where $c = 1$.

1.2 Related works

The main difference between our work and that of Mallinar et al. [2022] is that our work establishes super-linear growth of the squared norm of near-interpolators. Our work is motivated by the empirical evidence found by Wei et al. [2022] suggests that norms of kernel ridge regressors grow rapidly potentially beyond the purview of norm-based bound. We confirm that bounds similar to the ones in Koehler et al. [2021] are indeed vacuous for power-law spectra. Therefore, our work suggests that explaining the generalization capability of near-interpolators will likely require new tools.

Ghosh and Belkin [2022] provides a lower bound on the *test error* for near-interpolators, demonstrating a fundamental trade-off between training and testing error. Our work derives a lower bound on the *norm* for near-interpolators. Therefore, our work complements both Mallinar et al. [2022] and Ghosh and Belkin [2022].

Our result is reminiscent of the result [Belkin et al., 2018, Theorem 1] in *classification*, which establishes that the RKHS norm of a "near-interpolating" classifier grows at rate $\Omega(\exp(n^{1/p}))$. Note that if the number of samples $n = \Theta(\text{poly}(p))$, then the lower bound does not grow to infinity and thus is only meaningful when $n = \Omega(\exp(p))$. In contrast, our result is for *regression*. While our results are not directly comparable, our lower bound is meaningful in the more practical $n \propto p$ regime.

For more related works, see Appendix Section E.

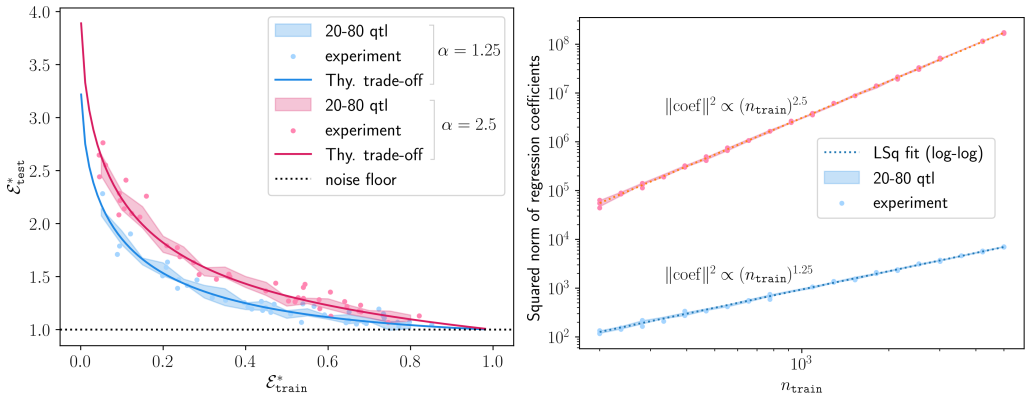

Figure 1: **Left**: *Trade-off between the testing and training errors from Proposition 2.9.* The solid lines are the parametrized curves $(x, y) = (\mathcal{E}^*_{\texttt{train}}, \mathcal{E}^*_{\texttt{test}})$ traced out by varying k (equivalently r). The resulting estimators can achieve a continuum regimes of overfitting. The scatter points are empirical results from synthetic experiments on the HDA model (Example 2.4). The value for r are tuned according to the tuning scheme in Remark A.1 for prescribed training error $\tau \approx \mathcal{E}^*_{\texttt{train}}$. The parameters are $n_{\texttt{train}} = n_{\texttt{test}} = 1000$, $\gamma_* = 0.5$, $\alpha \in \{1.25, 2.5\}$ and $\sigma^2 = 1$. See Appendix D for experimental details. **Right**: *Synthetic experiments validating the norm lower bound given by Theorem 2.3*. See Appendix D for additional experiment details. The squared norms are log-transformed then fitted by least squares to estimate the exponent α. The estimated exponents matches the true α's. Note that the trade-off is better (left) when the corresponding norm growth exponent is smaller (right).

2 Main results

Assumptions on the data distribution. Let n denote the number of samples, treated as the fundamental parameter. The feature dimension p is assumed to depend on n implicitly. The sample-to-feature ratio is denoted $\gamma := n/p \in (0, 1]$ and the asymptotic sample-to-feature ratio is denoted $\gamma_* := \lim_{n\to\infty} \gamma \in [0, 1]$. When $\gamma_* = 0$, p grows much faster than n. Denote by $X = [x_1, \ldots, x_n] \in \mathbb{R}^{p\times n}$ the data matrix and $y \in \mathbb{R}^n$ the training labels. Suppose that there exists a function $f : \mathbb{R}^p \to \mathbb{R}$ (depending on n) such that $y_i = \varepsilon_i + f(x_i)$ where $\varepsilon_i \in \mathbb{R}^n$ denote the noise. For instance, the well-specified case corresponds to when $f(x) = x^\top \beta^\star$ for some $\beta^\star \in \mathbb{R}^p$. Both y and ε are viewed as column vectors.

Assumptions on the noise. Suppose that the noise are independent across samples, has zero mean $0 = \mathbb{E}[\varepsilon_1]$ and variance $\sigma^2 = \mathbb{E}[\varepsilon_1^2] > 0$. For a positive integer p, let $\mathbb{I}_p$ denote the $p \times p$ identity matrix. Thus we have $\mathbb{E}[\varepsilon\varepsilon^\top] = \sigma^2\mathbb{I}_n$. Moreover, suppose that $\varepsilon \perp X$, i.e., the noise and the data are independent.

Definition 2.1. Ridge regression with regularizer $\varrho > 0$ is the vector $\hat{\beta}_\varrho$ defined via the optimization:

$$\hat{\beta}_\varrho := \operatorname{argmin}_{\beta\in\mathbb{R}^p} \tfrac{1}{n}\|X^\top\beta - y\|_2^2 + \varrho\|\beta\|_2^2. \tag{1}$$

Let $\hat{\Sigma} := n^{-1}XX^\top$ denote the sample covariance matrix, $\Sigma := \mathbb{E}[\hat{\Sigma}]$ the population covariance and $\check{G} := n^{-1}X^\top X$ the (scaled) gram matrix.

2.1 Super-linear growth of the squared norm

Our main result is that the expected squared norm of the KRR with $\varrho := rn^{-\alpha}$ regularizer grows at least on the order of n^α under suitable assumptions which we now introduce:

Assumption 2.2. Let $\alpha > 1$. The *exact eigenvalue decay (EVD)* condition with exponent α assumes that $\Sigma = \operatorname{diag}(\lambda_1, \cdots, \lambda_p)$ where $\lambda_i = i^{-\alpha}$.

Assumption 2.2 has been analyzed in many different context, most notably recently in being the setting for the so-called *tempered overfitting* phenomenon [Mallinar et al., 2022]. See the related works section for a detailed discussion.

Theorem 2.3. Assume that the exact EVD (Assumption 2.2) and certain random matrix-theoretic conditions hold. Define regularizers $\varrho := rn^{-\alpha}$ for the ridge regression (Definition 2.1) where $r > 0$ is a positive number. Then, we have $\mathbb{E}[\|\hat{\beta}_\varrho\|_2^2] = \Omega(n^\alpha)$.

See Figure 1-Left for experimental validation of the lower bound. Below, we will use the term "regularizer" to refer to both ϱ and r interchangeably.

The assumptions made in Theorem 2.3 are satisfied by the so-called HDA model, defined below. This is proved later in Lemma B.5.

Example 2.4. Bai and Silverstein [2010], Dobriban and Wager [2018]. The following is sometimes referred to as the *high-dimensional asymptotic (HDA)* model: *1.* $X = \Sigma^{1/2}Z$ where the entries of $Z = \{Z_{ij}\} \in \mathbb{R}^{p\times n}$ are i.i.d, have zero mean $\mathbb{E}[Z_{ij}] = 0$ and unit variance $\mathbb{E}[Z_{ij}^2] = 1$. The matrix Σ is positive semidefinite. *2.* $n/p \to \gamma_* \in (0, \infty)$, and *3.* Spectral distribution of Σ converges to a distribution H supported on $\mathbb{R}_{\geq 0}$.

Remark 2.5. When the conditions of Theorem 2.3 are met, the expected norm $\|\hat{\beta}_\varrho\|_2^2 = \Omega(n^\alpha)$. The current state-of-the-art uniform convergence generalization bound [Koehler et al., 2021, Corollary 1] are of the form $\|\beta\|_2/\sqrt{n}$ and are thus vacuous when $\|\beta\|_2^2 = \Omega(n^\alpha)$ when $\alpha > 1$. We note that the aforementioned results are for *perfect*-interpolators that achieve zero training error, rather than near-interpolators. To our knowledge, no analogous theory for near-interpolators is known. Whether the techniques of [Koehler et al., 2021] can be extended to explain near-interpolators is left as future work.

While stated for the ridge regressor as in Definition 2.1, our lower bound holds for any β that is "as good of an interpolator as $\hat{\rho}_\varrho$", i.e., β has training error less than that of $\hat{\rho}_\varrho$.

Definition 2.6. Let $\tau \geq 0$ arbitrary. The *minimum norm τ-near-interpolator* is defined as

$$\underline{\beta}_\tau := \operatorname{argmin}_{\beta\in\mathbb{R}^p} \|\beta\|_2^2 \quad s.t. \quad \tfrac{1}{n}\|X^\top\beta - y\|_2^2 \leq \tau. \tag{2}$$

A τ-*near-interpolator* is any $\beta \in \mathbb{R}^p$ that is feasible for Equation (2).

Proposition 2.7. Let $\rho > 0$ be arbitrary, $\hat{\beta}_\varrho \in \mathbb{R}^p$ be as in Definition 2.1, and $\tau := \frac{1}{n}\|X^\top \hat{\beta}_\varrho - y\|_2^2$. Consider $\underline{\beta}_\tau$ as in Definition 2.6. Then $\|\hat{\beta}_\varrho\|_2 = \|\underline{\beta}_\tau\|_2$. Consequently, if $\beta \in \mathbb{R}^p$ has less training error than $\|\hat{\beta}_\varrho\|_2$, then $\|\hat{\beta}_\varrho\|_2 \leq \|\beta\|_2$.

For the proof, see Appendix Section F.

2.2 Near-overfitting: Benign, tempered and everything in between

Simon et al. [2022] analyzed certain approximations of the testing and training errors of kernel ridge regression. While these approximations, dubbed the *eigenlearning framework*, are non-rigorous [Mallinar et al., 2022], they have been shown to be highly predictive in practice [Jacot et al., 2020, Bordelon et al., 2020, Canatar et al., 2021].

Following Mallinar et al. [2022], we use the eigenlearning framework to calculate the training and testing error of the estimators in Theorem 2.3 in terms of the *effective regularizer* [Wei et al., 2022], denoted by k.

Definition 2.8. Let $\alpha > 1$ and $\gamma_* \in [0, \infty)$. Define functions $\mathcal{I}(\cdot) \equiv \mathcal{I}_{\alpha,\gamma_*}(\cdot)$ and $\mathcal{J}(\cdot) \equiv \mathcal{J}_{\alpha,\gamma_*}(\cdot)$ as $\mathcal{I}(k) := \int_0^{1/\gamma_*} \frac{dx}{1+kx^\alpha}$, and $\mathcal{J}(k) := \int_0^{1/\gamma_*} \frac{dx}{(1+kx^\alpha)^2}$. When $\gamma_* = 0$, we assume that $1/\gamma_* = +\infty$.

Under Assumption 2.2, these functions from Definition 2.8 can be solved in closed-form given in Appendix G. The reason we work with the effective regularizer k rather than the regularizer r is that it is easier to calculate the approximations $\mathcal{E}^*_{\texttt{test}}, \mathcal{E}^*_{\texttt{train}}$ of the testing and training errors in the eigenlearning framework:

Proposition 2.9. In the setting of Section 2, assume further that f is well-specified, i.e., $f(x) = x^\top \beta^\star$ for some $\beta^\star$. Moreover, suppose that $\sup_{n=1,2\ldots} \|\beta^\star\|_2 < +\infty$. Assume the exact polynomial EVD condition (Assumption 2.2) with exponent $\alpha > 1$. For the estimator in Theorem 2.3 we have

$$\mathcal{E}^*_{\texttt{test}} \equiv \lim_{n\to\infty} \mathcal{E}_{\texttt{test}} = \sigma^2 \cdot \frac{1}{1-\mathcal{J}(k)}, \quad \text{and} \quad \mathcal{E}^*_{\texttt{train}} \equiv \lim_{n\to\infty} \mathcal{E}_{\texttt{train}} = \sigma^2 \cdot \frac{(1-\mathcal{I}(k))^2}{1-\mathcal{J}(k)}$$

Moreover, there exists $k_{\texttt{crit}} \in \mathbb{R}_{\geq 0}$ such that *1.* For each $r > 0$, there exists a unique $k \in (k_{\texttt{crit}}, +\infty)$ such that $r = \mathcal{R}(k) := k(1-\mathcal{I}(k))$, *2.* $\mathcal{R}$ is monotonically increasing on $(k_{\texttt{crit}}, +\infty)$, *3.* $\mathcal{E}^*_{\texttt{test}} > \sigma^2$ for all $k \in (k_{\texttt{crit}}, +\infty)$, *4.* $\lim_{k\to+\infty} \mathcal{E}^*_{\texttt{test}} = \sigma^2$, and *5.* $\frac{d}{d\alpha}\mathcal{E}^*_{\texttt{test}} > 0$ for any fixed $k > 0$.

For the proof of Proposition 2.9, see Appendix J. Thus, $\mathcal{R}$ is a bijection that relates the effective regularizer k and the (ordinary) regularizer r. Furthermore, note that $\lim_{k\to+\infty} \mathcal{E}^*_{\texttt{test}} = \sigma^2$ precisely states that the test error can be made arbitrarily close to the noise floor as k (equivalently, r) goes to infinity (See Proposition J.1 and Figure 2-Left).

Remark 2.10 (Trade-off quality vs norm growth exponent). Note that item 5 of Proposition 2.9 makes rigorous the observation that in Figure 1-left, the trade-off is better when the corresponding norm growth exponent is smaller (see Figure 1-right).

3 Discussion and limitations

Connection to early stopping. Typically, early stopping prevents the trained algorithm from perfectly interpolating the data. Can early stopped learning theory results, e.g., Ji et al. [2021], Kuzborskij and Szepesvári [2022], be applied to analyze near-interpolators?

Near-interpolators and uniform convergence generalization bound. Is possible to use uniform convergence-based approach to give non-vacuous generalization bound under the setting studied in this work? This question has already been raised by Dobriban and Wager [2018] in the context of classification.

Limitations. Our work is restricted to analyzing a random matrix model. Understanding the phenomenon uncovered in this paper in more general models and real world settings will be needed. Moreover, our work does not rule out the existence of uniform convergence generalization bound.

References

Yasaman Bahri, Ethan Dyer, Jared Kaplan, Jaehoon Lee, and Utkarsh Sharma. Explaining neural scaling laws. *arXiv preprint arXiv:2102.06701*, 2021.

Zhidong Bai and Jack W Silverstein. *Spectral analysis of large dimensional random matrices*, volume 20. Springer, 2010.

Zhidong Bai and Wang Zhou. Large sample covariance matrices without independence structures in columns. *Statistica Sinica*, pages 425–442, 2008.

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, Siyuan Ma, and Soumik Mandal. To understand deep learning we need to understand kernel learning. In *International Conference on Machine Learning*, pages 541–549. PMLR, 2018.

Raphaël Berthier, Francis Bach, and Pierre Gaillard. Tight nonparametric convergence rates for stochastic gradient descent under the noiseless linear model. *Advances in Neural Information Processing Systems*, 33:2576–2586, 2020.

Alberto Bietti, Luca Venturi, and Joan Bruna. On the sample complexity of learning with geometric stability. In *Advances in Neural Information Processing Systems*, 2021.

Blake Bordelon, Abdulkadir Canatar, and Cengiz Pehlevan. Spectrum dependent learning curves in kernel regression and wide neural networks. In *International Conference on Machine Learning*, pages 1024–1034. PMLR, 2020.

Sébastien Bubeck and Mark Sellke. A universal law of robustness via isoperimetry. *Advances in Neural Information Processing Systems*, 34:28811–28822, 2021.

Abdulkadir Canatar, Blake Bordelon, and Cengiz Pehlevan. Spectral bias and task-model alignment explain generalization in kernel regression and infinitely wide neural networks. *Nature communications*, 12(1):1–12, 2021.

Romain Couillet and Mérouane Debbah. Signal processing in large systems: A new paradigm. *IEEE Signal Processing Magazine*, 30(1):24–39, 2012.

Hugo Cui, Bruno Loureiro, Florent Krzakala, and Lenka Zdeborová. Generalization error rates in kernel regression: The crossover from the noiseless to noisy regime. In *Advances in Neural Information Processing Systems*, pages 10131–10143, 2021.

Edgar Dobriban and Stefan Wager. High-dimensional asymptotics of prediction: Ridge regression and classification. *The Annals of Statistics*, 46(1):247–279, 2018.

Nikhil Ghosh and Mikhail Belkin. A universal trade-off between the model size, test loss, and training loss of linear predictors. *arXiv preprint arXiv:2207.11621*, 2022.

Arthur Jacot, Berfin Simsek, Francesco Spadaro, Clément Hongler, and Franck Gabriel. Kernel alignment risk estimator: Risk prediction from training data. *Advances in Neural Information Processing Systems*, 33:15568–15578, 2020.

Ziwei Ji, Justin Li, and Matus Telgarsky. Early-stopped neural networks are consistent. *Advances in Neural Information Processing Systems*, 34:1805–1817, 2021.

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

Frederic Koehler, Lijia Zhou, Danica J Sutherland, and Nathan Srebro. Uniform convergence of interpolators: Gaussian width, norm bounds and benign overfitting. *Advances in Neural Information Processing Systems*, 34:20657–20668, 2021.

Ilja Kuzborskij and Csaba Szepesvári. Learning lipschitz functions by gd-trained shallow overparameterized relu neural networks. *arXiv preprint arXiv:2212.13848*, 2022.

Neil Rohit Mallinar, James B Simon, Amirhesam Abedsoltan, Parthe Pandit, Misha Belkin, and Preetum Nakkiran. Benign, tempered, or catastrophic: Toward a refined taxonomy of overfitting. In *Advances in Neural Information Processing Systems*, 2022.

Michael Murray, Hui Jin, Benjamin Bowman, and Guido Montufar. Characterizing the spectrum of the NTK via a power series expansion. *arXiv preprint arXiv:2211.07844*, 2022.

Courtney Paquette, Kiwon Lee, Fabian Pedregosa, and Elliot Paquette. Sgd in the large: Average-case analysis, asymptotics, and stepsize criticality. In *Conference on Learning Theory*, pages 3548–3626. PMLR, 2021.

Courtney Paquette, Bart van Merriënboer, Elliot Paquette, and Fabian Pedregosa. Halting time is predictable for large models: A universality property and average-case analysis. *Foundations of Computational Mathematics*, pages 1–77, 2022.

Jack W Silverstein and Sang-Il Choi. Analysis of the limiting spectral distribution of large dimensional random matrices. *Journal of Multivariate Analysis*, 54(2):295–309, 1995.

James B Simon, Madeline Dickens, Dhruva Karkada, and Michael R DeWeese. The eigenlearning framework: A conservation law perspective on kernel regression and wide neural networks. *arXiv preprint arXiv:2110.03922*, 2022.

Terrence Tao. Intuitive understanding of the Stieltjes transform. MathOverflow, 2011. URL `https://mathoverflow.net/q/79129`. Version: 2011-10-25.

Alexander Tsigler and Peter L Bartlett. Benign overfitting in ridge regression. *arXiv preprint arXiv:2009.14286*, 2020.

Maksim Velikanov and Dmitry Yarotsky. Explicit loss asymptotics in the gradient descent training of neural networks. *Advances in Neural Information Processing Systems*, 34:2570–2582, 2021.

Maksim Velikanov and Dmitry Yarotsky. Tight convergence rate bounds for optimization under power law spectral conditions. *arXiv preprint arXiv:2202.00992*, 2022.

Alexander Wei, Wei Hu, and Jacob Steinhardt. More than a toy: Random matrix models predict how real-world neural representations generalize. In *Proceedings of the 39th International Conference on Machine Learning*, pages 23549–23588. PMLR, 2022.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning (still) requires rethinking generalization. *Communications of the ACM*, 64(3):107–115, 2021.

Hongyang Zhang, Yihan Wu, and Heng Huang. How many data are needed for robust learning? *arXiv preprint arXiv:2202.11592*, 2022.

A Additional discussion on implications of Proposition 2.9

Using Proposition 2.9, we illustrate the trade-off between the training error versus the testing error in Figure 1-Right using closed-form expression for $\mathcal{E}^*_{\texttt{train}}$ and $\mathcal{E}^*_{\texttt{test}}$ are presented in Appendix G. Figure 2-Left demonstrates that empirical traing and test errors from synthetic experiments concentrate around the theoretical values $(\mathcal{E}^*_{\texttt{train}}, \mathcal{E}^*_{\texttt{test}})$ with growing n.

Remark A.1 (Tuning the regularizer). Proposition 2.9 allows for tuning the ridge parameter $\varrho := rn^\alpha$ to achieve a user-specified value of training error τ via the following procedure: First, use a binary search algorithm to find k_τ such that $\tau = \mathcal{E}^*_{\texttt{train}}$. Next, set $r := \mathcal{R}(k_\tau)$. Finally, set $\varrho := rn^\alpha$.

Remark A.2. The upshot of Proposition 2.9 is that any trade-off $(\mathcal{E}^*_{\texttt{train}}, \mathcal{E}^*_{\texttt{test}})$ on along the blue curve in Figure 1-Right can be achieved by the tuning algorithm in Remark A.1. For perfect-interpolators, Mallinar et al. [2022] shows that estimators with tempered overfitting achieve test error of exactly $\alpha\sigma^2$. In contrast, *near*-overfitting can achieve a continuum of test errors, i.e., $c\sigma^2$ where $c \in (1, c_{\max})$ belongs to a interval.

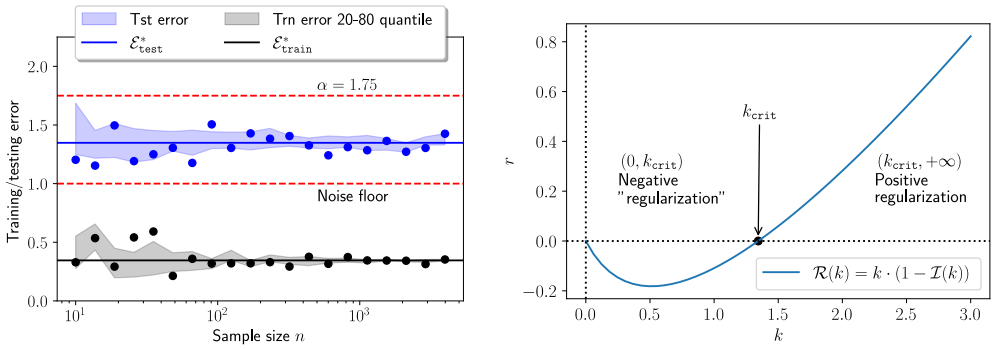

Figure 2: **Left**: *Synthetic experiment validating the approximations given by Proposition 2.9 using the same setup as in Figure 1.* By setting $r \approx 3.54$, we get a test error of ≈ 1.35 which is significantly below the tempered overfitting test error of $\alpha = 1.75$ in [Mallinar et al., 2022, Theorem 3.1]. See Figure 1 and Appendix D for experimental details. **Right**: *The $\mathcal{R}(k)$ function from Proposition 2.9.* The x-axis is the input k. Note that for $k < k_{\texttt{crit}}$ the regularizer r is negative. Although we are only interested in the $(k_{\texttt{crit}}, +\infty)$ portion, negative regularizers have been studied by Tsigler and Bartlett [2020] in the context of benign overfitting.

B Random Matrix Theory and Assumptions

In this section, we review and define the random matrix theory-based assumptions used to establish our results. These assumptions, while seemingly restrictive, are common in random matrix theory and showing their universality is an ongoing research area. See Remark B.11.

For $c \in \mathbb{R}$, let δ_c denote the *Dirac-delta measure* on $\mathbb{R}$ at c. In other words, for a Borel-measurable set $E \subseteq \mathbb{R}$, we have $\delta_c(E) = 1$ if $c \in E$ and $\delta_c(E) = 0$ otherwise.

Definition B.1 (Empirical spectral measure)**.** Let $M \in \mathbb{R}^{p\times p}$ be a matrix with real eigenvalues $\lambda_1, \ldots, \lambda_p$. The *empirical spectral measure* of M, denoted by $\texttt{esd}(M)$, is the measure on $\mathbb{R}$ given by $\texttt{esd}(M) = \frac{1}{p}\sum_{i=1}^p \delta_{\lambda_i}$.

We now state the weaker eigenvalue decay assumption sufficient for Theorem 2.3:

Assumption B.2 (Asymptotic EVD)**.** Let $\alpha > 1$. Suppose that $\texttt{esd}(n^\alpha\Sigma)$ converges to a distribution H on $\mathbb{R}_{\geq 0}$.

In Proposition C.7, we show that Assumption B.2 generalizes the earlier Assumption 2.2.

Random matrix theory are primarily concerned with analysis of the spectra of large random matrices. A key analytic tool is the *Stieltjes transform* of the empirical spectral measures of matrices:

Definition B.3. Let μ be a measure on $\mathbb{R}$. The *Stieltjes transform* of μ is the (complex-valued) function with input $z \in \mathbb{C}$ given by $\mathcal{S}_\mu(z) := \int \frac{\mu(t)dt}{t-z}$.

See Bai and Silverstein [2010, Appendix B.2] for reference.

For a matrix $M \in \mathbb{R}^{p\times p}$ with p real eigenvalues (e.g., when M is real and symmetric), the following holds:

$$\mathcal{S}_{\mathtt{esd}(M)}(z) = p^{-1}\mathrm{tr}\left((M - z\mathbb{I}_p)^{-1}\right).$$

We now state the other assumption made in Theorem 2.3:

Assumption B.4 (Positivity condition)**.** For every $r > 0$, suppose that $\lim_{n\to\infty} \mathbb{E}\big[\frac{d}{dr}(r\mathcal{S}_{\mathtt{esd}(n^\alpha \check{G})}(-r))\big] > 0$.

By leveraging the results of Silverstein and Choi [1995], we prove in Appendix C.1 the following:

Lemma B.5. Under the HDA model (Example 2.4) and the EVD condition (Assumption B.2), we have that $\lim_{n\to\infty} \mathbb{E}\big[\frac{d}{dr}(r\mathcal{S}_{\mathtt{esd}(n^\alpha \check{G})}(-r))\big] > 0$.

Next, we state what is sometimes referred to as the *self-consistent equation* [Tao, 2011]:

Assumption B.6. For each $r > 0$, there exists a unique $k \equiv k(r) \in \mathbb{R}$ such that the limit exists,

$$\tilde{\mathcal{I}}(k) := \lim_{n\to\infty} \tfrac{1}{n} \sum_{i=1}^{p} \tfrac{1}{1+kn^{-\alpha}\lambda_i^{-1}} \in \mathbb{R},$$

and the tuple (r, k) satisfies the equation

$$1 = \tfrac{r}{k} + \tilde{\mathcal{I}}(k). \tag{3}$$

Remark B.7. The functional relationship between r and k can be computed explicitly under the exact eigenvalue decay condition. As we will see in the proof of Proposition J.1, the expression $\tilde{\mathcal{I}}$ coincide with $\mathcal{I}$ from Definition 2.8.

Next, we state a version of the classical Marchenko-Pastur law for a random matrix ensemble X.

Assumption B.8 (Marchenko-Pastur law)**.** In the setting of Assumption B.6, further assume that almost surely

$$\lim_{n\to\infty} r\mathcal{S}_{\mathtt{esd}(n^\alpha \check{G})}(-r) = k\mathcal{S}_H(-k)$$

and $\lim_{n\to\infty} \frac{d}{dr}\left(r\mathcal{S}_{\mathtt{esd}(n^\alpha \check{G})}(-r)\right) = \frac{d}{dr}\left(k\mathcal{S}_H(-k)\right)$. We note that the k on the RHS depends on r.

Remark B.9. While we assume that the data is generated from the HDA model $X = \Sigma Z$, we note that, when $\Sigma = \mathrm{diag}(\{\lambda_i = i^{-\alpha}\})$ (Assumption 2.2), the empirical spectral measure of the *scaled* covariance $n^\alpha\Sigma$ converges to a limiting distribution. On the other hand, than the unscaled Σ matrix does not. Thus, the above Assumption B.8 reduces to the standard Marchenko-Pastur law when we consider the "fictitious" scaled HDA model $n^{\alpha/2}\sqrt{\Sigma}Z$ which is used in the analysis. The scaling of the regularizer $\varrho = rn^{-\alpha}$ in Definition 2.1 is chosen specifically to allow us to shift our analysis to this "fictitious" scaled HDA model.

The following is well-known [Dobriban and Wager, 2018]:

Theorem B.10 (Marchenko-Pastur theorem)**.** Under Example 2.4, both Assumption B.6 and Assumption B.8 hold.

Remark B.11. Many works have demonstrated these so-called universality phenomena for a broad range of random matrix ensemble beyond the simple HDA model. For instance, the Marchenko-Pastur law (Assumption B.8) and their variants has been extended to the setting where certain independence assumptions are dropped [Bai and Zhou, 2008] and when $\gamma_* = \lim_{n\to\infty} n/p = 0$ [Knowles and Yin, 2017, Wei et al., 2022]. As such, we expect Assumption B.4 to hold in these broader contexts as well. We leave this as an important future direction.

Having introduced the necessary assumptions, we now turn to proving Theorem 2.3.

C Norm lower bound in RMT settings

The goal of this section is to sketch the proof for Theorem 2.3. Complete proofs of all results are included in the Appendix. Throughout, we assume the setting of Section 2. The first key technical step the following:

Proposition C.1. $\mathbb{E}\|\hat{\beta}_\varrho\|_2^2 \geq n^{-1}\sigma^2\mathbb{E}[\mathrm{tr}((\hat{\Sigma}+\varrho\mathbb{I}_p)^{-2}\hat{\Sigma})]$.

Proof sketch of Proposition C.1. We first simplify $\|\hat{\beta}_\varrho\|_2^2$ using the well-known formula for ridge regression:

Lemma C.2. The closed-form solution for Equation (1) is given by the formula $\hat{\beta}_\varrho := (\hat{\Sigma}+\varrho\mathbb{I}_p)^{-1}\frac{1}{n}Xy$.

Next, let $M := (\hat{\Sigma}+\varrho\mathbb{I}_p)^{-1}\frac{1}{n}X$. Using the independence of X and ε, we get $\mathbb{E}[\|\hat{\beta}_\varrho\|_2^2] \geq \mathbb{E}[\mathrm{tr}(M^\top M\varepsilon\varepsilon^\top)]$. Since $M^\top M$ and $\varepsilon\varepsilon^\top$ are also independent, we have

$$\mathbb{E}[\mathrm{tr}(M^\top M\varepsilon\varepsilon^\top)] = \sigma^2\mathbb{E}[\mathrm{tr}(M^\top M)].$$

By $M^\top M = \frac{1}{n}(\hat{\Sigma}+\varrho\mathbb{I}_p)^{-1}\hat{\Sigma}(\hat{\Sigma}+\varrho\mathbb{I}_p)^{-1}$ and the cyclic property of trace, we get the desired inequality. □

The next step towards proving Theorem 2.3 is the following:

Proposition C.3. Let $\varrho := rn^{-\alpha}$. Then we have $\mathbb{E}\|\hat{\beta}_\varrho\|_2^2 \geq n^\alpha\sigma^2 \cdot \mathbb{E}\big[\frac{d}{dr}(r\mathcal{S}_{\mathtt{esd}(n^\alpha\check{G})}(-r))\big]$.

Proof sketch of Proposition C.3. We first relate the quantity $\mathrm{tr}((\hat{\Sigma}+\varrho\mathbb{I}_p)^{-2}\hat{\Sigma})$ inside the lower bound in Proposition C.1 to $\mathcal{S}_{\mathtt{esd}(n^\alpha\hat{\Sigma})}$, the Stieltjes transform of $n^\alpha\hat{\Sigma}$:

Lemma C.4. Let $M \in \mathbb{R}^{p\times p}$ be any symmetric matrix and $z \in \mathbb{R}$. Then we have

$$\tfrac{d}{dz}\mathrm{tr}(z(M+z\mathbb{I}_p)^{-1}) = \mathrm{tr}(M(M+z\mathbb{I}_p)^{-2}).$$

Next, we use the following well-known result for relating $\mathcal{S}_{\mathtt{esd}(n^\alpha\hat{\Sigma})}$ and $\mathcal{S}_{\mathtt{esd}(c\check{G})}$. For the sake of completeness, we include the proof in the Appendix.

Lemma C.5 (Gram-to-covariance)**.** Let $c \in \mathbb{R}$ and $z \in \mathbb{C}$ be arbitrary, then $\mathcal{S}_{\mathtt{esd}(c\hat{\Sigma})}(z) = \gamma \cdot \mathcal{S}_{\mathtt{esd}(c\check{G})}(z) - \frac{(1-\gamma)}{z}$.

Using Proposition C.1 and the two preceding Lemmas, the desired inequality follows from algebraic manipulation. □

Given the lower bound in Proposition C.3, our goal now is to relate the random quantity $\mathcal{S}_{\mathtt{esd}(n^\alpha\hat{\Sigma})}(\cdot)$ with the deterministic quantity $\mathcal{S}_{\mathtt{esd}(n^\alpha\Sigma)}(\cdot)$ using random matrix theory. Later, we will see that a consequence of Proposition C.7 is that $\mathbb{E}\big[\frac{d}{dr}(r\mathcal{S}_{\mathtt{esd}(n^\alpha\hat{\Sigma})}(-r))\big]$ is positive. This implies that $\mathbb{E}[\|\hat{\beta}\|_2^2] \geq o(n^\alpha)$. We now conclude with the proof of Theorem 2.3.

Proof of Theorem 2.3. Let $L := \lim_{n\to\infty}\mathbb{E}\big[\frac{d}{dr}(r\mathcal{S}_{\mathtt{esd}(n^\alpha\check{G})}(-r))\big] > 0$ be as in Assumption B.4. Thus, for all $n \gg 0$ sufficiently large, we have $\mathbb{E}\big[\frac{d}{dr}(r\mathcal{S}_{\mathtt{esd}(n^\alpha\check{G})}(-r))\big] > L/2 > 0$. By Proposition C.3, we get that $\mathbb{E}\|\hat{\beta}_\varrho\|_2^2 \geq n^\alpha\sigma^2 \cdot \frac{L}{2}$ for all $n \gg 0$, as desired. □

C.1 Positivity condition for the HDA model

This section will focus on the proof of Lemma B.5. Thus, throughout this section, we assume the setting of Example 2.4. Using the Marchenko-Pastur law (Assumption B.8) and calculus, we first show that

$$\lim_{n\to\infty}\mathbb{E}\big[\tfrac{d}{dr}(r\mathcal{S}_{\mathtt{esd}(n^\alpha\check{G})}(-r))\big] = \big(\tfrac{dr}{dk}\big)^{-1} \cdot \tfrac{d}{dk}\left(k\mathcal{S}_H(-k)\right)$$

where r and k are as in Assumption B.6. Thus, we reduce to showing the positivity of $\frac{dr}{dk}$ and $\frac{d}{dk}\left(k\mathcal{S}_H(-k)\right)$. See Appendix I. □

C.2 Convergence to limiting distribution

It remains to check that the exact eigenvalue decaying assumption (Assumption 2.2) indeed satisfy the condition 3 of Example 2.4.

Definition C.6. Given a measure μ on $\mathbb{R}$, we let $\texttt{cdf}[\mu]$ denote the cumulative distribution function of μ.

Proposition C.7. Under Assumption 2.2, we have the following:

$$\lim_{n\to\infty} \texttt{cdf}[\texttt{esd}(n^\alpha\Sigma)](t) = \begin{cases} 1-\gamma_* t^{-1/\alpha} & : t \geq \gamma_*^\alpha \\ 0 & \text{otherwise.} \end{cases}$$

Proof of Proposition C.7. The set of eigenvalues of $n^\alpha\Sigma$

$$\{(n/i)^\alpha\}_{i=1,\dots,p} = \{\underbrace{(\tfrac{n}{p})^\alpha}_{=\gamma^\alpha}, \dots, (\tfrac{n}{n+1})^\alpha, \underbrace{\tfrac{n}{n}}_{=1}, (\tfrac{n}{n-1})^\alpha, \dots, \underbrace{(\tfrac{n}{1})^\alpha}_{=n^\alpha}\}.$$

Thus, $\texttt{cdf}[\texttt{esd}(n^\alpha\Sigma)](t) = 0$ if $t < \gamma^\alpha$ and $= 1$ if $t > n^\alpha$.

Below, let $t \in [\gamma^\alpha, n^\alpha]$ and $j(t) \in \{1, \dots, p\}$ be the index such that $t \approx (n/j(t))^\alpha$ is as close as possible. Solving for $j(t)$, we have $j(t) \approx nt^{-1/\alpha}$. Thus, there are (approximately) $p - j(t)$ indices, denoted by i, such that $(n/i)^\alpha < (n/j(x))^\alpha$. Divide by p, we get the relative frequency of such indices i, which is $\texttt{cdf}[\texttt{esd}(n^\alpha\Sigma)](t) = 1 - (j(t)/p) \approx 1 - \gamma t^{-1/\alpha}$. This approximation becomes exact as $n \to \infty$. □

D Experiment

We run experiment with $\alpha = 1.75$ and $n/p = \gamma = 0.5$. We sample $\beta^\star \in \mathbb{R}^p$ such that $\beta_i^\star$ are i.i.d Gaussian with zero mean and variance $= 10/p$. For the data, we sample $X = \sqrt{\Sigma}Z$ as in the HDA model Example 2.4 where Z_{ij} are i.i.d standard Gaussian random variables and Σ is as in Assumption 2.2. The same set up is used for Figure 1. All code for the experiments are included in Appendix K.

E Expanded related works

Trade-offs in interpolation-based learning. In addition to Mallinar et al. [2022], Ghosh and Belkin [2022], Belkin et al. [2018], previous works have also studied the fundamental trade-off in learning algorithms between overparametrization and (Lipschitz) smoothness [Bubeck and Sellke, 2021] robustness and smoothness [Zhang et al., 2022].

Power law spectra. Many works reviewed in this section study the eigenvalues of kernel/gram matrices, while we are primarily interested in the covariance matrix spectra. However, we note that the covariance matrix have the same eigenvalues. Thus, results regarding the spectra applies to both kernel/gram and covariance matrices. Below, we will review works in this area using the term used by the original authors.

Power-law spectra datasets. Synthetic data with artificial power law EVD covariance have been used frequently as toy examples [Berthier et al., 2020, Mallinar et al., 2022]. On real datasets, power law EVD is often observed to describe neural tangent kernels (NTK) well in practice, including on MNIST ([Bahri et al., 2021, Fig, 4] and [Velikanov and Yarotsky, 2022, Fig. 2]), FASHION-MNIST [Cui et al., 2021, Fig. 7] CALTECH 101 [Murray et al., 2022, Fig. 1], CIFAR-100 [Wei et al., 2022, Fig. 3].

Theoretical machine learning works using power-law spectra. Bordelon et al. [2020] shows that power law EVD implies power law learning curve. Velikanov and Yarotsky [2021, §6.2] computes the power law EVD exponent for certain NTKs with ReLU to be $\alpha = 1 + \frac{1}{d}$. Murray et al. [2022] computes the EVD for NTKs with several different activations. The EVD condition is also known as the *capacity condition* in the kernel ridge regression literature. See Bietti et al. [2021] and the references there-in.

Bartlett et al. [2020, Theorem 6] shows that benign overfitting occurs when the covariance matrix eigenvalues $\lambda_i = i^{-1}\log^{-b}(i+1)$ for $b > 1$. Mallinar et al. [2022] studies power law decay for $\alpha \geq 1$ and proposes a taxonomy of overfitting into three categories: catastropic, tempered and benign.

Random matrix theory (RMT). The signal processing research community have long been using RMT for theoretical analysis [Couillet and Debbah, 2012]. Increasingly RMT has been applied to machine learning as well as a key tool for analysis. In particular, Dobriban and Wager [2018], Jacot et al. [2020] have applied RMT for (kernel) ridge regression. Paquette et al. [2022, 2021] uses the so-called local Marchenko-Pastur law [Knowles and Yin, 2017] to analyze gradient-based algorithms. Wei et al. [2022] also applies such local law to analyze the so-called *generalized cross- validation (GCV) estimator*.

F Proof for Proposition 2.7

Proof of Proposition 2.7. By definition, $\hat{\beta}_\varrho$ is feasible for the optimization in Equation (2) and thus $\|\hat{\beta}_\varrho\|_2 \geq \|\underline{\beta}_\tau\|_2$. Now, suppose for the sake of contradiction that $\|\underline{\beta}_\tau\|_2 < \|\hat{\beta}_\rho\|_2$. Then we have

$$\begin{aligned}
&\varrho\|\underline{\beta}_\tau\|_2^2 + \tfrac{1}{n}\|X\underline{\beta}_\tau - y\|_2^2 \\
&\leq \varrho\|\underline{\beta}_\tau\|_2^2 + \tau \quad \because \underline{\beta}_\tau \text{ is feasible for Equation (2)} \\
&< \varrho\|\hat{\beta}_\varrho\|_2^2 + \tau \quad \because \text{assumption } \|\underline{\beta}_\tau\|_2 < \|\hat{\beta}_\rho\|_2 \\
&= \varrho\|\hat{\beta}_\varrho\|_2^2 + \tfrac{1}{n}\|X\hat{\beta}_\varrho - y\|_2^2 \quad \because \text{Definition of } \tau
\end{aligned}$$

This contradicts the minimality of $\hat{\beta}_\rho$ for Equation (1). Thus, we've shown that $\|\underline{\beta}_\tau\|_2 = \|\hat{\beta}_\rho\|_2$. □

G Closed-form expression for Proposition 2.9

Let ${}_2F_1(a, b; c; z)$ be the *Gauss hypergeometric function*, implemented in SciPy as `scipy.special.hyp2f1`.

Lemma G.1. Let $\alpha > 1$ and $\gamma_* \in \mathbb{R}_{\geq 0}$ be fixed. The functions $\mathcal{I}, \mathcal{J}$ from Definition 2.8 can be written in closed form as:

$$\begin{aligned}
\mathcal{I}(k) &= \gamma_*^{-1} \times {}_2F_1(1, 1/\alpha; 1 + 1/\alpha; -k\gamma_*^{-\alpha}) \\
\mathcal{J}(k) &= \gamma_*^{-1} \times {}_2F_1(2, 1/\alpha; 1 + 1/\alpha; -k\gamma_*^{-\alpha}).
\end{aligned}$$

When $\gamma_* = 0$, we have

$$\begin{aligned}
\mathcal{I}(k) &= \frac{\pi}{\alpha} k^{-1/\alpha} \csc(\pi/\alpha) \\
\mathcal{J}(k) &= \frac{\pi(\alpha - 1)}{\alpha^2} k^{-1/\alpha} \csc(\pi/\alpha).
\end{aligned}$$

The above expressions can be obtained using computer algebra softwares such as MATHEMATICA. Note that the expression in the $\gamma_* = 0$ case has appeared previously in [Mallinar et al., 2022, Eqn. (22)] in a similar context. To the best of our knowledge, the expressions in the $\gamma_* \neq 0$ case are new, at least in the machine learning literature.

H Proofs for supporting lemmas of Theorem 2.3

Proof of Lemma C.2. Start with the objective function $\mathcal{F}(\beta) := \frac{1}{n}\|X^\top\beta - y\|_2^2 + \varrho\|\beta\|_2^2$. Take derivative with respect to β, we have

$$\frac{1}{2}\nabla_\beta\left(\frac{1}{n}\|X^\top\beta - y\|_2^2 + \varrho\|\beta\|_2^2\right) = \frac{1}{2}\nabla_\beta\left(\beta^\top(\hat{\Sigma} + \varrho\mathbb{I}_p)\beta - \frac{2}{n}\beta^\top Xy\right) = (\hat{\Sigma} + \varrho\mathbb{I}_p)\beta - \frac{1}{n}Xy.$$

Since $\nabla_\beta\mathcal{F}(\hat{\beta}_\varrho) = 0$, we are done. □

Lemma H.1 (Special case of Woodbury formula)**.** Let $M \in \mathbb{R}^{p\times n}$ be an arbitrary matrix and $\varrho \in (0,\infty)$. Then

$$(MM^\top + \varrho\mathbb{I}_p)^{-1}M = M(M^\top M + \varrho\mathbb{I}_n)^{-1} \in \mathbb{R}^{n\times p}$$

Proof of Lemma H.1. It suffices to prove Lemma H.1 for the special case when $\varrho = 1$, which we assume below. By the Woodbury matrix identity, we have

$$(MM^\top + \mathbb{I}_p)^{-1} = \mathbb{I} - M(M^\top M + \mathbb{I}_n)^{-1}M^\top \tag{4}$$

For brevity, let $P := MM^\top + \mathbb{I}_p$ and let $N := M^\top M + \mathbb{I}_n$. To proceed, we have

$$\begin{aligned}
&P^{-1}M \\
&= M - MN^{-1}M^\top M \quad \because \text{Multiplying (4) by } M \text{ on the right} \\
&= M(\mathbb{I}_n - N^{-1}M^\top M) \quad \because \text{Factoring out } M \text{ on the left} \\
&= M(\mathbb{I}_n - (\mathbb{I}_n - N^{-1})) \quad \because \mathbb{I}_n = N^{-1}N = N^{-1} + N^{-1}M^\top M \\
&= MN^{-1}
\end{aligned}$$

as desired. □

Proof of Lemma C.4. Without the loss of generality, suppose that $M = \operatorname{diag}(\lambda_i, \dots, \lambda_p)$. Then we have $f(z) := \operatorname{tr}(z(M + zI_p)^{-1}) = \sum_{i=1}^p \frac{z}{\lambda_i + z}$. Now, from elementary calculus, we have

$$\frac{d}{dx}\frac{x}{y+x} = (y+x)^{-1} - x(y+x)^{-2} = (y+x)^{-2}((y+x) - x) = \frac{y}{(y+x)^2}.$$

From this, we recover the fact that $\frac{d}{dz}f(z) = \sum_{i=1}^n \frac{\lambda_i}{(\lambda_i+z)^2} = \operatorname{tr}(M(M + z\mathbb{I}_p)^{-2})$, as desired. □

Proof of Lemma C.5. Without the loss of generality, we may assume that $c = 1$. Let $\hat\lambda_1 \geq \cdots \geq \hat\lambda_p$ be the eigenvalues of $\hat\Sigma$. Since $p > n$, we necessarily have that $\hat\lambda_{n+1} = \cdots = \hat\lambda_p = 0$. Moreover, $\hat\lambda_1, \dots, \hat\lambda_n$ are the eigenvalues of $\check G$. Now, unwinding the definition, we have

$$\mathcal{S}_{\mathtt{esd}(\hat\Sigma)}(z) = \frac{1}{p}\sum_{i=1}^p \frac{1}{\hat\lambda_i - z}$$

and

$$\mathcal{S}_{\mathtt{esd}(\check G)}(z) = \frac{1}{n}\sum_{i=1}^n \frac{1}{\hat\lambda_i - z}.$$

Thus,

$$\begin{aligned}
\mathcal{S}_{\mathtt{esd}(\hat\Sigma)}(z) &= \frac{1}{p}\left(\sum_{i=1}^n \frac{1}{\hat\lambda_i - z} + \sum_{i=n+1}^p \frac{1}{-z}\right) \\
&= \left(\frac{n}{p}\frac{1}{n}\sum_{i=1}^n \frac{1}{\hat\lambda_i - z}\right) - \frac{p-n}{p}\frac{1}{z} \\
&= \gamma \cdot \mathcal{S}_{\mathtt{esd}(\check G)}(z) - \frac{(1-\gamma)}{z}
\end{aligned}$$

as desired. □

Proof of Proposition C.1. Below, for brevity we let $a := f(X)$ and $M := (\hat\Sigma + \varrho\mathbb{I}_p)^{-1}\frac{1}{n}X$. We recall from the previous lemma that

$$\hat\beta_\varrho = (\hat\Sigma + \varrho\mathbb{I}_p)^{-1}\frac{1}{n}Xy = (\hat\Sigma + \varrho\mathbb{I}_p)^{-1}\frac{1}{n}X(f(X) + \varepsilon) = M(a + \varepsilon).$$

Thus,

$$\|\hat{\beta}_\varrho\|_2^2 = (a+\varepsilon)^\top M^\top M(a+\varepsilon) \geq \varepsilon^\top M^\top M\varepsilon + 2\varepsilon^\top M^\top Ma$$

Note that $\varepsilon \perp M^\top Ma$ since $\varepsilon \perp X$. Thus, since $\mathbb{E}[\varepsilon] = 0$, we have

$$\mathbb{E}[\|\hat{\beta}_\varrho\|_2^2] = \mathbb{E}[(a+\varepsilon)^\top M^\top M(a+\varepsilon)] \geq \mathbb{E}[\varepsilon^\top M^\top M\varepsilon] = \mathbb{E}[\mathrm{tr}(M^\top M\varepsilon\varepsilon^\top)]$$

Since $M^\top M \perp \varepsilon\varepsilon^\top$, we have

$$\mathbb{E}[\mathrm{tr}(M^\top M\varepsilon\varepsilon^\top)] = \mathrm{tr}(\mathbb{E}[M^\top M]\mathbb{E}[\varepsilon\varepsilon^\top]) = \mathrm{tr}(\mathbb{E}[M^\top M\sigma^2\mathbb{I}_n]) = \sigma^2\mathbb{E}[\mathrm{tr}(M^\top M)].$$

On the other hand, $M^\top M = \frac{1}{n}(\hat{\Sigma}+\varrho\mathbb{I}_p)^{-1}\hat{\Sigma}(\hat{\Sigma}+\varrho\mathbb{I}_p)^{-1}$. Using the cyclic property of trace, we get the desired inequality. □

Proof of Proposition C.3. Recall from Proposition C.1 that $\mathbb{E}\|\hat{\beta}\|_2^2 \geq n^{-1}\sigma^2\mathbb{E}[\mathrm{tr}((\hat{\Sigma}+\varrho\mathbb{I}_p)^{-2}\hat{\Sigma})]$. Below, we analyze the term inside the expectation. By the definition of the Stieltjes transform, we have

$$\mathrm{tr}(\varrho(\hat{\Sigma}+\varrho\mathbb{I}_p)^{-1}) = \mathrm{tr}(rn^{-\alpha}(\hat{\Sigma}+rn^{-\alpha}\mathbb{I}_p)^{-1}) = \mathrm{tr}(r(n^\alpha\hat{\Sigma}+r\mathbb{I}_p)^{-1}) = pr\mathcal{S}_{\mathtt{esd}(n^\alpha\hat{\Sigma})}(-r).$$

Therefore, by Lemma C.4, we have

$$\frac{d}{dr}\left(pr\mathcal{S}_{\mathtt{esd}(n^\alpha\hat{\Sigma})}(-r)\right) = \frac{d}{dr}\mathrm{tr}(\varrho(\hat{\Sigma}+\varrho\mathbb{I}_p)^{-1}) = \frac{d\varrho}{dr}\cdot\frac{d}{d\varrho}\mathrm{tr}(\varrho(\hat{\Sigma}+\varrho\mathbb{I}_p)^{-1}) = n^{-\alpha}\mathrm{tr}((\hat{\Sigma}+\varrho\mathbb{I}_p)^{-2}\hat{\Sigma}).$$

By Lemma C.5, we have

$$pr\mathcal{S}_{\mathtt{esd}(n^\alpha\hat{\Sigma})}(-r) = pr\left(\gamma\cdot\mathcal{S}_{\mathtt{esd}(n^\alpha\check{G})}(-r) + \frac{(1-\gamma)}{r}\right) = nr\mathcal{S}_{\mathtt{esd}(n^\alpha\check{G})}(-r) + p(1-\gamma)$$

Thus, we have

$$\frac{d}{dr}\left(pr\mathcal{S}_{\mathtt{esd}(n^\alpha\hat{\Sigma})}(-r)\right) = n\frac{d}{dr}\left(r\mathcal{S}_{\mathtt{esd}(n^\alpha\check{G})}(-r)\right)$$

from which we conclude that

$$\mathrm{tr}((\hat{\Sigma}+\varrho\mathbb{I}_p)^{-2}\hat{\Sigma}) = n^{\alpha+1}\frac{d}{dr}\left(r\mathcal{S}_{\mathtt{esd}(n^\alpha\check{G})}(-r)\right).$$

In view of $\mathbb{E}\|\hat{\beta}\|_2^2 \geq n^{-1}\sigma^2\mathbb{E}[\mathrm{tr}((\hat{\Sigma}+\varrho\mathbb{I}_p)^{-2}\hat{\Sigma})]$ from Proposition C.1, we get the desired inequality. □

I Continued from Appendix C.1

Before proceeding, we recall several definitions and notations adapted from Dobriban and Wager [2018]:

$$\lim_{n\to\infty}\mathbb{E}\left[\mathcal{S}_{\mathtt{esd}(n^\alpha\check{G})}(z)\right] = v(z) \tag{5}$$

is analogous to the $v(z)$ defined in the paragraph immediately following [Dobriban and Wager, 2018, Eqn. (2)]. The difference is our Equation (5) is for the limit of the n^α-scaled matrices $n^\alpha\check{G}$, rather than for $\check{G}$ as in Dobriban and Wager [2018].

Let $H = \lim_{n\to\infty}\mathtt{cdf}[\mathtt{esd}(n^\alpha\Sigma)]$ be the limiting distribution as in Assumption B.2. Plugging in $z = -r$ into Dobriban and Wager [2018, Eqn. (A.1)], we have

$$-\frac{1}{v(-r)} = -r - \frac{1}{\gamma}\int\frac{tdH(t)}{1+tv(-r)}.$$

Letting $k \equiv k(r) := \frac{1}{v(-r)}$, we can rewrite the above as

$$1 = \frac{r}{k} + \frac{1}{\gamma}\int\frac{tdH(t)}{k+t}. \tag{6}$$

By construction, we have

$$\frac{1}{\gamma}\int \frac{tdH(t)}{k+t} = \lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{p} \frac{1}{1+kn^{-\alpha}\lambda_i^{-1}}$$

where the RHS is as in Assumption B.6. Consequently, the tuple r, k from Assumption B.6 coincide with the earlier definition of $k := \frac{1}{v(-r)}$ right before Equation (6). Having established the above, we now proceed to:

Proof of Lemma B.5. By the product rule, we have

$$\tfrac{d}{dr}\left(r\mathcal{S}_{\mathtt{esd}(n^\alpha \check{G})}(-r)\right) = \mathcal{S}_{\mathtt{esd}(n^\alpha \check{G})}(r) - r\mathcal{S}'_{\mathtt{esd}(n^\alpha \check{G})}(-r)$$

Now, taking the limit of the above equation on both side, we have

$$\begin{aligned}
\lim_{n\to\infty} \mathbb{E}\left[\tfrac{d}{dr}\left(r\mathcal{S}_{\mathtt{esd}(n^\alpha \check{G})}(-r)\right)\right] &= \lim_{n\to\infty} \mathbb{E}\left[\mathcal{S}_{\mathtt{esd}(n^\alpha \check{G})}(-r) - r\mathcal{S}'_{\mathtt{esd}(n^\alpha \check{G})}(-r)\right] \\
&= v(-r) - rv'(-r) \qquad \because \text{Definition of } v \text{ and } v' \\
&= \tfrac{d}{dr}\left(rv(-r)\right) \qquad \because \text{Product rule} \\
&= \tfrac{d}{dr}\left(k\mathcal{S}_H(-k)\right) \qquad \because \text{ Marchenko-Pastur law (Assumption B.8)} \\
&= \tfrac{dk}{dr}\cdot\tfrac{d}{dk}\left(k\mathcal{S}_H(-k)\right) \qquad \because \text{Chain rule} \\
&= \left(\tfrac{dr}{dk}\right)^{-1}\cdot\tfrac{d}{dk}\left(k\mathcal{S}_H(-k)\right) \qquad \because \text{Inverse function theorem}
\end{aligned}$$

To complete the proof, it suffices to show that both $\frac{dr}{dk}$ and $\frac{d}{dk}\left(k\mathcal{S}_H(-k)\right)$ are positive which will be checked in the next two lemmas. □

Lemma I.1. The function $\frac{dr}{dk}$ evaluated at k is positive.

Proof of Lemma I.1. Recall that $k = \frac{1}{v(-r)}$. Thus, we have

$$\frac{dk}{dr}(r) = (-1)\frac{1}{v(-r)^2}(-1)\cdot v'(-r) = \frac{v'(-r)}{v(-r)^2}.$$

From the proof of Silverstein and Choi [1995, Theorem 4.1], we see that $v'(\cdot) > 0$ for all negative inputs. In particular, $v'(-r) > 0$ which implies that $\frac{dk}{dr}$ is positive. By the inverse function theorem, we have $\frac{dr}{dk} = (\frac{dk}{dr})^{-1}$ is also positive. □

Lemma I.2. The quantity $\frac{d}{dk}\left(k\mathcal{S}_H(-k)\right)$ is positive.

Proof of Lemma I.2. Plugging in $z = -r$ into Dobriban and Wager [2018, Eqn. (3)], we have

$$v(-r) - \frac{1}{r} = \frac{1}{\gamma}\left(m(-r) - \frac{1}{r}\right). \tag{7}$$

Now,

$$\begin{aligned}
rm(-r) &= \gamma r v(-r) + (1-\gamma) \quad \because \text{Equation (7)} && (8)\\
&= \gamma\frac{r}{k} + (1-\gamma) \quad \because \text{Definition of } k && (9)\\
&= \left(\gamma - \int \frac{tdH(t)}{k+t}\right) + (1-\gamma) \quad \because \text{Equation (6)} && (10)\\
&= 1 - \int \frac{tdH(t)}{k+t} && (11)\\
&= \int \frac{kdH(t)}{k+t} \quad \because 1 = \int dH(t) = \int \frac{k+t}{k+t}dH(t) && (12)\\
&= k\mathcal{S}_H(-k). && (13)
\end{aligned}$$

Thus, differentiating under the integral, we have

$$\frac{d}{dk}(k\mathcal{S}_H(-k)) = \int \frac{d}{dk}\left(\frac{k}{k+t}\right)dH(t) = \int \frac{tdH(t)}{(k+t)^2} > 0$$

as desired. □

J Proof of Proposition 2.9

We begin by analyzing the functions defined in Definition 2.8 and prove the items 1 and 2 of the "Moreover" part of Proposition 2.9:

Proposition J.1. Let $\mathcal{I}$ and $\mathcal{J}$ be functions as defined in Definition 2.8. Under Assumption 2.2 and Assumption B.6, we have that $r = \mathcal{R}(k) := k \cdot (1 - \mathcal{I}(k))$ and $\frac{dr}{dk} = 1 - \mathcal{J}(k)$.

Furthermore, the following holds:

1. $\mathcal{R}(k) \asymp k$ for $k \gg 0$,
2. There exists $k_{\texttt{crit}} > 0$ such that $\mathcal{R}(k_{\texttt{crit}}) = 0$, $\mathcal{R}$ is increasing and positive on $(k_{\texttt{crit}}, +\infty)$.
3. $\mathcal{J}(k) < 1$ for $k \in (k_{\texttt{crit}}, +\infty)$ and $\mathcal{J}(+\infty) = 0$.

Proof of Proposition J.1. We begin by proving the first part: that $r = \mathcal{R}(k) := k \cdot (1 - \mathcal{I}(k))$ and $\frac{dr}{dk} = 1 - \mathcal{J}(k)$. Rewrite the limit in Equation (3) as

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{p} \frac{1}{1 + kn^{-\alpha}\sigma_i^{-1}} = \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n/\gamma} \frac{1}{1 + k(i/n)^{\alpha}} = \int_0^{1/\gamma_*} \frac{dx}{1 + kx^{\alpha}}$$

The right-most equality follows from the definition of the (Riemann) integral. If $\gamma_* = 0$, then $1/\gamma_* = +\infty$ and the above is interpreted as an improper Riemann integral. Now, rearranging Equation (3), we get the desired formula of $r = \mathcal{R}(k) := k \cdot (1 - \mathcal{I}(k))$. The formula for $\frac{dr}{dk}$ follows from differentiating under the integral theorem. Note that this also proves the assertion made in Remark B.7.

For the first item of the "Furthermore" part, it suffices to show that $\lim_{k\to+\infty} \mathcal{I}(k) = 0$. This follows from the fact that $\lim_{k\to+\infty} \frac{1}{1+kx^{\alpha}} = 0$ for all $x > 0$, integrability of the function $(1 + x^{\alpha})^{-1}$ over $\mathbb{R}_{\geq 0}$, and the dominated convergence theorem. Likewise, $\lim_{k\to\infty} \mathcal{J}(k) = 0$ as well.

For the second item of the "Furthermore" part, we note that for all x sufficiently large, we have $\frac{dr}{dk} > 0$ since $\lim_{k\to\infty} \mathcal{J}(k) = 0$. Now, let $k_{\texttt{crit}}$ be the largest real number such that $\mathcal{R}(k_{\texttt{crit}}) = 0$. Since $\mathcal{R}(0) = 0$, we must have $k_{\texttt{crit}} \geq 0$.

For all $k > k_{\texttt{crit}}$, we claim that $\mathcal{I}(k) < 1$. To see this, assume the contrary. Then by the fact that $\lim_{k\to+\infty} \mathcal{I}(k) = 0$ and the intermediate value theorem, there must exists k' such that $k' > k$ such that $\mathcal{I}(k') = 1$ which implies that $\mathcal{R}(k') = 0$. This contradicts the maximality of $k_{\texttt{crit}}$.

Finally, since $1 + kx^{\alpha} \leq (1 + kx^{\alpha})^2$ for all $k \geq 0$ and $x \geq 0$, we have that $\mathcal{I}(k) \geq \mathcal{J}(k)$ for all such k's. Thus, by the previous claim, for all $k > k_{\texttt{crit}}$, we have $1 > \mathcal{I}(k) \geq \mathcal{J}(k)$. This proves that $\frac{dr}{dk} > 0$ for all $k > k_{\texttt{crit}}$, as desired. □

J.1 Review of the eigenlearning framework

Before proceeding with finishing the proof of Proposition 2.9, we briefly review the eigenlearning framework. Simon et al. [2022] calculates the test error for the estimator

$$\check{\beta}_\delta := X(X^\top X + \delta \mathbb{I}_n)^{-1} y = X(n\check{G} + \delta \mathbb{I}_n)^{-1} y \tag{14}$$

for kernel ridge regression using the so-called *eigenlearning equations* [Simon et al., 2022, Section 4.1]. Below, we recall some relevant parts of the framework:

Definition J.2 (Eigenlearning eqn. specialized to setting in Section 2)**.** Suppose that the ground truth regression function is linear, i.e., $f(x) = x^\top \beta^\star$ for some $\beta^\star \in \mathbb{R}^p$. Let δ and κ satisfy the equation

$$n = \frac{\delta}{\kappa} + \sum_{i=1}^{p} \frac{\lambda_i}{\lambda_i + \kappa}. \tag{15}$$

Define the following n-dependent quantities:

1. *Overfitting coefficient*: $\mathcal{E}_{\texttt{coef}} := n\frac{d\kappa}{d\delta}$
2. *Testing error*: $\mathcal{E}_{\texttt{test}} := \mathcal{E}_{\texttt{coef}}(\sigma^2 + C)$ where
$$C = \sum_{i=1}^{p} (1 - \mathcal{L}_i)(\beta_i^\star)^2 \quad \text{and} \quad \mathcal{L}_i := \frac{\lambda_i}{\lambda_i + \kappa}.$$
3. *Training error*: $\mathcal{E}_{\texttt{train}} := \frac{\delta^2}{n^2\kappa^2}\mathcal{E}_{\texttt{test}}$.

J.2 Completing the proof of Proposition 2.9

Throughout this section, we assume that we are in the situation of Proposition 2.9. Now, Simon et al. [2022] uses a different scaling for ridge regression than the one we use. We first resolve this discrepancy. Comparing Equation (14) with the expression in Lemma C.2, if we let $\delta := n\varrho$, then the expressions are equivalent, i.e., $\check{\beta}_\delta = \hat{\beta}_\varrho$. To see this, note that

$$\begin{aligned}\check{\beta}_\delta = \check{\beta}_{n\varrho} &= X(X^\top X + n\varrho \mathbb{I}_n)^{-1} y \\ &= (XX^\top + n\varrho \mathbb{I}_p)^{-1} Xy \quad \because \text{Lemma H.1} \\ &= (n(n^{-1}XX^\top + \varrho \mathbb{I}_p))^{-1} Xy \\ &= (\hat{\Sigma} + \varrho \mathbb{I}_p)^{-1} \tfrac{1}{n} Xy = \hat{\beta}_\varrho \quad \because \text{Definition of } \hat{\beta}_\varrho\end{aligned}$$

Furthermore, we claim that as $n \to \infty$, we have r, k satisfies Equation (3) if and only if $(\delta = nrn^{-\alpha}, \kappa = kn^{-\alpha})$ satisfies Equation (15):

$$n = \frac{\delta}{\kappa} + \sum_{i=1}^{p} \frac{\lambda_i}{\lambda_i + \kappa} \iff n = \frac{nrn^{-\alpha}}{kn^{-\alpha}} + \sum_{i=1}^{p} \frac{\lambda_i}{\lambda_i + kn^{-\alpha}} \iff 1 = \frac{r}{k} + \frac{1}{n} \sum_{i=1}^{p} \frac{1}{1 + kn^{-\alpha}\lambda_i^{-1}}.$$

Taking limit as $n \to \infty$, we have proved the claim.

Next, we show that $\lim_{n\to\infty} C = 0$ where C is as in Definition J.2. We have $\mathcal{L}_i := \frac{\lambda_i}{\lambda_i + \kappa} = \frac{1}{1 + k(i/n)^\alpha}$. Note that $\lim_{n\to\infty} \mathcal{L}_i = 1$ for all fixed i. On the other hand, since $\sup_{n=1,2\ldots} \|\beta^\star\|_2 < +\infty$, dominated convergence theorem implies that $\lim_{n\to\infty} C = 0$

We claim that the following asymptotic expression for the testing and training error hold:

$$\lim_{n\to\infty} \mathcal{E}_{\texttt{test}} = \sigma^2 \cdot \tfrac{dk}{dr} \quad \text{and} \quad \lim_{n\to\infty} \mathcal{E}_{\texttt{train}} = \sigma^2 \cdot \tfrac{r^2}{k^2} \cdot \tfrac{dk}{dr} \tag{16}$$

where r and k are defined as in Assumption B.6.

To see this, first note that the overfitting coefficient satisfies

$$\mathcal{E}_{\texttt{coef}} := n\tfrac{d\kappa}{d\delta} = n\tfrac{d\kappa}{d\varrho}\tfrac{d\varrho}{d\delta} = n\tfrac{d\kappa}{d\varrho}\tfrac{1}{n} = \tfrac{d\kappa}{d\varrho} = \tfrac{dk}{dr}.$$

Thus, we obtain the following asymptotic expression

$$\lim_{n\to\infty} \mathcal{E}_{\texttt{test}} = \mathcal{E}_{\texttt{coef}} \cdot \sigma^2 = \sigma^2 \cdot \tfrac{dk}{dr}.$$

On the other hand, the training error is given by

$$\mathcal{E}_{\texttt{train}} = \tfrac{\delta^2}{n^2\kappa^2}\mathcal{E}_{\texttt{test}} = \tfrac{\varrho^2}{\kappa^2}\mathcal{E}_{\texttt{test}} = \mathcal{E}_{\texttt{test}} \cdot \tfrac{r^2}{k^2}.$$

Therefore, $\lim_{n\to\infty} \mathcal{E}_{\texttt{train}} = \sigma^2 \cdot \frac{r^2}{k^2} \cdot \frac{dk}{dr}$. This proves (16), as desired.

Finally, we show that $\frac{d}{d\alpha}\mathcal{E}^*_{\texttt{test}} > 0$ for any $k > 0$. To this end, we use the expression derived in the previous step that $\mathcal{E}^*_{\texttt{test}} = \sigma^2 \cdot \frac{1}{1 - \mathcal{J}(k)}$. Taking derivative of both side w.r.t α, we have

$$\tfrac{d}{d\alpha}\mathcal{E}^*_{\texttt{test}} = \sigma^2 \tfrac{-1}{(1-\mathcal{J}(k))^2} \tfrac{d}{d\alpha}\mathcal{J}(k)$$

Now, we recall from Definition 2.8 that $\mathcal{J}(k) := \int_0^{1/\gamma_*} \frac{dx}{(1+kx^\alpha)^2}$. Thus, by differentiating under the integral sign, we have

$$\tfrac{d}{d\alpha}\mathcal{J}(k) = \int_0^{1/\gamma_*} \frac{-2kx^\alpha \log(x) dx}{(1 + kx^\alpha)^3}.$$

Putting it all together, we have

$$\tfrac{d}{d\alpha}\mathcal{E}^*_{\texttt{test}} = 2k\sigma^2 \tfrac{1}{(1-\mathcal{J}(k))^2} \int_0^{1/\gamma_*} \frac{x^\alpha \log(x) dx}{(1 + kx^\alpha)^3}.$$

Since the integrand is positive, the integral is positive as well. Moreover, since $k > 0$, we have $\frac{d}{d\alpha}\mathcal{E}^*_{\texttt{test}} > 0$ as desired. □

K Code

Implementation of the $\mathcal{I}$ and $\mathcal{J}$ functions from Definition 2.8:

```
import scipy.special as sc
gamma = 0.5
alpha = 1.75

# I generator
I_gen = lambda x,k, alpha: x*sc.hyp2f1(1,(1/alpha), 1 + (1/alpha), -k*x**alpha)
# J generator
J_gen = lambda x,k, alpha: x*sc.hyp2f1(2,(1/alpha), 1 + (1/alpha), -k*x**alpha)

I = lambda k : I_gen(1/gamma, k, alpha) #\mathcal{I}
J = lambda k : J_gen(1/gamma, k, alpha) #\mathcal{J}

N = lambda k : 1 - I(k) # helper
D = lambda k : 1 - J(k) # helper

Etst = lambda k : 1/D(k) #\mathcal{E}_{\mathtt{test}}/\sigma^2
Etrn = lambda k : N(k)**2/D(k) #\mathcal{E}_{\mathtt{train}}/\sigma^2
R  = lambda k : k*(1-I(k)) # \mathcal{R}
```

For the experiments in Figure 1-Right:

```
import numpy as np
gamma = 0.5
alpha = 1.75

k_grid = [ 1.34,  1.99,  2.45,  2.92,  3.44,  4.03,  4.71,  5.5 ,  6.44,
       7.55,  8.9 , 10.54, 12.58, 15.15, 18.46, 22.8 , 28.67, 36.87,
      48.82, 67.2 ]

n_tst = 1000
def get_norms(n,r):
    p = int(n/gamma)

    idx = np.arange(1,p+1) # feature indices

    pop_evs = idx**(-alpha) # population level eigenvalues

    X = np.multiply(np.sqrt(pop_evs[:,None]), np.random.normal(size= (p, n)) )
    X_tst = np.multiply(np.sqrt(pop_evs[:,None]), np.random.normal(size= (p, n_tst)) )

    beta_true = np.sqrt(10)*np.random.normal(size= (p,1))/np.sqrt(p)

    y = X.T@beta_true + np.random.normal(size= (n,1))
    y_tst =  X_tst.T@beta_true + np.random.normal(size= (n_tst,1))


    hatSig = (1/n)*X@X.T # sample covariance matrix

    beta = (1/n)*np.linalg.solve(hatSig + r*n**(-alpha)*np.eye(p), X@y)

    norm = np.linalg.norm(beta)**2
    Etrn = np.mean(np.square(y-X.T@beta))
    Etst = np.mean(np.square(y_tst-X_tst.T@beta))
    return {"norm": norm, "Etrn":Etrn, "Etst":Etst}
```

```
rs = R(np.array(k_grid))

n = 1000

Etrns = []
Etsts = []
for r in rs:
    result = get_norms(n,r)
    Etrns.append(result["Etrn"])
    Etsts.append(result["Etst"])
```

For the experiments in Figure 1:

```
# run the previous block first!
r = 3.5433549686341

ns = np.logspace(1,3.6,num=20)
categories = ["norm","Etrn","Etst"]
n_trials = 10

results = {cat : [[] for _ in range(n_trials)] for cat in categories}

for t in range(n_trials):
    for n in ns:
        out = get_norms(int(n),r)
        for cat in categories:
            results[cat][t].append(out[cat])
```