Under-Parameterized Double Descent for Ridge Regularized Least Squares Denoising of Data on a Line

Anonymous Author(s)
Affiliation
Address
email

Abstract

In this paper, we present a simple example that provably exhibits double descent in the under-parameterized regime. For simplicity, we look at the ridge regularized least squares denoising problem with data on a line embedded in high-dimension space. By deriving an asymptotically accurate formula for the generalization error, we observe sample-wise and parameter-wise double descent with the peak in the under-parameterized regime rather than at the interpolation point or in the over-parameterized regime. Further, the peak of the sample-wise double descent curve corresponds to a peak in the curve for the norm of the estimator, and adjusting μ, the strength of the ridge regularization, shifts the location of the peak. We observe that parameter-wise double descent occurs for this model for small μ. For larger values of μ, we observe that the curve for the norm of the estimator has a peak but that this no longer translates to a peak in the generalization error.

1 Introduction

This paper aims to demonstrate interesting new phenomena that suggest that our understanding of the relationship between the number of data points, the number of parameters, and the generalization error is incomplete, even for simple linear models with data on a line. The classical bias-variance theory postulates that the generalization risk versus the number of parameters for a fixed number of training data points is U-shaped. However, modern machine learning showed that if we keep increasing the number of parameters, the generalization error eventually starts decreasing again [1, 2]. This second descent has been termed as *double descent* and occurs in the *over-parameterized regime*, that is when the number of parameters exceeds the number of data points. Understanding the location and the cause of such peaks in the generalization error is of significant importance. Hence many recent works have theoretically studied the generalization error for linear regression [3–12] and kernelized regression [13–21] and show that there exists a peak at the boundary between the under and over-parameterized regimes. Further works such as [10, 22–25] show that there can be multiple descents in the over-parameterized regime and [26] shows that any shaped generalization error curve can occur in the over-parameterized regime. However, all prior works assume that the classical bias-variance trade-off is true in the under-parameterized regime.

The implicit bias of the learning algorithm is a possible reason that the error decreases in the over-parameterized regime [27–32]. In the under-parameterized regime, there is exactly one solution that minimizes the loss. However, once in the over-parameterized regime, there are many different solutions, and the training algorithm implicitly picks one that generalizes well. For linear models, the generalization error and the variance are very closely related to the norm of the estimator [11, 33]. Then, using the well-known fact that the pseudo-inverse solution to the least squares problem is the minimum norm solution, we see that the training algorithm picks solutions with the minimum norm. Hence this learning algorithm minimizes the variance and lowers the generalization error.

Table 1: Table showing various assumptions on the data and the location of the double descent peak for linear regression and denoising. We only present a subset of references for each problem setting.

Noise	Ridge Reg.	Dimension	Peak Location	Reference
Input	Yes	1	Under-parameterized	This paper.
Input	No	Low	Over-parameterized/interpolation point	[33, 37]
Output	No	Full	Over-parameterized/interpolation point	[5, 8, 11]
Output	Yes	Full	Over-parameterized/interpolation point	[11, 24]
Output	No	Low	Over-parameterized/interpolation point	[34, 35]
Output	Yes	Low	Over-parameterized/interpolation point	[36]

Main Contributions. In contrast with prior work, this paper shows that double descent can occur in the under-parameterized regime. Specifically, when denoising data on a line embedded in high-dimensional space using a denoiser obtained as the pseudo-inverse solution for the ridge regularized least squares problem, we show that a peak in the generalization error curve occurs in the under-parameterized regime. We also show that changing the ridge regularization strength changes the location of the peak. The major contributions of this paper are as follows.[1]

- (**Generalization error**) We derive a theoretical formula for the generalization error (Theorem 1).
- (**Under-parameterized double descent**) We prove (Theorem 2) and empirically demonstrate that the generalization error versus the number of data points curve has double descent in the under-parameterized regime.
- (**Location of the peak**) The peak location depends on the regularization strength. We provide evidence (Theorem 6) that the peak is near $c = \frac{1}{\mu^2+1}$ for the sample-wise double descent curves.
- (**Norm of the estimator**) We show that the peak in the curve for the generalization error versus the number of training data points corresponds to a peak in the norm of the estimator. However, versus the number of parameters, we show that there is still a peak in the curve for the norm of the estimator (Theorem 5), but this no longer corresponds to a peak in the generalization error.

Low-Dimensional Data. It is important to highlight that using low-rank data does not immediately imply that a peak occurs in the under-parameterized regime. Specifically, [33–37] look at a variety of different problems with low rank data and see that the peak occurs at the interpolation point or in the over-paramterized regime. Table 1 compares common assumptions and the location of the peak.

2 Background and Model Assumptions

Throughout the paper, we assume that noiseless training data x_i live in $\mathbb{R}^d$ and that we have access to a $d \times N_{trn}$ matrix X_{trn} of training data. Then given new data $X_{tst} \in \mathbb{R}^{d \times N_{tst}}$, we are interested in the least squares generalization (or test) error. Two scenarios for the generalization error curve are considered; data scaling and parameter scaling.

Definition 1. • Data scaling *refers to the regime in which we fix the dimension d of the input data and vary the number of training data points N_{trn}. This is also known as the sample-wise regime.*

- Parameter scaling *refers to the regime in which we fix the number of training data points N_{trn} and vary the dimension d of the input data. This is also known as the parameter-wise regime.*
- *A linear model is* under-parameterized, *if $d < N_{trn}$. A linear model is* over-parameterized, *if $d > N_{trn}$. The boundary of the under and over-parameterized regimes is when $d = N_{trn}$.*
- *Given N_{trn}, the* interpolation point *is the smallest d for the which the model has zero training error.*
- *A curve has* double descent *if the curve has a local maximum or peak.*
- *The* aspect ratio *of an $m \times n$ matrix is $c := m/n$.*

Prior Double Descent We present a *baseline model from prior work on double descent*. This is to highlight *prior* important phenomena related to double descent in the literature. Concretely, consider the following simple linear model that is a special case of the general models studied in [5, 8, 11, 24] amongst many other works. Let $x_i \sim \mathcal{N}(0, I_d)$ and let $\beta \in \mathbb{R}^d$ be a linear model with $\|\beta\| = 1$. Let $y_i = \beta^T x_i + \xi_i$ where $\xi \sim \mathcal{N}(0, 1)$. Then, let $\beta_{opt} := \arg\min_{\tilde{\beta}} \|\beta^T X_{trn} - \tilde{\beta} X_{trn} + \xi_{trn}\|$, where $\xi_{trn} \in \mathbb{R}^{N_{trn} \times 1}$. Then the excess risk, when taking the expectation over the new test data point, can be expressed as $\mathcal{R} = \|\beta - \beta_{opt}\|^2 = \|\beta\|^2 + \|\beta_{opt}\|^2 - 2\beta^T \beta_{opt}$. Let c be the aspect ratio of the

[1] All code is available anonymized at [Github Repo]

data matrix. That is, $c = d/N_{trn}$. Then it can be shown that[2]

$$\mathbb{E}_{X_{trn},\xi_{trn}}[\|\beta_{opt}\|^2] = \begin{cases} 1 + \frac{c}{1-c} & c < 1 \\ \frac{1}{c} + \frac{1}{c-1} & c > 1 \end{cases} \quad \text{and} \quad \mathbb{E}_{X_{trn},\xi_{trn}}[\beta^T \beta_{opt}] = \begin{cases} 1 & c < 1 \\ \frac{1}{c} & c > 1 \end{cases}$$

Then, the excess risk can be expressed as $\mathcal{R} = \begin{cases} \frac{c}{1-c} & c < 1 \\ \frac{c-1}{c} + \frac{1}{c-1} & c > 1 \end{cases}$. There are a few important features that are considered staple in many prior double descent curves that are present in this model.

1. The peak happens at $c = 1$, on the border between the under and over-parameterized regimes.
2. Further, at $c = 1$ the training error equals zero. Hence this is the interpolation point.
3. The peak occurs due to the norm of the estimator β_{opt} blowing up near the interpolation point.

Further, [26] proved risk is monotonic in the under-parameterized regime for the above model.

For the ridge regularized version of the regression problem, as shown in [11, 24], *the peak is always at* $c = 1$ (see Figure 1 in [24]). Further, as seen in Figure 1 in [24], changing the strength of the regularization changes the magnitude of the peak. *Not the location of the peak.* Building on this, [23] looks at the model where $y_i = f(x_i) + \xi_i$ and shows that triple descent occurs for the random features model [38] in the *over-parameterized* regime. Further [26] shows that by considering a variety of product data distributions, any shaped risk curve can be observed in the *over-parameterized* regime.

Assumptions for Denoising Model With the context from the previous section in mind, we are now ready to present the assumptions for the input noise model with double descent in the under-parameterized regime. For the denoising problem, let $A_{trn} \in \mathbb{R}^{d \times N_{trn}}$ be the noise matrix, then the ridge regularized least square denoiser W_{opt} is the minimum norm solution to

$$W_{opt} := \arg\min_{W} \|X_{trn} - W(X_{trn} + A_{trn})\|_F^2 + \mu^2 \|W\|_F^2. \tag{1}$$

Given test data X_{tst}, the mean squared generalization error is given by

$$\mathcal{R}(W_{opt}) = \mathbb{E}_{A_{trn},A_{tst}} \left[\frac{1}{N_{tst}} \|X_{tst} - W_{opt}(X_{tst} + A_{tst})\|_F^2 \right]. \tag{2}$$

The reason we consider linear models with the pseudo-inverse solution is that this eliminates other factors, such as the initialization of the network that could be a cause of the double descent [23]. We assume that the data lies on a line embedded in high-dimensional space.

Assumption 1. *Let $\mathcal{U} \subset \mathbb{R}^d$ be a one dimensional space with a unit basis vector u. Then let $X_{trn} = \sigma_{trn} u v_{trn}^T \in \mathbb{R}^{d \times N_{trn}}$ and $X_{tst} = \sigma_{tst} u v_{tst}^T \in \mathbb{R}^{d \times N_{tst}}$ be the respective SVDs for the training data and test data matrices. We further assume that $\sigma_{trn} = O(\sqrt{N_{trn}})$ and $\sigma_{tst} = O(\sqrt{N_{tst}})$.*

In [26], it was shown that by considering specific data distributions, any shaped generalization error curve could be observed in the over-parameterized regime. Hence to limit the effect of the data, we consider data on a line with norm restrictions.

Assumption 2. *The entries of the noise matrices $A \in \mathbb{R}^{d \times N}$ are I.I.D. from $\mathcal{N}(0, 1/d)$.*

Notational note. One final piece of technical notation is the following function definition.

$$\begin{aligned} p(\mu) := {} & (4\mu^{15} + 48\mu^{13} + 204\mu^{11} + 352\mu^9 + 192\mu^7)\sqrt{\mu^2 + 4} \\ & - (4\mu^{16} + 56\mu^{14} + 292\mu^{12} + 680\mu^{10} + 640\mu^8 + 128\mu^6). \end{aligned} \tag{3}$$

3 Under-Parameterized Regime Peak

We begin by providing a formula for the generalization error given by Equation 2 for the least squares solution given by Equation 1. The over-parameterized case can be found in Appendix F.2. See Appendix A for more discussion. All proofs are in Appendix F.

Theorem 1 (Generalization Error Formula)**.** *Suppose the training data X_{trn} and test data X_{tst} satisfy Assumption 1 and the noise A_{trn}, A_{tst} satisfy Assumption 2. Let μ be the regularization*

[2]The proofs are in Appendix F.1.

parameter. Then for the under-parameterized regime (i.e., $c < 1$) for the solution W_{opt} to Problem 1, the generalization error or risk given by Equation 2 is given by

$$\mathcal{R}(c, \mu) = \tau^{-2}\left(\frac{\sigma_{tst}^2}{N_{tst}} + \frac{c\sigma_{trn}^2(\sigma_{trn}^2 + 1))}{2d}\left(\frac{1 + c + \mu^2 c}{\sqrt{(1 - c + \mu^2 c)^2 + 4\mu^2 c^2}} - 1\right)\right) + o\left(\frac{1}{d}\right),$$

where $\tau^{-1} = \dfrac{2}{2 + \sigma_{trn}^2(1 + c + \mu^2 c - \sqrt{(1 - c + \mu^2 c) + 4\mu^2 c^2})}$.

Data Scaling. We prove that the risk curve in Theorem 1 has a peak for $c \in (0, 1)$. Theorem 2 tells us that under certain conditions, we are guaranteed to have a peak in the under-parameterized regime. This contrasts with prior work such as [3, 5, 8–11, 14, 25]. Further, we conjecture that the peak occurs near $c = (\mu^2 + 1)^{-1}$ (Appendix B). Figure 1 shows that the theoretically predicted risk matches the numerical risk. Moreover, the curve has a single peak for $c < 1$. Thus, *verifying that double descent occurs in the under-parameterized regime.* Finally, Figure 1 shows that the location of the peak is near the conjectured location of $\frac{1}{\mu^2+1}$. See Appendix D for the training error curves.

Theorem 2 (Under-Parameterized Peak)**.** *If $\mu \in \mathbb{R}_{>0}$ is such that $p(\mu) < 0$, $\sigma_{trn}^2 = N_{trn} = d/c$ and $\sigma_{tst}^2 = N_{tst}$, and d is sufficiently large, then the risk $\mathcal{R}(c)$ from Theorem 1, as a function of c, has a local maximum in the under-parameterized regime ($c \in (0, 1)$).*

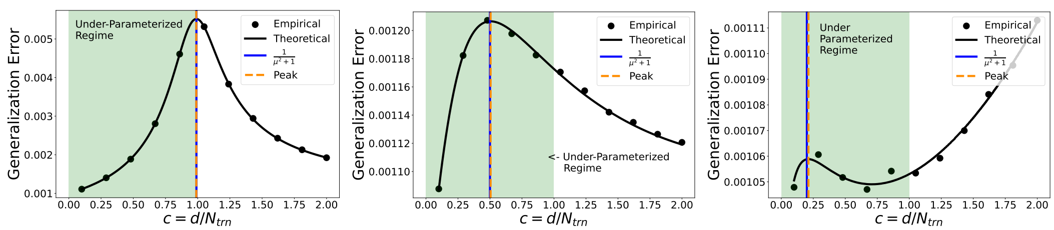
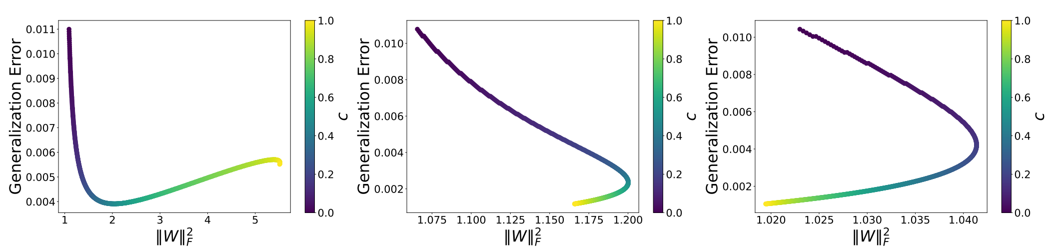


Figure 1: Figure showing the risk curve in the data scaling regime for different values of μ [(L) $\mu = 0.1$, (C) $\mu = 1$, (R) $\mu = 2$]. Here $\sigma_{trn} = \sqrt{N_{trn}}, \sigma_{tst} = \sqrt{N_{tst}}, d = 1000, N_{tst} = 1000$. For each empirical point, we ran at least 100 trials. More details can be found in Appendix G.



(a) $\mu = 0.1$ (b) $\mu = 1$ (c) $\mu = 2$

Figure 2: Figure showing generalization error versus $\|W_{opt}\|_F^2$ for the parameter scaling regime for three different values of μ. More details can be found in Appendix B.

Parameter Scaling. For many prior models, the data and parameter scaling regimes are analogous in that the shape of the risk is primarily governed by the aspect ratio c of the data matrix. However, we see significant differences between the parameter scaling and data scaling regimes for our setup. Figure 2 shows that for small values of μ, double descent occurs in the under-parameterized regime, for larger values of μ, the risk is monotonically decreasing.[3] Further, Figure 2 shows that for larger values of μ, *there is still a peak* in the curve for the norm of the estimator $\|W_{opt}\|_F^2$. However, this *does not* translate to a peak in the risk curve.

Theorem 3 ($\|W_{opt}\|_F$ Peak)**.** *If $\sigma_{tst} = \sqrt{N_{tst}}$, $\sigma_{trn} = \sqrt{N_{trn}}$ and μ is such that $p(\mu) < 0$, then for N_{trn} large enough and $d = cN_{trn}$, we have that $\|W_{opt}\|_F$ has a local maximum in the under-parameterized regime. Specifically for $c \in ((\mu^2 + 1)^{-1}, 1)$.*

[3]This is verified for more values of μ in Appendix B.

References

[1] Manfred Opper and Wolfgang Kinzel. Statistical Mechanics of Generalization. *Models of Neural Networks III: Association, Generalization, and Representation*, 1996 (cited on page 1).

[2] Mikhail Belkin, Daniel J. 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*, 2019 (cited on page 1).

[3] Madhu S. Advani, Andrew M. Saxe, and Haim Sompolinsky. High-dimensional Dynamics of Generalization Error in Neural Networks. *Neural Networks*, 2020 (cited on pages 1, 4, 9).

[4] Chen Cheng and Andrea Montanari. Dimension Free Ridge Regression. *arXiv preprint arXiv:2210.08571*, 2022 (cited on page 1).

[5] Edgar Dobriban and Stefan Wager. High-dimensional asymptotics of prediction: ridge regression and classification. *The Annals of Statistics*, 2018 (cited on pages 1, 2, 4, 9).

[6] Gabriel Mel and Surya Ganguli. A Theory of High Dimensional Regression with Arbitrary Correlations Between Input Features and Target Functions: Sample Complexity, Multiple Descent Curves and a Hierarchy of Phase Transitions. In *Proceedings of the 38th International Conference on Machine Learning*, 2021 (cited on page 1).

[7] Vidya Muthukumar, Kailas Vodrahalli, and Anant Sahai. Harmless Interpolation of Noisy Data in Regression. *2019 IEEE International Symposium on Information Theory (ISIT)*, 2019 (cited on page 1).

[8] Peter Bartlett, Philip M. Long, Gábor Lugosi, and Alexander Tsigler. Benign Overfitting in Linear Regression. *Proceedings of the National Academy of Sciences*, 2020 (cited on pages 1, 2, 4, 9).

[9] Mikhail Belkin, Daniel J. Hsu, and Ji Xu. Two Models of Double Descent for Weak Features. *SIAM Journal on Mathematics of Data Science*, 2020 (cited on pages 1, 4, 9).

[10] Michal Derezinski, Feynman T Liang, and Michael W Mahoney. Exact Expressions for Double Descent and Implicit Regularization Via Surrogate Random Design. In *Advances in Neural Information Processing Systems*, 2020 (cited on pages 1, 4, 9).

[11] Trevor Hastie, Andrea Montanari, Saharon Rosset, and Ryan J. Tibshirani. Surprises in High-Dimensional Ridgeless Least Squares Interpolation. *The Annals of Statistics*, 2022 (cited on pages 1–4, 9).

[12] Bruno Loureiro, Gabriele Sicuro, Cedric Gerbelot, Alessandro Pacco, Florent Krzakala, and Lenka Zdeborova. Learning Gaussian Mixtures with Generalized Linear Models: Precise Asymptotics in High-dimensions. In *Advances in Neural Information Processing Systems*, 2021 (cited on page 1).

[13] Song Mei, Theodor Misiakiewicz, and Andrea Montanari. Generalization Error of Random Feature and Kernel Methods: Hypercontractivity and Kernel Matrix Concentration. *Applied and Computational Harmonic Analysis*, 2022 (cited on page 1).

[14] Song Mei and Andrea Montanari. The Generalization Error of Random Features Regression: Precise Asymptotics and the Double Descent Curve. *Communications on Pure and Applied Mathematics*, 75, 2021 (cited on pages 1, 4, 9).

[15] Song Mei, Andrea Montanari, and Phan-Minh Nguyen. A Mean Field View of the Landscape of Two-layer Neural Networks. *Proceedings of the National Academy of Sciences of the United States of America*, 2018 (cited on page 1).

[16] Nilesh Tripuraneni, Ben Adlam, and Jeffrey Pennington. Covariate Shift in High-Dimensional Random Feature Regression. *ArXiv*, 2021 (cited on page 1).

[17] Federica Gerace, Bruno Loureiro, Florent Krzakala, Marc Mezard, and Lenka Zdeborova. Generalisation Error in Learning with Random Features and the Hidden Manifold Model. In *Proceedings of the 37th International Conference on Machine Learning*, 2020 (cited on page 1).

[18] Blake Woodworth, Suriya Gunasekar, Jason D. Lee, Edward Moroshko, Pedro Savarese, Itay Golan, Daniel Soudry, and Nathan Srebro. Kernel and Rich Regimes in Overparametrized Models. In *Proceedings of Thirty Third Conference on Learning Theory*, 2020 (cited on page 1).

[19] Bruno Loureiro, C'edric Gerbelot, Hugo Cui, Sebastian Goldt, Florent Krzakala, Marc M'ezard, and Lenka Zdeborov'a. Learning Curves of Generic Features Maps for Realistic Datasets with a Teacher-Student Model. In *NeurIPS*, 2021 (cited on page 1).

[20] Behrooz Ghorbani, Song Mei, Theodor Misiakiewicz, and Andrea Montanari. Limitations of Lazy Training of Two-layers Neural Network. In *Advances in Neural Information Processing Systems*, 2019 (cited on page 1).

[21] Behrooz Ghorbani, Song Mei, Theodor Misiakiewicz, and Andrea Montanari. When Do Neural Networks Outperform Kernel Methods? In *Advances in Neural Information Processing Systems*, 2020 (cited on page 1).

[22] Ben Adlam and Jeffrey Pennington. The Neural Tangent Kernel in High Dimensions: Triple Descent and a Multi-Scale Theory of Generalization. In *International Conference on Machine Learning*, 2020 (cited on page 1).

[23] Stéphane d'Ascoli, Levent Sagun, and Giulio Biroli. Triple Descent and the Two Kinds of Overfitting: Where and Why Do They Appear? In *Advances in Neural Information Processing Systems*, 2020 (cited on pages 1, 3).

[24] Preetum Nakkiran, Prayaag Venkat, Sham M. Kakade, and Tengyu Ma. Optimal Regularization can Mitigate Double Descent. In *International Conference on Learning Representations*, 2020 (cited on pages 1–3, 13).

[25] Behrooz Ghorbani, Song Mei, Theodor Misiakiewicz, and Andrea Montanari. Linearized Two-layers Neural Networks in High Dimension. *The Annals of Statistics*, 2021 (cited on pages 1, 4, 9).

[26] Lin Chen, Yifei Min, Mikhail Belkin, and Amin Karbasi. Multiple Descent: Design Your Own Generalization Curve. *Advances in Neural Information Processing Systems*, 2021 (cited on pages 1, 3).

[27] Alnur Ali, Edgar Dobriban, and Ryan J. Tibshirani. The Implicit Regularization of Stochastic Gradient Flow for Least Squares. In *International Conference on Machine Learning*, 2020 (cited on page 1).

[28] Arthur Jacot, Berfin Simsek, Francesco Spadaro, Clement Hongler, and Franck Gabriel. Implicit Regularization of Random Feature Models. In *Proceedings of the 37th International Conference on Machine Learning*, 2020 (cited on page 1).

[29] Ohad Shamir. The Implicit Bias of Benign Overfitting. In *Proceedings of Thirty Fifth Conference on Learning Theory*, 2022 (cited on page 1).

[30] Behnam Neyshabur. Implicit Regularization in Deep Learning. *ArXiv*, abs/1709.01953, 2017 (cited on page 1).

[31] Behnam Neyshabur, Srinadh Bhojanapalli, David Mcallester, and Nati Srebro. Exploring Generalization in Deep Learning. In *Advances in Neural Information Processing Systems*, 2017 (cited on page 1).

[32] Behnam Neyshabur, Ryota Tomioka, and Nathan Srebro. In Search of the Real Inductive Bias: On the Role of Implicit Regularization in Deep Learning. *CoRR*, abs/1412.6614, 2015 (cited on page 1).

[33] Rishi Sonthalia and Raj Rao Nadakuditi. Training data size induced double descent for denoising feedforward neural networks and the role of training noise. *Transactions on Machine Learning Research*, 2023 (cited on pages 1, 2, 13, 16, 19, 24).

[34] Ningyuan Huang, David W. Hogg, and Soledad Villar. Dimensionality reduction, regularization, and generalization in overparameterized regressions. *SIAM Journal on Mathematics of Data Science*, 2022 (cited on page 2).

[35] Ji Xu and Daniel J Hsu. On the number of variables to use in principal component regression. *Advances in neural information processing systems*, 2019 (cited on page 2).

[36] Denny Wu and Ji Xu. On the Optimal Weighted $\backslash ell$_2 Regularization in Overparameterized Linear Regression. *Advances in Neural Information Processing Systems*, 2020 (cited on page 2).

[37] Risi Sonthalia Chinmaya Kausik Kashvi Srivastva. Generalization Error without Independence: Denoising, Linear Regression, and Transfer Learning, 2023 (cited on page 2).

[38] Ali Rahimi and Benjamin Recht. Random Features for Large-Scale Kernel Machines. In *Advances in Neural Information Processing Systems*, 2007 (cited on page 3).

[39] Andrew Ng. CS229 Lecture notes. *CS229 Lecture notes*, 2000 (cited on page 10).

[40] Carl D. Meyer Jr. Generalized Inversion of Modified Matrices. *SIAM Journal on Applied Mathematics*, 1973 (cited on page 17).

[41] Friedrich Götze and Alexander Tikhomirov. The Rate of Convergence for Spectra of GUE and LUE Matrix Ensembles. *Central European Journal of Mathematics*, 2005 (cited on page 21).

[42] Friedrich Götze and Alexander Tikhomirov. Rate of Convergence to the Semi-Circular Law. *Probability Theory and Related Fields*, 2003 (cited on page 21).

[43] Friedrich Götze and Alexander Tikhomirov. Rate of Convergence in Probability to the Marchenko-Pastur Law. *Bernoulli*, 2004 (cited on page 21).

[44] Vladimir Marcenko and Leonid Pastur. Distribution of Eigenvalues for Some Sets of Random Matrices. *Mathematics of The Ussr-sbornik*, 1967 (cited on page 21).

[45] Z. Bai, Baiqi. Miao, and Jian-Feng. Yao. Convergence Rates of Spectral Distributions of Large Sample Covariance Matrices. *SIAM Journal on Matrix Analysis and Applications*, 2003 (cited on page 21).

Contents

A Under-Parameterized Regime Peak

We begin by providing a formula for the generalization error given by Equation 2 for the least squares solution given by Equation 1. All proofs are in Appendix F.

Theorem 1 (Generalization Error Formula)**.** *Suppose the training data X_{trn} and test data X_{tst} satisfy Assumption 1 and the noise A_{trn}, A_{tst} satisfy Assumption 2. Let μ be the regularization parameter. Then for the under-parameterized regime (i.e., $c < 1$) for the solution W_{opt} to Problem 1, the generalization error or risk given by Equation 2 is given by*

$$\mathcal{R}(c,\mu) = \tau^{-2}\left(\frac{\sigma_{tst}^2}{N_{tst}} + \frac{c\sigma_{trn}^2(\sigma_{trn}^2+1))}{2d}\left(\frac{1+c+\mu^2 c}{\sqrt{(1-c+\mu^2 c)^2+4\mu^2 c^2}} - 1\right)\right) + o\left(\frac{1}{d}\right),$$

where $\tau^{-1} = \dfrac{2}{2+\sigma_{trn}^2(1+c+\mu^2 c-\sqrt{(1-c+\mu^2 c)+4\mu^2 c^2})}$.

Since the focus is on the under-parameterized regime, Theorem 1 only presents the under-parameterized case. The over-parameterized case can be found in Appendix F.2.

Data Scaling. Looking at the formula in Theorem 1, the risk curve's shape is unclear. In this section, we prove that the risk curve in Theorem 1 has a peak for $c \in (0, 1)$. Theorem 2 tells us that under certain conditions, we are theoretically guaranteed to have a peak in the under-parameterized regime. This contrasts with prior work such as [3, 5, 8–11, 14, 25] where double descent occurs in the over-parameterized regime or on the boundary between the two regimes.

Theorem 2 (Under-Parameterized Peak)**.** *If $\mu \in \mathbb{R}_{>0}$ is such that $p(\mu) < 0$, $\sigma_{trn}^2 = N_{trn} = d/c$ and $\sigma_{tst}^2 = N_{tst}$, and d is sufficiently large, then the risk $\mathcal{R}(c)$ from Theorem 1, as a function of c, has a local maximum in the under-parameterized regime ($c \in (0, 1)$).*

Since the peak no longer occurs at $c = 1$, one important question is to determine the location of the peak. Theorem 6 provides a method for estimating the location of the peak.

Theorem 4 (Peak Location)**.** *If $\mu \in \mathbb{R}_{>0}$ is such that $p(\mu) < 0$, $\sigma_{trn}^2 = N_{trn} = d/c$ and $\sigma_{tst}^2 = N_{tst}$, then the partial derivative with respect to c of the risk $\mathcal{R}(c)$ from Theorem 1 can be written as*

$$\frac{\partial}{\partial c}\mathcal{R}(c,\mu) = \frac{(\mu^2 c + c - 1)P(c,\mu,T(c,\mu),d) + 4d\mu^2 c^2(2\mu^2 c - T(c,\mu))}{Q(c,\mu,T(c,\mu),d)},$$

where $T(c,\mu) = \sqrt{(1-c+\mu^2 c)^2+4\mu^2 c^2}$ and P, Q are polynomials in four variables.

Here, at $c = (\mu^2+1)^{-1}$, the first term in the numerator is zero. Hence we conjecture that the peak of the generalization error curve occurs near $c = (\mu^2+1)^{-1}$.

Remark 1. *Note that as $\mu \to 0$, we have that $4d\mu^2 c^2(2\mu^2 c - T(c,\mu)) \to 0$. We also note that, when $\mu = 1$, we have that $2c - T(c, 1) = 0$. Thus, we see that for μ near 0 or 1, we should expect our estimate of the location of the peak to be accurate.*

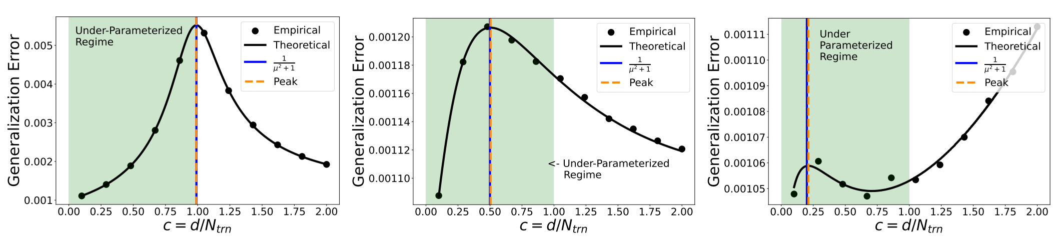
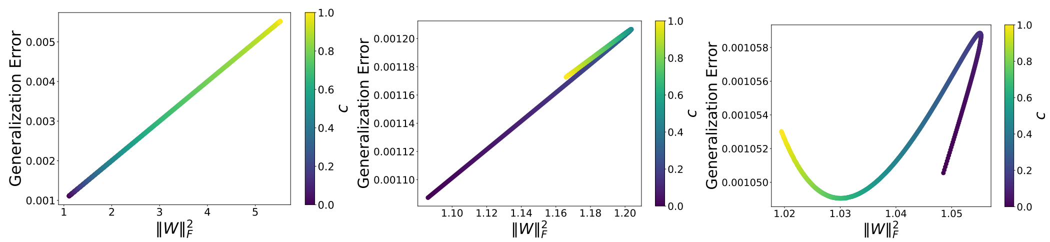


Figure 3: Figure showing the risk curve in the data scaling regime for different values of μ [(L) $\mu = 0.1$, (C) $\mu = 1$, (R) $\mu = 2$]. Here $\sigma_{trn} = \sqrt{N_{trn}}, \sigma_{tst} = \sqrt{N_{tst}}, d = 1000, N_{tst} = 1000$. For each empirical point, we ran at least 100 trials. More details can be found in Appendix G.

We numerically verify the predictions from Theorems 1, 2, 6. Figure 1 shows that the theoretically predicted risk matches the numerical risk. Moreover, the curve has a single peak for $c < 1$. Thus, *verifying that double descent occurs in the under-parameterized regime.* Finally, Figure 3 shows that the location of the peak is near the conjectured location of $\frac{1}{\mu^2+1}$. This conjecture is further tested



(a) $\mu = 0.1$ (b) $\mu = 1$ (c) $\mu = 2$

Figure 4: Figure showing generalization error versus $\|W_{opt}\|_F^2$ for the data scaling regime for three different values of μ. More details can be found in Appendix B and G.

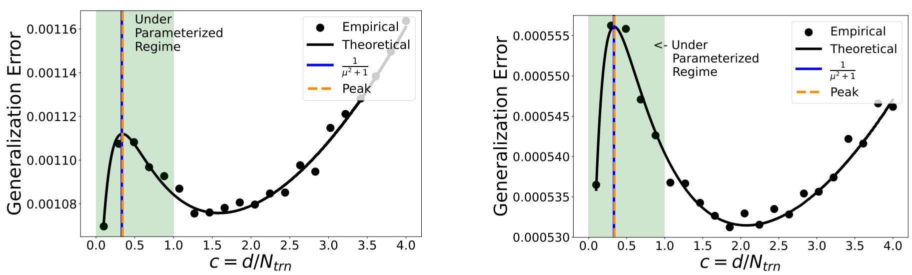


Figure 5: Figure showing that the shape of the risk curve in the data scaling regime depends on d [(L) $d = 1000$, (R) $d = 2000$]. Here $\mu = \sqrt{2}, \sigma_{trn} = \sqrt{N_{tst}}, \sigma_{trn} = \sqrt{N_{tst}}, N_{tst} = 1000$. Each empirical point is an average of at least 200 trials. More details can be found in Appendix G.

for a larger range of μ values in Appendix B. One similarity with prior work is that the peak in the generalization error or risk is corresponds to a peak in the norm of the estimator W_{opt} as seen in Figure 4 (i.e., the curve passes through the top right corner). The figure further shows, as conjectured in [39], that the double descent for the generalization error disappears when plotted as a function of $\|W_{opt}\|_F^2$ and, in some cases, recovers an approximation of the standard U shaped error curve.

Risk curve shape depends on d. Another interesting aspect of Theorem 2 is that it requires that d is large enough. Hence the shape of the risk curve depends on d. Most results for the risk are in the asymptotic regime. While Theorems 1, 2, and 6 are also in the asymptotic regime, we see that the results are accurate even for (relatively) small values of d, N_{trn}. Figure 5 shows that the shape of the risk curve depends on the value of d. Both curves still have a peak at the same location.

Parameter Scaling. For many prior models, the data scaling and parameter scaling regimes are analogous in that the shape of the risk curve does not depend on which one is scaled. The shape is primarily governed by the aspect ratio c of the data matrix. However, we see significant differences between the parameter scaling and data scaling regimes for our setup. Figure 6 shows risk curves that differ from those in Figure 3. Further, while for small values of μ, double descent occurs in the under-parameterized regime, for larger values of μ, the risk is monotonically decreasing.[4]

Even more astonishing, as shown in Figure 7, is the fact that for larger values of μ, *there is still a peak* in the curve for the norm of the estimator $\|W_{opt}\|_F^2$. However, this *does not* translate to a peak in the risk curve. Thus, showing that the norm of the estimator increasing cannot solely result in the generalization error increasing. The following theorem provides a local maximum in the $\|W_{opt}\|_F^2$ versus c curve for $c < 1$.

Theorem 5 ($\|W_{opt}\|_F$ Peak)**.** *If $\sigma_{tst} = \sqrt{N_{tst}}$, $\sigma_{trn} = \sqrt{N_{trn}}$ and μ is such that $p(\mu) < 0$, then for N_{trn} large enough and $d = cN_{trn}$, we have that $\|W_{opt}\|_F$ has a local maximum in the under-parameterized regime. Specifically for $c \in ((\mu^2+1)^{-1}, 1)$.*

[4] This is verified for more values of μ in Appendix B.

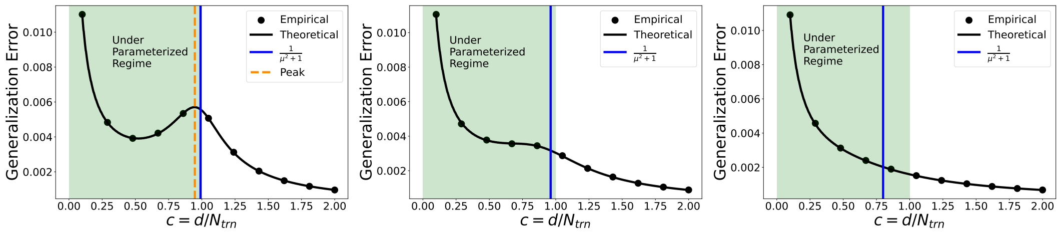
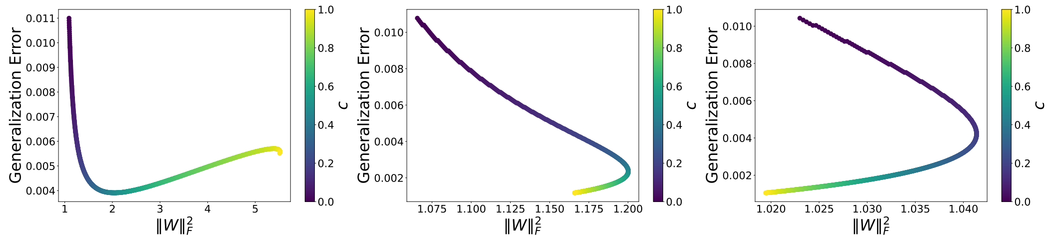


Figure 6: Figure showing the risk curves in the parameter scaling regime for different values of μ [(L) $\mu = 0.1$, (C) $\mu = 0.2$, (R) $\mu = 0.2$]. Here only the $\mu = 0.1$ has a local peak. Here $N_{trn} = N_{tst} = 1000$ and $\sigma_{trn} = \sigma_{tst} = \sqrt{1000}$. Each empirical point is an average of 100 trials.



(a) $\mu = 0.1$ (b) $\mu = 1$ (c) $\mu = 2$

Figure 7: Figure showing generalization error versus $\|W_{opt}\|_F^2$ for the parameter scaling regime for three different values of μ. More details can be found in Appendix B.

B Peak Location and $\|W_{opt}\|_F$

Theorem 6 (Peak Location)**.** *If $\mu \in \mathbb{R}_{>0}$ is such that $p(\mu) < 0$, $\sigma_{trn}^2 = N_{trn} = d/c$ and $\sigma_{tst}^2 = N_{tst}$, then the partial derivative with respect to c of the risk $\mathcal{R}(c)$ from Theorem 1 can be written as*

$$\frac{\partial}{\partial c}\mathcal{R}(c, \mu) = \frac{(\mu^2 c + c - 1)P(c, \mu, T(c, \mu), d) + 4d\mu^2 c^2 (2\mu^2 c - T(c, \mu))}{Q(c, \mu, T(c, \mu), d)},$$

where $T(c, \mu) = \sqrt{(1 - c + \mu^2 c)^2 + 4\mu^2 c^2}$ and P, Q are polynomials in four variables.

B.1 Peak Location for the Data Scaling Regime

We first look at the peak location conjecture for the data scaling regime. For this experiment, for 101 different values of $\mu \in [0.1, 10]$ we compute the generalization error at 101 equally spaced points for

$$c \in \left(\frac{1}{2(\mu^2 + 1)}, \frac{2}{\mu^2 + 1}\right).$$

We then pick the c value that has the maximum from amongst these 101 values of c. We notice that this did not happen at the boundary. Hence it corresponded to a true local maximum. We plot this value of c on Figure 8 and compare this against $\frac{1}{\mu^2+1}$. As we can see from Figure 8, our conjectured location of the peak is an accurate estimate.

C Generalization error - bias and variance

For both the data scaling and parameter scaling regimes, Figures 9 and 10 show the bias, $\|W_{opt}\|$ and the generalization error. Here we see that our estimate is accurate.

D Training Error

As seen in the prior section, the peak happens in the interior of the under-parameterized regime and not on the border between the under-parameterized and over-parameterized regimes. In many prior works, the peak aligns with the interpolation point (i.e., zero training error). Theorem 7 derives a formula for the training error in the under-parameterized regime. Figure 11 plots the location of

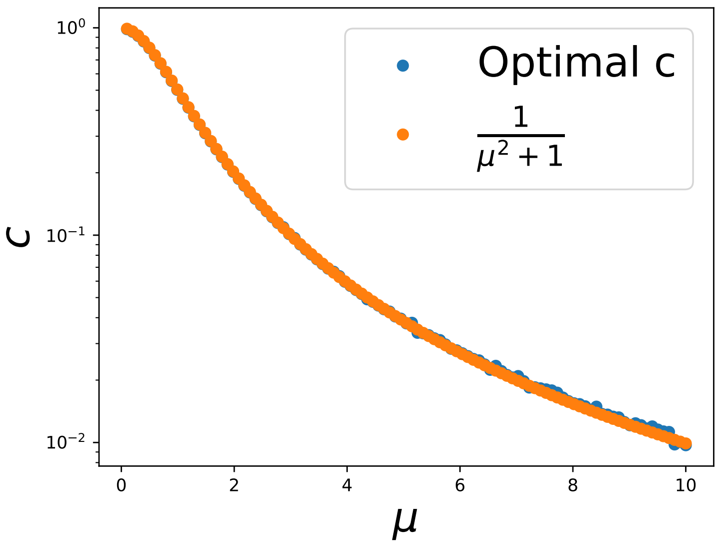
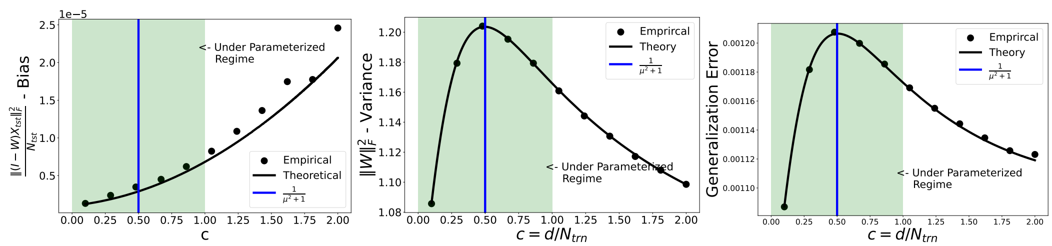
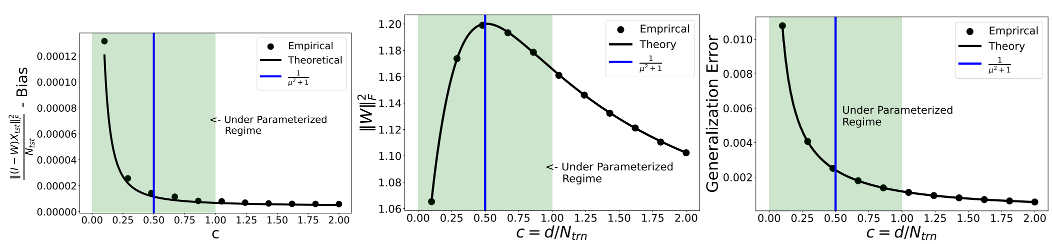


Figure 8: Figure showing the value of c where the peak occurs and the curve $1/(\mu^2+1)$



(a) Bias (b) $\|W_{opt}\|_F^2$ (c) Generalization Error

Figure 9: Figure showing the bias, $\|W_{opt}\|_F^2$, and the generalization error in the data scaling regime for $\mu = 1$, $\sigma_{trn} = \sqrt{N_{trn}}$, and $\sigma_{tst} = \sqrt{N_{tst}}$. Here $d = 1000$ and $N_{tst} = 1000$. For each empirical data point, we ran at least 100 trials. More details can be found in Appendix G.



(a) Bias (b) $\|W_{opt}\|_F^2$ (c) Generalization Error

Figure 10: Figure showing the $\|W_{opt}\|_F^2$, and the generalization error in the parameter scaling regime for $\mu = 1$, $\sigma_{trn} = \sqrt{N_{trn}}$, and $\sigma_{tst} = \sqrt{N_{tst}}$. Here $N_{trn} = 1000$ and $N_{tst} = 1000$. For each empirical data point, we ran at least 100 trials. More details can be found in Appendix G.

the peak, the training error, and the third derivative of the training error. Here the figure shows that the training error curve does not signal the location of the peak in the generalization error curve. However, it shows that for the data scaling regime, the peak roughly corresponds to a local minimum of the third derivative of the training error.

Theorem 7 (Training Error)**.** *Let τ be as in Theorem 1. The training error for $c < 1$ is given by*

$$\mathbb{E}_{A_{trn}}[\|X_{trn} - W_{opt}(X_{trn} + A_{trn})\|_F^2] = \tau^{-2}\left(\sigma_{trn}^2\left(1 - c \cdot T_1\right) + \sigma_{trn}^4 T_2\right) + o(1),$$

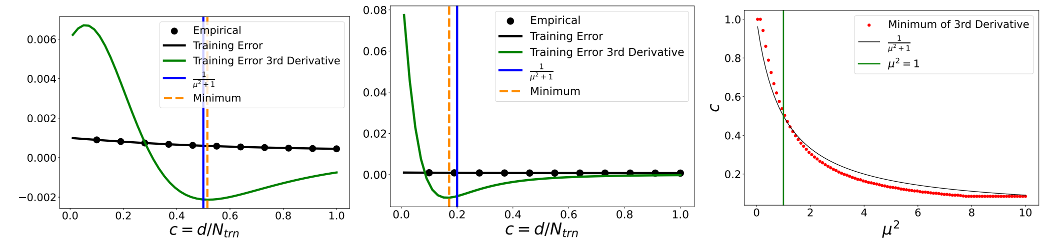


Figure 11: Figure showing the training error, the third derivative of the training error, and the location of the peak of the generalization error for different values of μ [(L) $\mu = 1$, (C) $\mu = 2$] for the data scaling regime. (R) shows the location of the local minimum of the third derivative and $\frac{1}{\mu^2+1}$.

where $T_1 = \dfrac{\mu^2}{2}\left(\dfrac{1+c+\mu^2 c}{\sqrt{(1-c+\mu^2 c)^2+4\mu^2 c^2}} - 1\right) + \dfrac{1}{2} + \dfrac{1+\mu^2 c - \sqrt{(1-c+\mu^2 c)^2 + 4c^2\mu^2}}{2c}$,

and $T_2 = \dfrac{(\mu^2 c + c - 1 - \sqrt{(1-c+\mu^2 c)^2+4c^2\mu^2})^2(\mu^2 c + c + 1 - \sqrt{(1-c+\mu^2 c)^2+4c^2\mu^2})}{2\sqrt{(1-c+\mu^2 c)^2+4c^2\mu^2}}$.

E Regularization Trade-off

We analyze the trade-off between the two regularizers and the generalization error.

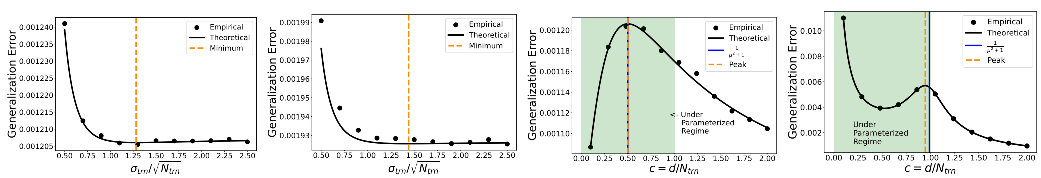


Figure 12: The first two figures show the σ_{trn} versus risk curve for $c = 0.5, \mu = 1$ and $c = 2, \mu = 0.1$ with $d = 1000$. The second two figures show the risk when training using the optimal σ_{trn} for the data scaling and parameter scaling regimes.

Optimal σ_{trn}. First, we fix μ and determine the optimal σ_{trn}. Figure 12 displays the generalization error versus σ_{trn}^2 curve. The figure shows that the error is initially large but then decreases until the optimal generalization error. The generalization error when using the optimal σ_{trn} is also shown in Figure 12. Here, unlike [24], picking the optimal value of σ_{trn} does not mitigate double descent.

Proposition 1 (Optimal σ_{trn})**.** *The optimal value of σ_{trn}^2 for $c < 1$ is given by*

$$\sigma_{trn}^2 = \frac{\sigma_{tst}^2 d[2c(\mu^2+1)^2 - 2T(c\mu^2+c+1) + 2(c\mu^2 - 2c + 1)] + N_{tst}(\mu^2 c^2 + c^2 + 1 - T)}{N_{tst}(c^3(\mu^2+1)^2 - T(\mu^2 c^2 + c^2 - 1) - 2c^2 - 1)}.$$

Additionally, it is interesting to determine how the optimal value of σ_{trn} depends on both μ and c. Figure 13 shows that for small values of μ (0.1,0.5), as c changes, there exists an (inverted) double descent curve for the optimal value of σ_{trn}. However, unlike [33], for the data scaling

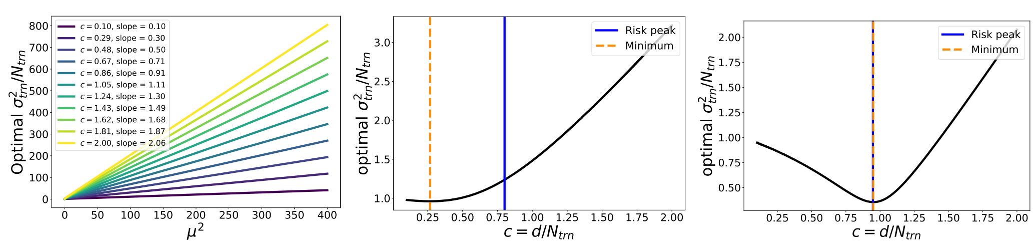


Figure 13: The first figure plots the optimal σ_{trn}^2/N_{trn} versus μ curve. The middle figure plots the optimal σ_{trn}^2/N_{trn} versus c in the data scaling regime for $\mu = 0.5$, and the last figure plots the optimal σ_{trn}^2/N_{trn} versus c in the parameter scaling regime for $\mu = 0.1$.

regime, the minimum of this double descent curve *does not match the location for the peak of the generalization error*. Further, as the amount of ridge regularization increases, the optimal amount of noise regularization decreases proportionally; optimal $\sigma_{trn}^2 \approx d\mu^2$. Thus, for higher values of ridge regularization, it is preferable to have higher-quality data.

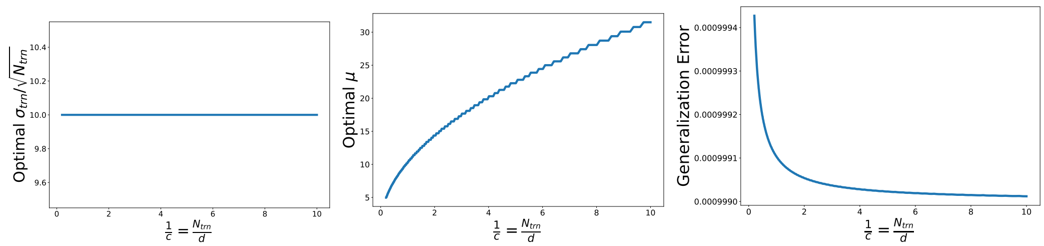
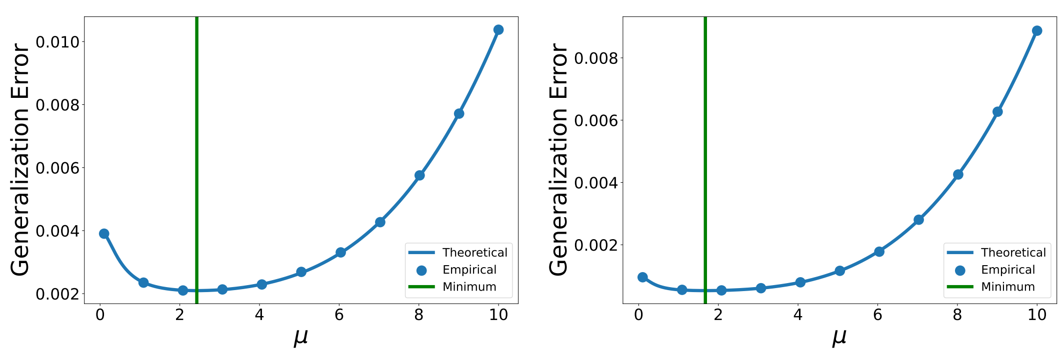


Figure 14: Trade-off between the regularizers. The left column is the optimal σ_{trn}, the central column is the optimal μ, and the right column is the generalization error for these parameter restrictions.

Interaction Between the Regularizers. The optimal values of μ and σ_{trn} are jointly computed using grid search for $\mu \in (0, 100]$ and $\sigma_{trn}/\sqrt{N_{trn}} \in (0, 10]$. Figure 14 shows the results. Specifically, σ_{trn} is at the highest possible value (so best quality data), and then the model regularizes purely using the ridge regularizer. This results in a monotonically decreasing generalization error curve. Thus, in the data scaling model, *there is an implicit bias that favors one regularizer over the other.* Specifically, the model's implicit bias *is to use higher quality data while using ridge regularization to regularize the model appropriately*. It is surprising that the two regularizers are not balanced.

E.1 Optimal Value of μ



(a) $c = 0.5$

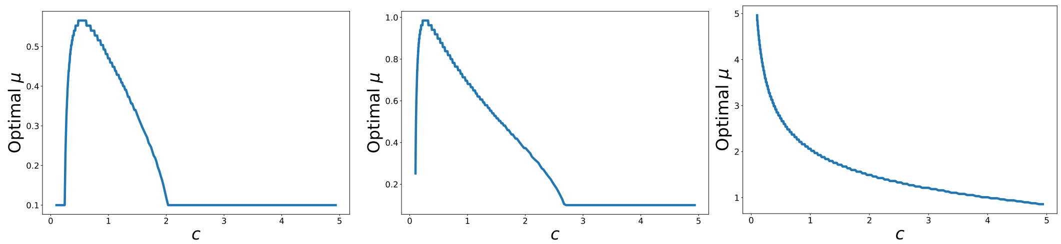
(b) $c = 2$

Figure 15: Figure showing the generalization error versus μ for $\sigma_{trn}^2 = N_{trn}$ and $\sigma_{tst}^2 = N_{tst}$.

We now explore the effect of fixing σ_{trn} and then changing μ. Figure 15, shows a U shaped curve for the generalization error versus μ, suggesting that there is an optimal value of μ, which should be used to minimize the generalization error.

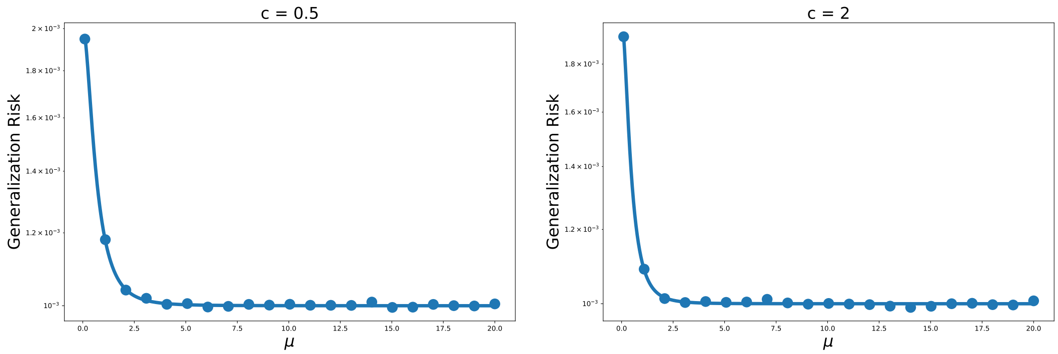
Next, we compute the optimal value of μ using grid search and plot it against c. Figure 16 shows double descent for the optimal value of μ for small values of σ_{trn}. Thus for low SNR data we see double descent, but we do not for high SNR data.

Finally, for a given value of μ and c, we compute the optimal σ_{trn}. We then compute the generalization error (when using the optimal σ_{trn}) and plot the generalization error versus μ curve. Figure 17 displays a very different trend from Figure 15. Instead of having a U-shaped curve, we have a monotonically decreasing generalization error curve. *This suggests that we can improve generalization by using higher-quality training while compensating for this by increasing the amount of ridge regularization.*



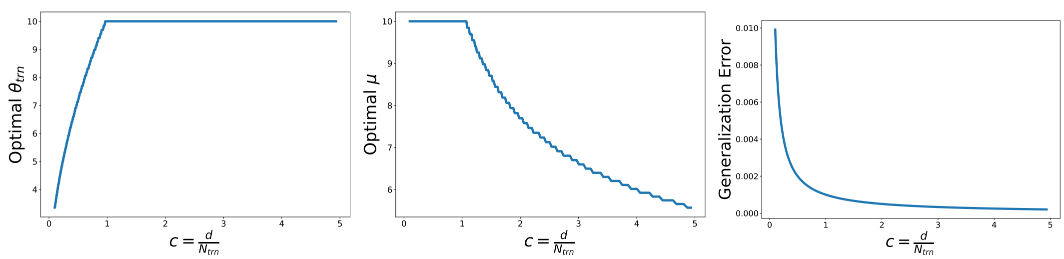
(a) $\sigma_{trn} \approx 0.2$ (b) $\sigma_{trn} \approx 0.3$ (c) $\sigma_{trn} \approx 1$

Figure 16: Figure for the optimal value of μ versus for different values of σ_{trn}

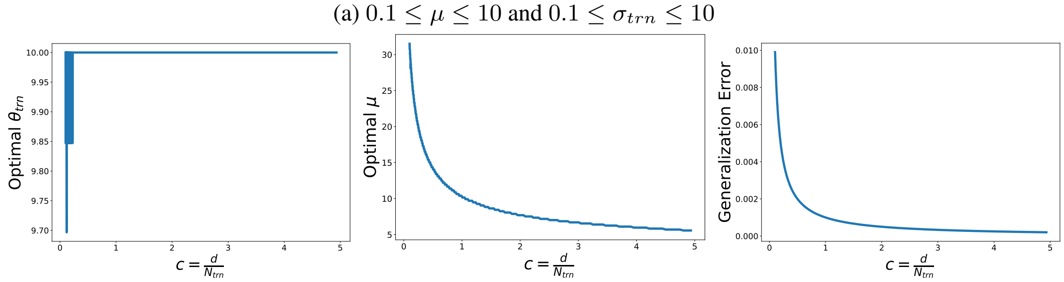


(a) $c = 0.5$ (b) $c = 2$

Figure 17: Figure showing the generalization error versus μ for the optimal σ_{trn}^2 and $\sigma_{tst}^2 = N_{tst}$.



(a) $0.1 \leq \mu \leq 10$ and $0.1 \leq \sigma_{trn} \leq 10$



(b) $0.1 \leq \mu \leq 100$ and $0.1 \leq \sigma_{trn} \leq 10$

Figure 18: Trade-off between the regularizers. The left column is the optimal σ_{trn}, the central column is the optimal μ, and the right column is the generalization error for these parameter restrictions

E.2 Trade-off in Parameter Scaling Regime

Here we look at the trade-off between σ_{trn} and μ for the parameter scaling regime. We again see that the model implicitly prefers regularizing via ridge regularization and not via input data noise regularizer.

F Proofs

F.1 Linear Regression

We begin by noting,

$$\beta^T = (\beta_{opt}^T X + \xi_{trn}) X_{trn}^\dagger.$$

Thus, we have,

$$\begin{aligned}\|\beta\|^2 &= \operatorname{Tr}(\beta^T \beta) \\ &= \operatorname{Tr}(\beta_{opt}^T X_{trn} X_{trn}^\dagger (X_{trn}^\dagger)^T X_{trn} \beta_{opt}) + \operatorname{Tr}(\xi_{trn} X_{trn}^\dagger (X_{trn}^\dagger)^T \xi_{trn}^T) + 2\operatorname{Tr}(\beta_{opt}^T X_{trn} X_{trn}^\dagger X_{trn}^\dagger)^T \xi_{trn}^T.\end{aligned}$$

Taking the expectation, with respect to ξ_{trn}, we see that the last term vanishes.

Letting $X_{trn} = U_X \Sigma_X V_X^T$. We see that using the rotational invariance of X, U_X, V_X are independent and uniformly random. Thus, $s := \beta_{opt}^T U_X$ is a uniformly random unit vector.

Thus, we see,

$$\mathbb{E}_{X_{trn},\xi_{trn}}\left[\operatorname{Tr}(\beta_{opt}^T X_{trn} X_{trn}^\dagger (X_{trn}^\dagger)^T X_{trn} \beta_{opt})\right] = \sum_{i=1}^{\min(d,N_{trn})} \mathbb{E}[s_i^2] = \min\left(1, \frac{1}{c}\right)$$

Similarly, we see,

$$\mathbb{E}_{X_{trn},\xi_{trn}}\left[\xi_{trn} X_{trn}^\dagger (X_{trn}^\dagger)^T \xi_{trn}^T\right] = \sum_{i=1}^{\min(d,N_{trn})} \mathbb{E}\left[\frac{1}{\sigma_i(X_{trn})^2}\right]$$

Multiplying and dividing by d, normalizes the singular values squared of X_{trn} so that the limiting distribution is the Marchenko Pastur distribution with shape c. Thus, we can estimate using Lemma 5 from Sonthalia and Nadakuditi [33] to get,

$$\begin{cases} \frac{c}{1-c} + o(1) & c < 1 \\ \frac{1}{c-1} + o(1) & c > 1 \end{cases}.$$

Finally, the cross-term has an expectation equal to zero. Thus,

$$\mathbb{E}_{X_{trn},\xi_{trn}}[\|\beta_{opt}\|^2] = \begin{cases} 1 + \frac{c}{1-c} & c < 1 \\ \frac{1}{c} + \frac{1}{c-1} & c > 1 \end{cases}$$

Then we have,

$$\beta^T \beta_{opt} = \beta_{opt}^T X_{trn} X_{trn}^\dagger \beta_{opt} + \xi_{trn} X_{trn}^\dagger \beta_{opt}$$

The second term has an expectation equal to zero, and the first term is similar to before and has an expectation equal to $\min\left(1, \dfrac{1}{c}\right)$.

F.2 Proofs for Theorem 1

The proof structure closely follows that of [33].

F.2.1 Step 1: Decompose the error into bias and variance terms.

First, we decompose the error. Since we are not in the supervised learning setup, we do not have standard definitions of bias/variance. However, we will call the following terms the bias/variance of the model. First, we recall the following from [33].

Lemma 1 (Sonthalia and Nadakuditi [33])**.** *If A_{tst} has mean 0 entries and A_{tst} is independent of X_{tst} and W, then*

$$\mathbb{E}_{A_{tst}}[\|X_{tst} - WY_{tst}\|_F^2] = \underbrace{\mathbb{E}_{A_{tst}}[\|X_{tst} - WX_{tst}\|_F^2]}_{Bias} + \underbrace{\mathbb{E}_{A_{tst}}[\|WA_{tst}\|_F^2]}_{Variance}. \tag{4}$$

F.2.2 Step 2: Formula for W_{opt}

Here, we compute the explicit formula for W_{opt} in Problem 1. Let $\hat{A}_{trn} = [A_{trn} \quad \mu I]$, $\hat{X}_{trn} = [X_{trn} \quad 0]$, and $\hat{Y}_{trn} = \hat{X}_{trn} + \hat{A}_{trn}$. Then solving $\arg\min_W \|X_{trn} - WY_{trn}\|_F^2 + \mu^2\|W\|_F^2$ is equivalent to solving $\arg\min_W \|\hat{X}_{trn} - W\hat{Y}_{trn}\|_F^2$. Thus, $W_{opt} = \arg\min_W \|\hat{X}_{trn} - W\hat{Y}_{trn}\|_F^2 = \hat{X}_{trn}\hat{Y}_{trn}^\dagger$. Expanding this out, we get the following formula for $\hat{W}$. Let $\hat{u}$ be the left singular vector and $\hat{v}_{trn}$ be the right singular vectors of $\hat{X}_{trn}$. Note that the left singular does not change after ridge regularization, so $\hat{u} = u$. Let $\hat{h} = \hat{v}_{trn}^T\hat{A}_{trn}^\dagger$, $\hat{k} = \hat{A}_{trn}^\dagger u$, $\hat{s} = (I - \hat{A}_{trn}\hat{A}_{trn}^\dagger)u$, $\hat{t} = \hat{v}_{trn}(I - \hat{A}_{trn}^\dagger\hat{A}_{trn})$, $\hat{\gamma} = 1 + \sigma_{trn}\hat{v}_{trn}^T\hat{A}_{trn}^\dagger u$, $\hat{\tau} = \sigma_{trn}^2\|\hat{t}\|^2\|\hat{k}\|^2 + \hat{\gamma}^2$.

Proposition 2. *If $\hat{\gamma} \neq 0$ and A_{trn} has full rank then*

$$W_{opt} = \frac{\sigma_{trn}\hat{\gamma}}{\hat{\tau}}u\hat{h} + \frac{\sigma_{trn}^2\|\hat{t}\|^2}{\hat{\tau}}u\hat{k}^T\hat{A}_{trn}^\dagger.$$

Proof. Here we know that u is arbitrary. We have that $\hat{A}_{trn}$ has full rank. Thus, the rank of $\hat{A}_{trn}$ is d, and the range of $\hat{A}_{trn}$ is the whole space. Thus, u lives in the range of $\hat{A}_{trn}$. In this case, we want Theorem 3 from [40]. We define

$$\hat{p} = -\frac{\sigma_{trn}^2\|\hat{k}\|^2}{\hat{\gamma}}\hat{t}^T - \sigma_{trn}\hat{k} \text{ and } \hat{q}^T = -\frac{\sigma_{trn}\|\hat{t}\|^2}{\hat{\gamma}}\hat{k}^T\hat{A}_{trn}^\dagger - \hat{h}.$$

Then we have,

$$(\hat{A}_{trn} + \sigma_{trn}u\hat{v}_{trn}^T)^\dagger = \hat{A}_{trn}^\dagger + \frac{\sigma_{trn}}{\hat{\gamma}}\hat{t}^T\hat{k}^T\hat{A}_{trn}^\dagger - \frac{\hat{\gamma}}{\hat{\tau}}\hat{p}\hat{q}^T.$$

Note that, by our assumptions, we have $\hat{t} = \hat{v}_{trn}(I - \hat{A}_{trn}^\dagger\hat{A}_{trn})$, and $(I - \hat{A}_{trn}^\dagger\hat{A}_{trn})$ is a projection matrix, thus

$$\begin{aligned}
\hat{v}_{trn}^T\hat{t}^T &= \hat{v}_{trn}^T(I - \hat{A}_{trn}^\dagger\hat{A}_{trn})^T\hat{v}_{trn}^T \\
&= \hat{v}_{trn}^T(I - \hat{A}_{trn}^\dagger\hat{A}_{trn})^T(I - \hat{A}_{trn}^\dagger\hat{A}_{trn})^T\hat{v}_{trn}^T.
\end{aligned}$$

To compute $W_{opt} = \hat{X}_{trn}(\hat{X}_{trn} + \hat{A}_{trn})^\dagger = \sigma_{trn}u\hat{v}_{trn}^T(\hat{A}_{trn} + \sigma_{trn}u\hat{v}_{trn}^T)^\dagger$, using $\hat{\gamma} - 1 = \sigma_{trn}\hat{v}_{trn}^T\hat{A}_{trn}^\dagger u = \sigma_{trn}\hat{h}u$, we multiply this through.

$$\begin{aligned}
\sigma_{trn}u\hat{v}_{trn}^T(\hat{A}_{trn} + \sigma_{trn}u\hat{v}_{trn}^T)^\dagger &= \sigma_{trn}u\hat{v}_{trn}^T\left(\hat{A}_{trn}^\dagger + \frac{\sigma_{trn}}{\hat{\gamma}}\hat{t}^T\hat{k}^T\hat{A}_{trn}^\dagger - \frac{\hat{\gamma}}{\hat{\tau}}\hat{p}\hat{q}^T\right) \\
&= \sigma_{trn}u\hat{h} + \frac{\sigma_{trn}^2\|\hat{t}\|^2}{\hat{\gamma}}u\hat{k}^T\hat{A}_{trn}^\dagger \\
&\quad + \frac{\sigma_{trn}\hat{\gamma}}{\hat{\tau}}u\hat{v}_{trn}^T\left(\frac{\sigma_{trn}^2\|\hat{k}\|^2}{\hat{\gamma}}\hat{t}^T + \sigma_{trn}\hat{k}\right)\hat{q}^T \\
&= \sigma_{trn}u\hat{h} + \frac{\sigma_{trn}^2\|\hat{t}\|^2}{\hat{\gamma}}u\hat{k}^T\hat{A}_{trn}^\dagger + \frac{\sigma_{trn}^3\|\hat{k}\|^2\|\hat{t}\|^2}{\hat{\tau}}u\hat{q}^T \\
&\quad + \frac{\sigma_{trn}\hat{\gamma}(\hat{\gamma} - 1)}{\hat{\tau}}u\hat{q}^T.
\end{aligned}$$

Then we have,

$$\begin{aligned}
\frac{\sigma_{trn}^3\|\hat{k}\|^2\|\hat{t}\|^2}{\hat{\tau}}u\hat{q}^T &= \frac{\sigma_{trn}^3\|\hat{k}\|^2\|\hat{t}\|^2}{\hat{\tau}}u\left(-\frac{\sigma_{trn}\|\hat{t}\|^2}{\hat{\gamma}}\hat{k}^T\hat{A}_{trn}^\dagger - \hat{h}\right) \\
&= -\frac{\sigma_{trn}^4\|\hat{k}\|^2\|\hat{t}\|^4}{\hat{\tau}\hat{\gamma}}u\hat{k}^T\hat{A}_{trn}^\dagger - \frac{\sigma_{trn}^3\|\hat{k}\|^2\|\hat{t}\|^2}{\hat{\tau}}u\hat{h}
\end{aligned}$$

and

$$\begin{aligned}\frac{\sigma_{trn}\hat{\gamma}(\hat{\gamma}-1)}{\hat{\tau}}u\hat{q}^T &= \frac{\sigma_{trn}\hat{\gamma}(\hat{\gamma}-1)}{\hat{\tau}}u\left(-\frac{\sigma_{trn}\|\hat{t}\|^2}{\hat{\gamma}}\hat{k}^T\hat{A}_{trn}^\dagger - \hat{h}\right)\\ &= -\frac{\sigma_{trn}^2\|\hat{t}\|^2(\hat{\gamma}-1)}{\hat{\tau}}u\hat{k}^T\hat{A}_{trn}^\dagger - \frac{\sigma_{trn}\hat{\gamma}(\hat{\gamma}-1)}{\hat{\tau}}u\hat{h}.\end{aligned}$$

Substituting back in and collecting like terms, we get,

$$\begin{aligned}\sigma_{trn}u\hat{v}_{trn}^T(\hat{A}_{trn}+\sigma_{trn}u\hat{v}_{trn}^T)^\dagger = \sigma_{trn}&\left(1-\frac{\sigma_{trn}^2\|\hat{k}\|^2\|\hat{t}\|^2}{\hat{\tau}}-\frac{\hat{\gamma}(\hat{\gamma}-1)}{\hat{\tau}}\right)u\hat{h}+\\ &\sigma_{trn}^2\left(\frac{\|\hat{t}\|^2}{\hat{\gamma}}-\frac{\sigma_{trn}^2\|\hat{k}\|^2\|\hat{t}\|^4}{\hat{\tau}\hat{\gamma}}-\frac{\|\hat{t}\|^2(\hat{\gamma}-1)}{\hat{\tau}}\right)u\hat{k}^T\hat{A}_{trn}^\dagger.\end{aligned}$$

We can then simplify the constants as follows.

$$1-\frac{\sigma_{trn}^2\|\hat{k}\|^2\|\hat{t}\|^2}{\hat{\tau}}-\frac{\hat{\gamma}(\hat{\gamma}-1)}{\hat{\tau}} = \frac{\hat{\tau}-\sigma_{trn}^2\|\hat{k}\|^2\|\hat{t}\|^2-\gamma^2+\gamma}{\hat{\tau}} = \frac{\hat{\gamma}}{\hat{\tau}}$$

and

$$\frac{\|\hat{t}\|^2}{\hat{\gamma}}-\frac{\sigma_{trn}^2\|\hat{k}\|^2\|\hat{t}\|^4}{\hat{\tau}\hat{\gamma}}-\frac{\|\hat{t}\|^2(\hat{\gamma}-1)}{\hat{\tau}} = \frac{\|\hat{t}\|^2\left(\hat{\tau}-\sigma_{trn}^2\|\hat{k}\|^2\|\hat{t}\|^2-\hat{\gamma}^2+\hat{\gamma}\right)}{\hat{\tau}\hat{\gamma}} = \frac{\|\hat{t}\|^2}{\hat{\tau}}.$$

This gives us the result. ☐

F.2.3 Step 3: Decompose the terms into a sum of various trace terms.

For the bias and variance terms, we have the following two Lemmas.

Lemma 2. *If W_{opt} is the solution to Equation 1, then*

$$X_{tst}-W_{opt}X_{tst} = \frac{\hat{\gamma}}{\hat{\tau}}X_{tst}.$$

Proof. To see this, note that we have $N_{trn}+M>M$.

$$\begin{aligned}X_{tst}-W_{opt}X_{tst} &= X_{tst}-\frac{\sigma_{trn}\hat{\gamma}}{\hat{\tau}}u\hat{h}uv_{tst}^T-\frac{\sigma_{trn}^2\|\hat{t}\|^2}{\hat{\tau}}u\hat{k}^T\hat{A}_{trn}^\dagger uv_{tst}^T\\ &= X_{tst}-\frac{\hat{\sigma}_{trn}\hat{\gamma}}{\hat{\tau}}u\hat{v}_{trn}^T\hat{A}_{trn}^\dagger uv_{tst}^T-\frac{\sigma_{trn}^2\|\hat{t}\|^2}{\hat{\tau}}u\hat{k}^T\hat{A}_{trn}^\dagger uv_{tst}^T.\end{aligned}$$

Note that $\hat{\gamma}=1+\sigma_{trn}\hat{v}_{trn}^T\hat{A}_{trn}^\dagger u$. Thus, we have that $\sigma_{trn}\hat{v}_{trn}^T\hat{A}_{trn}^\dagger u=\hat{\gamma}-1$. Substituting this into the second term, we get,

$$X_{tst}-W_{opt}X_{tst} = X_{tst}-\frac{\hat{\gamma}(\hat{\gamma}-1)}{\hat{\tau}}uv_{tst}^T-\frac{\sigma_{trn}^2\|\hat{t}\|^2}{\hat{\tau}}u\hat{k}^T\hat{A}_{trn}^\dagger uv_{tst}^T.$$

For the third term, since $\hat{k}=\hat{A}_{trn}^\dagger u$, $\hat{k}^T\hat{A}_{trn}^\dagger u=\hat{k}^T\hat{k}=\|\hat{k}\|^2$. Substituting this into the expression, we get that

$$X_{tst}-W_{opt}X_{tst} = X_{tst}-\frac{\hat{\gamma}(\hat{\gamma}-1)}{\hat{\tau}}uv_{tst}^T-\frac{\sigma_{trn}^2\|\hat{t}\|^2\|\hat{k}\|^2}{\hat{\tau}}uv_{tst}^T.$$

Since $X_{tst}=uv_{tst}^T$, we get,

$$X_{tst}-W_{opt}X_{tst} = X_{tst}\left(1-\frac{\hat{\gamma}(\hat{\gamma}-1)}{\hat{\tau}}-\frac{\sigma_{trn}^2\|\hat{t}\|^2\|\hat{k}\|^2}{\hat{\tau}}\right).$$

Simplify the constants using $\hat{\tau} = \sigma_{trn}^2 \|\hat{t}\|^2 \|\hat{k}\|^2 + \hat{\gamma}^2$, we get,

$$\frac{\hat{\tau} + \hat{\gamma} - \hat{\gamma}^2 - \sigma_{trn}^2 \|\hat{t}\|^2 \|\hat{k}\|^2}{\hat{\tau}} = \frac{\hat{\gamma}}{\hat{\tau}}.$$

□

Lemma 3 (Sonthalia and Nadakuditi [33]). *If the entries of A_{tst} are independent with mean 0, and variance $1/d$, then we have that $\mathbb{E}_{A_{tst}}[\|W_{opt} A_{tst}\|^2] = \frac{N_{tst}}{d} \|W_{opt}\|^2$.*

Lemma 4. *If $\hat{\gamma} \neq 0$ and A_{trn} has full rank, then we have that*

$$\|W_{opt}\|_F^2 = \frac{\sigma_{trn}^2 \hat{\gamma}^2}{\tau^2} \operatorname{Tr}(\hat{h}^T \hat{h}) + 2 \frac{\sigma_{trn}^3 \|\hat{t}\|^2 \hat{\gamma}}{\hat{\tau}^2} \operatorname{Tr}(\hat{h}^T \hat{k}^T \hat{A}_{trn}^\dagger) + \frac{\sigma_{trn}^4 \|\hat{t}\|^4}{\hat{\tau}^2} \underbrace{\operatorname{Tr}((\hat{A}_{trn}^\dagger)^T \hat{k} \hat{k}^T \hat{A}_{trn}^\dagger)}_{\rho}.$$

Proof. We have

$$\begin{aligned}
\|W_{opt}\|_F^2 &= \operatorname{Tr}(W_{opt}^T W_{opt}) \\
&= \operatorname{Tr}\left(\left(\frac{\sigma_{trn} \hat{\gamma}}{\hat{\tau}} u \hat{h} + \frac{\sigma_{trn}^2 \|\hat{t}\|^2}{\hat{\tau}} u \hat{k}^T \hat{A}_{trn}^\dagger \right)^T \left(\frac{\sigma_{trn} \hat{\gamma}}{\hat{\tau}} u \hat{h} + \frac{\sigma_{trn}^2 \|\hat{t}\|^2}{\hat{\tau}} u \hat{k}^T \hat{A}_{trn}^\dagger \right) \right) \\
&= \frac{\sigma_{trn}^2 \hat{\gamma}^2}{\hat{\tau}^2} \operatorname{Tr}(\hat{h}^T u^T u \hat{h}) + 2 \frac{\sigma_{trn}^3 \|\hat{t}\|^2 \hat{\gamma}}{\hat{\tau}^2} \operatorname{Tr}(\hat{h}^T u^T u \hat{k}^T \hat{A}_{trn}^\dagger) \\
&\quad + \frac{\sigma_{trn}^4 \|\hat{t}\|^4}{\hat{\tau}^2} \operatorname{Tr}((\hat{A}_{trn}^\dagger)^T \hat{k} u^T u \hat{k}^T \hat{A}_{trn}^\dagger) \\
&= \frac{\sigma_{trn}^2 \hat{\gamma}^2}{\hat{\tau}^2} \operatorname{Tr}(\hat{h}^T \hat{h}) + 2 \frac{\sigma_{trn}^3 \|\hat{t}\|^2 \hat{\gamma}}{\hat{\tau}^2} \operatorname{Tr}(\hat{h}^T \hat{k}^T \hat{A}_{trn}^\dagger) + \frac{\sigma_{trn}^4 \|\hat{t}\|^4}{\hat{\tau}^2} \operatorname{Tr}((\hat{A}_{trn}^\dagger)^T \hat{k} \hat{k}^T \hat{A}_{trn}^\dagger).
\end{aligned}$$

Where the last inequality is true due to the fact that $\|u\|^2 = 1$. □

F.2.4 Step 4: Estimate With Random Matrix Theory

Lemma 5. *Let A be a $p \times q$ matrix and let $\hat{A} = [A \quad \mu I] \in \mathbb{R}^{p \times q+p}$. Suppose $A = U \Sigma V^T$ be the singular value decomposition of A. If $\hat{A} = \hat{U} \hat{\Sigma} \hat{V}^T$ is the singular value decomposition of $\hat{A}$, then $\hat{U} = U$ and if $p < q$*

$$\hat{\Sigma} = \begin{bmatrix} \sqrt{\sigma_1(A)^2 + \mu^2} & 0 & \cdots & 0 \\ 0 & \sqrt{\sigma_2(A)^2 + \mu^2} & & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \cdots & \sqrt{\sigma_p(A)^2 + \mu^2} \end{bmatrix} \in \mathbb{R}^{p \times p},$$

and

$$\hat{V} = \begin{bmatrix} V_{1:p} \Sigma \hat{\Sigma}^{-1} \\ \mu U \hat{\Sigma}^{-1} \end{bmatrix} \in \mathbb{R}^{q+p \times p}.$$

Here $V_{1:p}$ are the first p columns of V.

Proof. Since $p < q$, we have that $U \in \mathbb{R}^{p \times p}$, $\Sigma \in \mathbb{R}^{p \times p}$ are invertible. Here also consider the form of the SVD in which $V^T \in \mathbb{R}^{p \times q}$.

We start by nothing that $\hat{U} \hat{\Sigma}^2 \hat{U}^T = \hat{A} \hat{A}^T = A A^T + \mu^2 I = U(\Sigma^2 + \mu^2 I_p) U^T$. Thus, we immediately see that $\sigma_i(\hat{A})^2 = \sigma_i(A)^2 + \mu^2$ and that $\hat{U} = U$.

Finally, we see,

$$\hat{V}^T = \hat{\Sigma}^{-1} U^T \hat{A} = \begin{bmatrix} \hat{\Sigma}^{-1} \Sigma V_{1:p}^T & \mu \hat{\Sigma}^{-1} U^T \end{bmatrix}$$

□

Lemma 6. *Let A be a $p \times q$ matrix and let $\hat{A} = [A \quad \mu I] \in \mathbb{R}^{p \times q+p}$. Suppose $A = U\Sigma V^T$ be the singular value decomposition of A. If $\hat{A} = \hat{U}\hat{\Sigma}\hat{V}^T$ is the singular value decomposition of $\hat{A}$, then $\hat{U} = U$ and if $p > q$*

$$\hat{\Sigma} = \begin{bmatrix} \sqrt{\sigma_1(A)^2+\mu^2} & 0 & \cdots & 0 & & \cdots & 0 \\ 0 & \sqrt{\sigma_2(A)^2+\mu^2} & & 0 & & & \\ \vdots & & \ddots & \vdots & & & \vdots \\ 0 & 0 & \cdots & \sqrt{\sigma_q(A)^2+\mu^2} & & & 0 \\ & & & & \mu & & \\ \vdots & & & & & \ddots & 0 \\ 0 & 0 & \cdots & 0 & \cdots & 0 & \mu \end{bmatrix} \in \mathbb{R}^{p \times p}.$$

Here we will denote the upper left $q \times q$ block by C. Further,

$$\hat{V} = \begin{bmatrix} V\Sigma_{1:q,1:q}^T C^{-1} & 0 \\ \mu U_{1:q} C^{-1} & U_{q+1:p} \end{bmatrix} \in \mathbb{R}^{q+p \times p}.$$

Proof. Since $p > q$, we have that $U \in \mathbb{R}^{p \times p}$ and we have that $\Sigma \in \mathbb{R}^{p \times q}$. Here $V^T \in \mathbb{R}^{q \times q}$ is invertible.

We start with nothing,

$$\hat{U}\hat{\Sigma}^2\hat{U}^T = \hat{A}\hat{A}^T = AA^T + \mu^2 I = U\left(\begin{bmatrix} \Sigma_{1:q,1:q}^2 & 0 \\ 0 & 0_{q-p} \end{bmatrix} + \mu^2 I_q\right)U^T.$$

Thus, we immediately see that for $i = 1, \ldots, p$ $\sigma_i(\hat{A})^2 = \sigma_i(A)^2 + \mu^2$ and for $i = p+1, \ldots, q$, we have that $\sigma_i(\hat{A})^2 = \mu^2$ and that $\hat{U} = U$.

Then, we see,

$$\hat{V}^T = \hat{\Sigma}^{-1}U^T\hat{A} = \begin{bmatrix} \hat{\Sigma}^{-1}\Sigma V^T & \mu\hat{\Sigma}^{-1}U^T \end{bmatrix}.$$

Note that Σ has 0 for the last $p - q$ entries. Thus,

$$\hat{\Sigma}^{-1}\Sigma V = \begin{bmatrix} C^{-1}\Sigma_{1:q,1:q}V \\ 0_{q-p,q} \end{bmatrix}.$$

Similarly, due to the structure of $\hat{\Sigma}$, we see,

$$\mu\hat{\Sigma}^{-1}U^T = [\mu C^{-1}U_{1:q}^T \quad \mu\frac{1}{\mu}U_{q+1:p}^T].$$

$\square$

Lemma 7. *Suppose A is an p by q matrix such that $p < q$, the entries of A are independent and have mean 0, variance $1/p$, and bounded fourth moment. Let $c = p/q$. Let $\hat{A} = [A \quad \mu I] \in \mathbb{R}^{p \times q+p}$. Let $W_p = \hat{A}\hat{A}^T$ and let $W_q = \hat{A}^T\hat{A}$. Suppose λ_p is a random non-zero eigenvalue from the largest p eigenvalues of W_p, and λ_q is a random non-zero eigenvalue of W_q. Then*

1. $\mathbb{E}\left[\frac{1}{\lambda_p}\right] = \mathbb{E}\left[\frac{1}{\lambda_q}\right] = \frac{\sqrt{(1+\mu^2c-c)^2+4\mu^2c^2}-1-\mu^2c+c}{2\mu^2c} + o(1).$
2. $\mathbb{E}\left[\frac{1}{\lambda_p^2}\right] = \mathbb{E}\left[\frac{1}{\lambda_q^2}\right] = \frac{\mu^2c^2+c^2+\mu^2c-2c+1}{2\mu^4c\sqrt{4\mu^2c^2+(1-c+\mu^2c)^2}} + \frac{1}{2\mu^4}\left(1-\frac{1}{c}\right) + o(1).$

Proof. First, we note that the non-zero eigenvalues of W_p and W_q are the same. Hence we focus on W_p. W_p is nearly a Wishart matrix but is not normalized by the correct value. However, cW_p does have the correct normalization.

Due to the assumptions on A, we have that the eigenvalues of cAA^T converge to the Marchenko-Pastur. Hence since the eigenvalues of cW_p are

$$(c\lambda_p)_i = c\sigma_i(A)^2 + c\mu^2,$$

we can estimate them by estimating $c\sigma_i(A)^2$ with the Marchenko-Pastur [41–45]. In particular, we want the expectation of the inverse. We need to use the Stieljes transform. We know that if $m_c(z)$ is the Stieljes transform for the Marchenko-Pastur with shape parameter c, then if λ is sampled from the Marchenko-Pastur distribution, then

$$m_c(z) = \mathbb{E}_\lambda \left[\frac{1}{\lambda - z}\right].$$

Thus, we have that the expected inverse of the eigenvalue can be approximated $m(-c\mu^2)$. We know that the Steiljes transform:

$$m_c(z) = -\frac{1 - z - c - \sqrt{(1-z-c)^2 - 4cz}}{-2zc}.$$

Thus, we have,

$$\mathbb{E}\left[\frac{1}{c\lambda_p}\right] = m(-c\mu^2) = \frac{\sqrt{(1+\mu^2 c - c)^2 + 4\mu^2 c^2} - 1 - \mu^2 c + c}{2\mu^2 c^2}.$$

Canceling $1/c$ from both sides, we get,

$$\mathbb{E}\left[\frac{1}{\lambda_p}\right] = \frac{\sqrt{(1+\mu^2 c - c)^2 + 4\mu^2 c^2} - 1 - \mu^2 c + c}{2\mu^2 c}.$$

Then for the estimate of $\mathbb{E}\left[1/\lambda_p^2\right]$, we need to compute the derivative of the $m_c(z)$ and evaluate it at $-c\mu^2$. Hence, we see,

$$m_c'(z) = \frac{(c - z + \sqrt{-4cz + (1-c-z)^2} - 1)(c + z + \sqrt{-4cz + (1-c-z)^2} - 1)}{4cz^2\sqrt{-4cz + (1-c-z)^2}}.$$

Thus,

$$\begin{aligned}\mathbb{E}\left[\frac{1}{c^2\lambda_p^2}\right] &= m_c'(-c\mu^2)\\ &= \frac{(c + \mu^2 c + \sqrt{4\mu^2c^2 + (1-c+\mu^2c)^2} - 1)(c - \mu^2 c + \sqrt{4\mu^2c^2 + (1-c+\mu^2c)^2} - 1)}{4\mu^4c^3\sqrt{4\mu^2c^2 + (1-c+\mu^2c)^2}}.\end{aligned}$$

Canceling the $1/c^2$ from both sides, we get,

$$\mathbb{E}\left[\frac{1}{\lambda_p^2}\right] = \frac{(c + \mu^2 c + \sqrt{4\mu^2c^2 + (1-c+\mu^2c)^2} - 1)(c - \mu^2 c + \sqrt{4\mu^2c^2 + (1-c+\mu^2c)^2} - 1)}{4\mu^4c\sqrt{4\mu^2c^2 + (1-c+\mu^2c)^2}}.$$

Multiplying out and simplifying

$$\mathbb{E}\left[\frac{1}{\lambda_p^2}\right] = \frac{\mu^2c^2 + c^2 + \mu^2 c - 2c + 1}{2\mu^4c\sqrt{4\mu^2c^2 + (1-c+\mu^2c)^2}} + \frac{1}{2\mu^4}\left(1 - \frac{1}{c}\right).$$

□

Lemma 8. *Suppose A is an p by q matrix such that $p > q$, the entries of A are independent and have mean 0, variance $1/p$, and bounded fourth moment. Let $c = p/q$. Let $\hat{A} = [A \quad \mu I] \in \mathbb{R}^{p\times q+p}$. Let $W_p = \hat{A}\hat{A}^T$ and let $W_q = \hat{A}^T\hat{A}$. Suppose λ_p is a random non-zero eigenvalue of W_p, and λ_q is a random eigenvalue from the largest q eigenvalues of W_q. Then*

1. $\mathbb{E}\left[\frac{1}{\lambda_q}\right] = \mathbb{E}\left[\frac{1}{\lambda_p}\right] = \frac{\sqrt{4\mu^2c + (1-c+\mu^2c)^2} - c - \mu^2 c + 1}{2\mu^2} + o(1).$
2. $\mathbb{E}\left[\frac{1}{\lambda_q^2}\right] = \mathbb{E}\left[\frac{1}{\lambda_p^2}\right] = \frac{1 - 2c + c^2 + \mu^2 c + \mu^2 c^2}{2\mu^4\sqrt{4\mu^2c + (-1+c+\mu^2c)^2}} + (1-c)\frac{1}{2\mu^4} + o(1).$

Proof. First, we note that the non-zero eigenvalues of W_p and W_q are the same. Hence we focus on W_p. Due to the assumptions on A, we have that the eigenvalues of A^TA converge to the Marchenko-Pastur with shape c^{-1}. Hence if λ_p is one of the first q eigenvalues of W_p, we see,

$$\mathbb{E}\left[\frac{1}{\lambda_p}\right] = m_{c^{-1}}(\mu^2) = \frac{\sqrt{(1+\mu^2-1/c)^2+4\mu^2/c}-1-\mu^2+1/c}{2\mu^2/c}.$$

Then for the estimate of $\mathbb{E}\left[1/\lambda_p^2\right]$, we need to compute the derivative of the $m_{c^{-1}}(z)$ and evaluate it at $-\mu^2$. Hence, we see,

$$\begin{aligned}
\mathbb{E}\left[\frac{1}{\lambda_p^2}\right] &= \frac{(1/c+\mu^2+\sqrt{4\mu^2/c+(1-1/c+\mu^2)^2}-1)(1/c-\mu^2+\sqrt{4\mu^2/c+(1-1/c+\mu^2)^2}-1)}{4\mu^4/c\sqrt{4\mu^2/c+(1-1/c+\mu^2)^2}} \\
&= \frac{(1+\mu^2c+c\sqrt{4\mu^2/c+(1-1/c+\mu^2)^2}-c)(1-\mu^2c+c\sqrt{4\mu^2/c+(1-1/c+\mu^2)^2}-c)}{4\mu^4c\sqrt{4\mu^2/c+(1-1/c+\mu^2)^2}} \\
&= \frac{(1+\mu^2c+\sqrt{4\mu^2c+(-1+c+\mu^2c)^2}-c)(1-\mu^2c+\sqrt{4\mu^2c+(-1+c+\mu^2c)^2}-c)}{4\mu^4\sqrt{4\mu^2c+(-1+c+\mu^2c)^2}}
\end{aligned}$$

This can be further simplified to

$$\frac{1-2c+c^2+\mu^2c+\mu^2c^2}{2\mu^4\sqrt{4\mu^2c+(-1+c+\mu^2c)^2}}+(1-c)\frac{1}{2\mu^4}+o(1)$$

□

We will also need to estimate some other terms.

Lemma 9. *Suppose A is an p by q matrix such that the entries of A are independent and have mean 0, variance $1/p$, and bounded fourth moment. Let $\hat{A} = [A \quad \mu I] \in \mathbb{R}^{p\times q+p}$. Let $W_p = \hat{A}\hat{A}^T$ and let $W_q = \hat{A}^T\hat{A}$. Suppose λ_p, λ_q are random non-zero eigenvalues of W_p, W_q from the largest $\min(p,q)$ eigenvalues of W_p, W_q. Then*

1. *If $p > q$, $\mathbb{E}\left[\frac{\lambda_p}{\lambda_p+\mu^2}\right] = c\left(\frac{1}{2}+\frac{1+\mu^2c-\sqrt{(-1+c+\mu^2c)^2+4\mu^2c}}{2c}\right)+o(1)$.*
2. *If $p < q$, $\mathbb{E}\left[\frac{\lambda_q}{\lambda_q+\mu^2}\right] = \frac{1}{2}+\frac{1+\mu^2c-\sqrt{(1-c+\mu^2c)^2+4c^2\mu^2}}{2c}+o(1)$.*
3. *If $p > q$, $\mathbb{E}\left[\frac{\lambda_p}{(\lambda_p+\mu^2)^2}\right] = c\left(\frac{1+c+\mu^2c}{2\sqrt{(-1+c+\mu^2c)^2+4\mu^2c}}-\frac{1}{2}\right)+o(1)$.*
4. *If $p < q$, $\mathbb{E}\left[\frac{\lambda_q}{(\lambda_q+\mu^2)^2}\right] = \frac{1+c+\mu^2c}{2\sqrt{(1-c+c\mu^2)^2+4c^2\mu^2}}-\frac{1}{2}+o(1)$.*

Proof. Notice that

$$\frac{\lambda}{\lambda+\mu^2} = 1-\frac{\mu^2}{\lambda+\mu^2} \text{ and } \frac{\lambda}{(\lambda+\mu^2)^2} = \frac{1}{\lambda+\mu^2}-\frac{\mu^2}{(\lambda+\mu^2)^2}$$

Then use Lemmas 7, and 8 to finish the proof. □

Bounding the Variance.

Lemma 10. *Let η_n be a uniform measure on n numbers $a_1, \ldots, a_n$ such that $\eta^n \to \eta$ weakly in probability. Then for any bounded continuous function f*

$$\frac{1}{n}\sum_{i=1}^{n-1} f(a_i) \to \mathbb{E}_{x\sim\eta}[f(x)].$$

Proof. Using weak convergence

$$\frac{1}{n}\sum_{i=1}^{n} f(a_i) \to \mathbb{E}_{x\sim\eta}[f(x)].$$

Then using the boundedness of f, we get,

$$\frac{1}{n}\sum_{i=1}^{n-1} f(a_i) - \frac{1}{n}\sum_{i=1}^{n} f(a_i) = -\frac{1}{n}f(a_n) \to 0.$$

□

Lemma 11. *Let η_n be a uniform measure on n numbers $a_1, \ldots, a_n$ such that $\eta_n \to \eta$ weakly in probability. Let s be a uniformly random unit vector in $\mathbb{R}^m$ independent of η_n. Suppose $n/m \to \zeta \in (0,1]$. Then for any bounded function f,*

$$\mathbb{E}_s\left[\sum_{i=1}^{n} s_i^2 f(a_i)\right] \to \zeta\mathbb{E}_{x\sim\eta}[f(x)]$$

and

$$\mathbb{E}_s\left[\left(\sum_{i=1}^{n} s_i^2 f(a_i)\right)^2\right] - \mathbb{E}_s\left[\sum_{i=1}^{n} s_i^2 f(a_i)\right]^2 \to 0.$$

Proof. The first limit comes directly from weak convergence.

For the second, notice,

$$\left(\sum_{i=1}^{n} s_i^2 f(a_i)\right)^2 = \sum_{i=1}^{n} s_i^4 f(a_i)^2 + \sum_{i\neq j} s_i^2 s_j^2 f(a_i)f(a_j) = \sum_{i=1}^{n} s_i^4 f(a_i)^2 + \sum_{i=1}^{n} s_i^2 f(a_i)\sum_{j\neq i} s_j^2 f(a_j).$$

Taking the expectation with respect to s we get,

$$\mathbb{E}_s\left[\left(\sum_{i=1}^{n} s_i^2 f(a_i)\right)^2\right] = \frac{1}{m^2+O(m)}\sum_{i=1}^{n} f(a_i)^2 + \frac{1}{m^2+O(m)}\sum_{i=1}^{n} f(a_i)\sum_{j\neq i} f(a_j)$$

Then using Lemma 10 for any fixed i, we have,

$$\frac{1}{m}\sum_{j\neq i} f(a_j) \to \zeta\mathbb{E}_{x\sim\eta}[f(x)].$$

Thus, as $n\to\infty$, we have,

$$\mathbb{E}_s\left[\left(\sum_{i=1}^{n} s_i^2 f(a_i)\right)^2\right] \to \zeta^2\mathbb{E}_{x\sim\eta}[f(x)]^2.$$

Then since

$$\mathbb{E}_s\left[\sum_{i=1}^{n} s_i^2 f(a_i)\right]^2 \to \zeta^2\mathbb{E}_{x\sim\eta}[f(x)]^2.$$

Thus, the variance goes to zero. □

The interpretation of the above Lemma is that the variance of the sum decays to zero as $m\to\infty$.

Lemma 12. *Suppose A is an p by q matrix such that the entries of A are independent and have mean 0, variance $1/p$, and bounded fourth moment. Let $\hat{A} = [A \quad \mu I] \in \mathbb{R}^{p\times q+p}$. Let $x\in\mathbb{R}^p$ and $\hat{y}\in\mathbb{R}^{p+q}$ be unit norm vectors such that $\hat{y}^T = [y^T \quad 0_p]$. Then*

1. *If $p<q$, then* $\mathbb{E}[\mathrm{Tr}(x^T(\hat{A}\hat{A}^T)^\dagger x] = \frac{\sqrt{(1-c+\mu^2c)^2+4\mu^2c^2}-1-\mu^2c+c}{2\mu^2c} + o(1)$.

2. *If $p > q$, then* $\mathbb{E}[\mathrm{Tr}(x^T(\hat{A}\hat{A}^T)^\dagger x] = \frac{\sqrt{(-1+c+\mu^2c)^2+4\mu^2c}-1-\mu^2c+c}{2\mu^2c} + o(1)$.

3. *If $p < q$, then* $\mathbb{E}[\mathrm{Tr}(\hat{y}^T(\hat{A}^T\hat{A})^\dagger \hat{y}] = c\left(\frac{1+c+\mu^2c}{2\sqrt{(1-c+\mu^2c)^2+4c^2\mu^2}} - \frac{1}{2}\right) + o(1)$.

4. *If $p > q$, then* $\mathbb{E}[\mathrm{Tr}(\hat{y}^T(\hat{A}^T\hat{A})^\dagger \hat{y}] = c\left(\frac{1+c+\mu^2c}{2\sqrt{(-1+c+\mu^2c)^2+4\mu^2c}} - \frac{1}{2}\right) + o(1)$.

The variance of each above is $o(1)$.

Proof. Let us start with $p < q$.

Let $\hat{A} = \hat{U}\hat{\Sigma}\hat{V}^T$, where $\hat{\Sigma}$ is $p \times p$. Then we see,

$$(\hat{A}\hat{A}^T)^\dagger = \hat{U}\hat{\Sigma}^{-2}\hat{U}^T.$$

Where $\hat{U}$ is uniformly random. Thus similar to [33], we can use Lemma 7 to get,

$$\mathbb{E}[\mathrm{Tr}(x^T(\hat{A}\hat{A}^T)^\dagger x] = \frac{\sqrt{(1+\mu^2c-c)^2+4\mu^2c^2}-1-\mu^2c+c}{2\mu^2c} + o(1).$$

On the other hand, for $p > q$, we have that only the first q eigenvalues have the expectation in Lemma 8 The other $p - q$ are equal to $\frac{1}{\mu^2}$. Thus, we see,

$$\begin{aligned}\mathbb{E}[\mathrm{Tr}(x^T(\hat{A}\hat{A}^T)^\dagger x] &= \frac{1}{c}\left(\frac{\sqrt{4\mu^2c+(-1+c+\mu^2c)^2}-c-\mu^2c+1}{2\mu^2}+o(1)\right)+\left(1-\frac{1}{c}\right)\frac{1}{\mu^2} \\ &= \frac{\sqrt{4\mu^2c+(-1+c+\mu^2c)^2}+c-\mu^2c-1}{2c\mu^2}.\end{aligned}$$

Again let us first consider the case when $p < q$. Then we have,

$$(\hat{A}^T\hat{A})^\dagger = \hat{V}\hat{\Sigma}^{-2}\hat{V}^T = \begin{bmatrix} V_{1:p}\Sigma\hat{\Sigma}^{-1} \\ \mu U\hat{\Sigma}^{-1}\end{bmatrix}\hat{\Sigma}^{-2}\begin{bmatrix}\hat{\Sigma}^{-1}\Sigma V_{1:p}^T & \mu\hat{\Sigma}^{-1}U^T\end{bmatrix}.$$

Since $\hat{y}$ has zeros in the last p coordinates, we see,

$$\hat{y}^T(\hat{A}^T\hat{A})^\dagger\hat{y} = y^T V_{1:p}\Sigma\hat{\Sigma}^{-4}\Sigma V_{1:p}^T y.$$

Thus, we can use Lemma 9 to estimate this as,

$$c\left(\frac{1+c+\mu^2c}{2\sqrt{(1-c+c\mu^2)^2+4c^2\mu^2}}-\frac{1}{2}\right)+o(1).$$

The extra factor of c comes from the sum of p coordinates of a uniformly unit vector in q dimensional space. And for $p > q$, we have that the estimate is

$$\frac{1+c+\mu^2c}{2\sqrt{(1+\mu^2-1/c)^2+4\mu^2/c}}-\frac{c}{2}+o(1).$$

For the variance term, use Lemma 11. For three of the cases, the limiting distribution is the Marchenko-Pastur distribution. For the other case, the limiting measure is a mixture of the Marchenko-Pastur and a dirac delta at $1/\mu^2$. □

The rest of the lemmas in this section are used to compute the mean and variance of the various terms that appear in the formula of W_{opt}.

Lemma 13. *We have that*

$$\mathbb{E}_{A_{trn}}\left[\|\hat{h}\|^2\right] = \begin{cases} c\left(\frac{1+c+\mu^2c}{2\sqrt{(1-c+\mu^2c)^2+4\mu^2c^2}}-\frac{1}{2}\right)+o(1) & c<1 \\ c\left(\frac{1+c+\mu^2c}{2\sqrt{(-1+c+\mu^2c)^2+4\mu^2c}}-\frac{1}{2}\right)+o(1) & c>1\end{cases}$$

and that $\mathbb{V}(\|\hat{h}\|^2) = o(1)$.

Proof. Here we see that
$$\|\hat{h}\|^2 = \mathrm{Tr}(\hat{v}_{trn}^T(\hat{A}_{trn}^T\hat{A}_{trn})^\dagger\hat{v}_{trn}^T).$$
Thus, using the Lemma 12 we get that if $c < 1$
$$\mathbb{E}[\|\hat{h}\|^2] = c\left(\frac{1+c+\mu^2 c}{2\sqrt{(1-c+\mu^2 c)^2+4\mu^2 c^2}} - \frac{1}{2}\right) + o(1)$$
and if $c > 1$
$$\mathbb{E}[\|\hat{h}\|^2] = c\left(\frac{1+c+\mu^2 c}{2\sqrt{(-1+c+\mu^2 c)^2+4\mu^2 c}} - \frac{1}{2}\right) + o(1).$$
□

Lemma 14. *We have*
$$\mathbb{E}_{A_{trn}}\left[\|\hat{k}\|^2\right] = \begin{cases} \frac{\sqrt{(1-c+\mu^2 c)^2+4\mu^2 c^2}-1-\mu^2 c+c}{2\mu^2 c} + o(1) & c < 1 \\ \frac{\sqrt{(-1+c+\mu^2 c)^2+4\mu^2 c}-1-\mu^2 c+c}{2\mu^2 c} + o(1) & c > 1 \end{cases}$$
and that $\mathbb{V}(\|\hat{k}\|^2) = o(1)$.

Proof. Since $\hat{k} = \hat{A}_{trn}^\dagger u$, we have that
$$\|\hat{k}\|^2 = \mathrm{Tr}(u^T(\hat{A}_{trn}\hat{A}_{trn}^T)^\dagger u).$$
According to the Lemma 12, if $c < 1$
$$\mathbb{E}[\|\hat{k}\|^2] = \frac{\sqrt{(1-c+\mu^2 c)^2+4\mu^2 c^2}-1-\mu^2 c+c}{2\mu^2 c} + o(1)$$
and if $c > 1$
$$\mathbb{E}[\|\hat{k}\|^2] = \frac{\sqrt{(-1+c+\mu^2 c)^2+4\mu^2 c}-1-\mu^2 c+c}{2\mu^2 c} + o(1).$$
□

Lemma 15. *We have that*
$$\mathbb{E}_{A_{trn}}\left[\|\hat{t}\|^2\right] = \begin{cases} \frac{1}{2}\left(1-c-\mu^2 c+\sqrt{(1-c+\mu^2 c)^2+4c^2\mu^2}\right) + o(1) & c < 1 \\ \frac{1}{2}\left(1-c-\mu^2 c+\sqrt{(-1+c+\mu^2 c)^2+4\mu^2 c}\right) + o(1) & c > 1 \end{cases}$$
and we have that $\mathbb{V}(\|\hat{t}\|^2) = o(1)$

Proof. Here we see that $\hat{t} = \hat{v}_{trn}(I - \hat{A}_{trn}^\dagger\hat{A}_{trn})$. Thus, we see that
$$\|\hat{t}\|^2 = \|v_{trn}\|^2 - \hat{v}_{trn}^T\hat{A}_{trn}^\dagger\hat{A}_{trn}\hat{v}_{trn} = 1 - \hat{v}_{trn}^T\hat{A}_{trn}^\dagger\hat{A}_{trn}\hat{v}_{trn}.$$
If $\hat{V} \in \mathbb{R}^{p+q\times p+q}$, we have that
$$\hat{A}_{trn}^\dagger\hat{A}_{trn} = \hat{V}\begin{bmatrix} I_p & 0 \\ 0 & 0_q \end{bmatrix}\hat{V}^T.$$
Then if $p < q$ using Lemma 6 and the fact that the last p coordinates of $\hat{v}_{trn}$ are 0, we see that
$$\hat{v}_{trn}^T\hat{A}_{trn}^\dagger\hat{A}_{trn}\hat{v}_{trn} = v_{trn}^T V_{1:p}\Sigma\hat{\Sigma}^{-2}\Sigma V_{1:p}^T v_{trn}.$$
Then using Lemma 9 to estimate the middle diagonal matrix, we get that
$$\begin{aligned}\mathbb{E}[\|\hat{t}\|^2] &= 1 - c\left(\frac{1}{2} + \frac{1+\mu^2 c - \sqrt{(1+\mu^2 c - c)^2 + 4c^2\mu^2}}{2c}\right) \\ &= \frac{1}{2}\left(1-c-\mu^2 c+\sqrt{(1-c+\mu^2 c)^2+4c^2\mu^2}\right) + o(1).\end{aligned}$$

Similarly for $c > 1$, we have that

$$\mathbb{E}[\|\hat{t}\|^2] = 1 - \left(\frac{1}{2} + \frac{c + \mu^2 c - c\sqrt{(1+\mu^2-1/c)^2 + 4\mu^2/c}}{2}\right) + o(1)$$
$$= \frac{1}{2}\left(1 - c - \mu^2 c + \sqrt{(-1+c+\mu^2 c)^2 + 4\mu^2 c}\right) + o(1).$$

The variance of $\hat{A}_{trn}^{\dagger}\hat{A}_{trn}$ is also $o(1)$ using Lemma 11. $\square$

Lemma 16. *We have that* $\mathbb{E}_{A_{trn}}[\hat{\gamma}] = 1$ *and* $\mathbb{V}(\gamma) = O(\sigma_{trn}^2/d)$.

Proof. Noting that $\hat{A} = U\hat{\Sigma}\hat{V}^T$, we have that

$$\hat{\gamma} = 1 + \sigma_{trn}\hat{v}_{trn}^T\hat{A}_{trn}^{\dagger}u = 1 + \sigma_{trn}\sum_{i=1}^{\min(N_{trn},d)}\sigma_i(\hat{A})^{-1}\hat{a}_i b_i.$$

Here $\hat{a}^T = \hat{v}_{trn}^T\hat{V}$ and $b = U^T u$. U is a uniformly random rotation matrix that is independent of $\hat{\Sigma}$ and $\hat{V}$. Thus, taking the expectation with respect to A_{trn}, we get that the expectation is equal to zero.

For the variance, let us first consider the case when $c < 1$. For this case, we have that

$$\hat{V} = \begin{bmatrix} V_{1:d}\Sigma\hat{\Sigma}^{-1} \\ \mu U\hat{\Sigma}^{-1} \end{bmatrix}.$$

Thus, letting $a^T = v_{trn}^T V_{1:d}$, we get that

$$\hat{\gamma} = 1 + \sum_{i=1}^{d}\frac{\sigma_i(A)}{\sigma_i^2(A) + \mu^2}a_i b_i.$$

Squaring and taking the expectation, we see that

$$\mathbb{E}[\gamma^2] = 1 + \frac{\sigma_{trn}^2}{N_{trn}}\mathbb{E}_{\lambda\sim\mu_c}\left[\frac{\lambda}{(\lambda+\mu^2)^2}\right] + o\left(\frac{\sigma_{trn}^2}{N_{trn}}\right).$$

Similarly for $c > 1$, we have that

$$\mathbb{E}[\gamma^2] = 1 + \frac{\sigma_{trn}^2}{d}\mathbb{E}_{\lambda\sim\mu_c}\left[\frac{\lambda}{(\lambda+\mu^2)^2}\right] + o\left(\frac{\sigma_{trn}^2}{d}\right).$$

$\square$

Lemma 17. *We have that*

$$\mathbb{E}\left[\mathrm{Tr}((\hat{A}_{trn}^{\dagger})^T\hat{k}\hat{k}^T\hat{A}_{trn}^{\dagger})\right] = \mathbb{E}[\rho] = \begin{cases} \frac{\mu^2c^2+c^2+\mu^2c-2c+1}{2\mu^4c\sqrt{4\mu^2c^2+(1-c+\mu^2c)^2}} + \frac{1}{2\mu^4}\left(1-\frac{1}{c}\right) + o(1) & c < 1 \\ \frac{1-2c+c^2+\mu^2c+\mu^2c^2}{2\mu^4c\sqrt{4\mu^2c+(-1+c+\mu^2c)^2}} + \left(1-\frac{1}{c}\right)\frac{1}{2\mu^4} + o(1) & \end{cases}$$

and that $\mathbb{V}(\rho) = o(1)$.

Proof. Here we have that

$$\rho = \mathrm{Tr}(\hat{k}^T(\hat{A}_{trn}^T\hat{A}_{trn})^{\dagger}\hat{k}) = \mathrm{Tr}(u^T(\hat{A}_{trn}\hat{A}_{trn}^T)^{\dagger}(\hat{A}_{trn}\hat{A}_{trn}^T)^{\dagger}u).$$

We first notice that

$$(\hat{A}_{trn}\hat{A}_{trn}^T)^{\dagger}(\hat{A}_{trn}\hat{A}_{trn}^T)^{\dagger} = \hat{U}^T\hat{\Sigma}^2\hat{U}.$$

Thus using Lemmas 7 and 8, we see that if $c < 1$

$$\mathbb{E}[\rho] = \frac{\mu^2c^2 + c^2 + \mu^2c - 2c + 1}{2\mu^4c\sqrt{4\mu^2c^2 + (1-c+\mu^2c)^2}} + \frac{1}{2\mu^4}\left(1 - \frac{1}{c}\right)$$

and if $c > 1$

$$\mathbb{E}[\rho] = \frac{1}{c}\left(\frac{1-2c+c^2+\mu^2c+\mu^2c^2}{2\mu^4\sqrt{4\mu^2c+(-1+c+\mu^2c)^2}} + (1-c)\frac{1}{2\mu^4}\right) + \left(1-\frac{1}{c}\right)\frac{1}{\mu^4}$$
$$= \frac{1-2c+c^2+\mu^2c+\mu^2c^2}{2\mu^4c\sqrt{4\mu^2c+(-1+c+\mu^2c)^2}} + \left(1-\frac{1}{c}\right)\frac{1}{2\mu^4}.$$

The variance being $o(1)$ comes from Lemma 11 again. $\square$

Lemma 18. *We have that*

$$\mathbb{E}_{A_{trn}}\left[\text{Tr}(\hat{h}^T\hat{k}^T\hat{A}_{trn}^{\dagger})\right] = 0$$

and the variance is $o(1)$.

Proof. Letting $\hat{A} = U\hat{\Sigma}\hat{V}^T$, we get that

$$\text{Tr}(\hat{h}^T\hat{k}^T\hat{A}^T) = u^T U\hat{\Sigma}^{-3}\hat{V}^T\hat{v}_{trn}^T.$$

Then again since U is uniformly random and independent of $\hat{\Sigma}$ and $\hat{V}$, the expectation is equal to zero. The variance is computed similarly to Lemma 16. □

F.2.5 Step 5: Putting it together

Lemma 19. *We have that*

$$\mathbb{E}\left[\frac{\tau}{\sigma_{trn}^2}\right] = \begin{cases} \frac{1}{\sigma_{trn}^2} + \frac{1}{2}\left(1+\mu^2c+c-\sqrt{(1-c+\mu^2c)^2+4\mu^2c^2}\right)+o(1) & c<1 \\ \frac{1}{\sigma_{trn}^2} + \frac{1}{2}\left(1+\mu^2c+c-\sqrt{(-1+c+\mu^2c)^2+4\mu^2c}\right)+o(1)) & c>1 \end{cases}$$

and that $\mathbb{V}(\tau/\sigma_{trn}^2) = o(1)$.

Proof. Using the fact that all of the quantities concentrate, we can use the previous estimates. Specifically, we use that

$$|\mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]| \le \sqrt{\mathbb{V}[X]\mathbb{V}[Y]}.$$

Thus, since our variances decay, we can use the product of the expectations. Further,

$$\begin{aligned} |\mathbb{V}[XY]| &= |\mathbb{V}[X]\mathbb{V}[Y] + \mathbb{E}[X]^2\mathbb{V}[Y] + \mathbb{E}[Y]^2\mathbb{V}[X] - 2\mathbb{E}[X]\mathbb{E}[Y]\text{Cov}(X,Y) + \text{Cov}(X^2,Y^2) - \text{Cov}(X,Y)^2| \\ &\le |\mathbb{V}[X]\mathbb{V}[Y] + \mathbb{E}[X]^2\mathbb{V}[Y] + \mathbb{E}[Y]^2\mathbb{V}[X]| + 2|\mathbb{E}[X]\mathbb{E}[Y]|\sqrt{\mathbb{V}[X]\mathbb{V}[Y]} + |\mathbb{V}[X]\mathbb{V}[Y]| + |\sqrt{\mathbb{V}[X^2]\mathbb{V}[Y^2]}|. \end{aligned}$$

Thus, since the variances individually go to 0, we see that the variance of the product also goes to 0. Then using Lemma 15 and 14, we have that if $c < 1$

$$\mathbb{E}\left[\|\hat{t}\|^2\|\hat{k}\|^2\right] = \frac{1}{2}\left(1+\mu^2c+c-\sqrt{(1-c+\mu^2c)^2+4\mu^2c^2}\right)+o(1)$$

and $\mathbb{V}(\|\hat{t}\|^2\|\hat{k}\|^2) = o(1)$. Then since

$$|\mathbb{V}[X+Y]| \le |\mathbb{V}[X] + \mathbb{V}[Y]| + 2\sqrt{\mathbb{V}[X]\mathbb{V}[Y]}$$

we have that using Lemma 16, that if $c < 1$

$$\mathbb{E}\left[\frac{\tau}{\sigma_{trn}^2}\right] = \frac{1}{\sigma_{trn}^2} + \frac{1}{2}\left(1+\mu^2c+c-\sqrt{(1-c+\mu^2c)^2+4\mu^2c^2}\right)+o(1)$$

and that that variance is $o(1)$. If $c > 1$

$$\mathbb{E}\left[\frac{\tau}{\sigma_{trn}^2}\right] = \frac{1}{\sigma_{trn}^2} + \frac{1}{2}\left(1+\mu^2c+c-\sqrt{(-1+c+\mu^2c)^2+4\mu^2c}\right)+o(1).$$

□

Lemma 20. *We have that*

$$\mathbb{E}_{A_{trn}}\left[\frac{1}{\sigma_{trn}^2}\|\hat{h}\|^2 + \|\hat{t}\|^4\rho\right] = \begin{cases} \frac{c(1+\sigma_{trn}^{-2})}{2}\left(\frac{\mu^2c+c+1}{\sqrt{(1-c+\mu^2c)^2+4\mu^2c^2}}-1\right)+o(1) & c<1 \\ \frac{c(1+\sigma_{trn}^{-2})}{2}\left(\frac{\mu^2c+c+1}{\sqrt{(-1+c+\mu^2c)^2+4\mu^2c}}-1\right)+o(1) & c>1 \end{cases}$$

and that the variance is $o(1)$.

Proof. Similar to Lemma 19, we can multiply the expectations since the variances are small. For $c < 1$, simplifying, we get that

$$\mathbb{E}_{A_{trn}}\left[\frac{1}{\sigma_{trn}^2}\|\hat{h}\|^2 + \|\hat{t}\|^4\rho\right] = \frac{c(1+\sigma_{trn}^{-2})}{2}\left(\frac{\mu^2c+c+1}{\sqrt{(1-c+\mu^2c)^2+4\mu^2c^2}} - 1\right) + o(1)$$

and if $c > 1$, we get that

$$\mathbb{E}_{A_{trn}}\left[\frac{1}{\sigma_{trn}^2}\|\hat{h}\|^2 + \|\hat{t}\|^4\rho\right] = \frac{c(1+\sigma_{trn}^{-2})}{2}\left(\frac{\mu^2c+c+1}{\sqrt{(-1+c+\mu^2c)^2+4\mu^2c}} - 1\right) + o(1)$$

and the variance decays since the variances decay individually. $\square$

Lemma 21. *We have that*

$$\mathbb{E}_{A_{trn}}\left[\|W_{opt}\|_F^2\right] = \frac{\sigma_{trn}^4}{\tau^2}\begin{cases}\frac{c(1+\sigma_{trn}^{-2})}{2}\left(\frac{\mu^2c+c+1}{\sqrt{(1-c+\mu^2c)^2+4\mu^2c^2}} - 1\right) + o(1) & c<1\\ \frac{c(1+\sigma_{trn}^{-2})}{2}\left(\frac{\mu^2c+c+1}{\sqrt{(-1+c+\mu^2c)^2+4\mu^2c}} - 1\right) + o(1) & c>1\end{cases}$$

and that $\mathbb{V}(\|W_{opt}\|_F^2) = o(1)$.

Proof. Follows immediately from Lemmas 4, 17, 18, and 20. $\square$

Theorem 1 (Generalization Error Formula)**.** *Suppose the training data X_{trn} and test data X_{tst} satisfy Assumption 1 and the noise A_{trn}, A_{tst} satisfy Assumption 2. Let μ be the regularization parameter. Then for the under-parameterized regime (i.e., $c < 1$) for the solution W_{opt} to Problem 1, the generalization error or risk given by Equation 2 is given by*

$$\mathcal{R}(c,\mu) = \tau^{-2}\left(\frac{\sigma_{tst}^2}{N_{tst}} + \frac{c\sigma_{trn}^2(\sigma_{trn}^2+1))}{2d}\left(\frac{1+c+\mu^2c}{\sqrt{(1-c+\mu^2c)^2+4\mu^2c^2}} - 1\right)\right) + o\left(\frac{1}{d}\right),$$

where $\tau^{-1} = \dfrac{2}{2+\sigma_{trn}^2(1+c+\mu^2c-\sqrt{(1-c+\mu^2c)+4\mu^2c^2})}$.

Proof. Rewriting $\frac{\hat{\gamma}^2}{\tau^2}$ as $\frac{\hat{\gamma}^2/\sigma_{trn}^4}{\tau^2/\sigma_{trn}^4}$, we can the concentration from Lemmas 16 and 19. Then using Lemma 21 we get the needed result. $\square$

Theorem 8. *For the over-parameterized case, we have that the generalization error is given by*

$$\mathcal{R}(c,\mu) = \tau^{-2}\left(\frac{\sigma_{tst}^2}{N_{tst}} + \frac{c\sigma_{trn}^2(\sigma_{trn}^2+1))}{2d}\left(\frac{1+c+\mu^2c}{\sqrt{(-1+c+\mu^2c)^2+4\mu^2c}} - 1\right)\right) + o\left(\frac{1}{d}\right),$$

where $\tau^{-1} = \dfrac{2}{2+\sigma_{trn}^2(1+c+\mu^2c-\sqrt{(-1+c+\mu^2c)+4\mu^2c})}$.

Proof. Rewriting $\frac{\hat{\gamma}^2}{\tau^2}$ as $\frac{\hat{\gamma}^2/\sigma_{trn}^4}{\tau^2/\sigma_{trn}^4}$, we can the concentration from Lemmas 16 and 19. Then using Lemma 21 we get the needed result. $\square$

F.3 Proof of Theorem 2

Theorem 2 (Under-Parameterized Peak)**.** *If $\mu \in \mathbb{R}_{>0}$ is such that $p(\mu) < 0$, $\sigma_{trn}^2 = N_{trn} = d/c$ and $\sigma_{tst}^2 = N_{tst}$, and d is sufficiently large, then the risk $\mathcal{R}(c)$ from Theorem 1, as a function of c, has a local maximum in the under-parameterized regime ($c \in (0, 1)$).*

Proof. First, we compute the derivative of the risk. We do so using SymPy and get the following expression.

$$\frac{\frac{\left(1+\frac{d}{c}\right)\left(\frac{\mu^2+1}{\sqrt{4c^2\mu^2+\left(c\mu^2-c+1\right)^2}}+\frac{\left(-4c\mu^2-\frac{\left(2\mu^2-2\right)\left(c\mu^2-c+1\right)}{2}\right)\left(c\mu^2+c+1\right)}{\left(4c^2\mu^2+\left(c\mu^2-c+1\right)^2\right)^{\frac{3}{2}}}\right)}{2}-\frac{d\left(-1+\frac{c\mu^2+c+1}{\sqrt{4c^2\mu^2+\left(c\mu^2-c+1\right)^2}}\right)}{2c^2}}{\left(1+\frac{d\left(c\mu^2+c-\sqrt{4c^2\mu^2+\left(c\mu^2-c+1\right)^2}+1\right)}{2c}\right)^2}$$

$$+\frac{\left(\frac{\left(-1+\frac{c\mu^2+c+1}{\sqrt{4c^2\mu^2+\left(c\mu^2-c+1\right)^2}}\right)\left(1+\frac{d}{c}\right)}{2}+1\right)\left(-\frac{d\left(\mu^2-\frac{4c\mu^2+\frac{\left(2\mu^2-2\right)\left(c\mu^2-c+1\right)}{2}}{\sqrt{4c^2\mu^2+\left(c\mu^2-c+1\right)^2}}+1\right)}{c}+\frac{d\left(c\mu^2+c-\sqrt{4c^2\mu^2+\left(c\mu^2-c+1\right)^2}+1\right)}{c^2}\right)}{\left(1+\frac{d\left(c\mu^2+c-\sqrt{4c^2\mu^2+\left(c\mu^2-c+1\right)^2}+1\right)}{2c}\right)^3}.$$

We can then compute the limit as $c \to 0^+$ and $c \to 1^-$. Again using SymPy we see that

$$\lim_{c\to 0^+} \frac{\partial}{\partial c}\mathcal{R}(c, \mu^2; \sigma_{trn}^2 = d/c) = \frac{4}{d+1} > 0.$$

Similarly, we can compute the limit as $c \to 1^-$ and get

$$\frac{2 \cdot \text{Expression}}{(\mu^4 + 4\mu^2)^{\frac{7}{2}} \left(d\mu^2 - d\mu\sqrt{\mu^2+4} + 2d + 2\right)^3}$$

where

$$\begin{aligned}
\text{Expression} = &-2d^2\mu^{16} + 2d^2\mu^{15}\sqrt{\mu^2+4} - 28d^2\mu^{14} + 24d^2\mu^{13}\sqrt{\mu^2+4} - 146d^2\mu^{12} \\
&+ 102d^2\mu^{11}\sqrt{\mu^2+4} - 340d^2\mu^{10} + 176d^2\mu^9\sqrt{\mu^2+4} - 320d^2\mu^8 \\
&+ 96d^2\mu^7\sqrt{\mu^2+4} - 64d^2\mu^6 - 2d\mu^{14} + 2d\mu^{13}\sqrt{\mu^2+4} - 26d\mu^{12} \\
&+ 30d\mu^{11}\sqrt{\mu^2+4} - 120d\mu^{10} + 144d\mu^9\sqrt{\mu^2+4} - 224d\mu^8 \\
&+ 224d\mu^7\sqrt{\mu^2+4} - 128d\mu^6 - 4\mu^{10} - 32\mu^8 - 64\mu^6.
\end{aligned}$$

Here using the arithmetic mean and geometric mean inequality, we see that

$$\mu^2 + 2 \geq \mu\sqrt{\mu^2+4}.$$

Thus, the denominator is always positive for $\mu > 0$. Thus, to determine the sign of the derivative, we need to determine the sign of the numerator. Here, we see that as a function of d, the numerator is a quadratic function of d, with the coefficient of d^2 is given by

$$\begin{aligned}
&(4\mu^{15} + 48\mu^{13} + 204\mu^{11} + 352\mu^9 + 192\mu^7)\sqrt{\mu^2+4} \\
&\qquad -(4\mu^{16} + 56\mu^{14} + 292\mu^{12} + 680\mu^{10} + 640\mu^8 + 128\mu^6).
\end{aligned}$$

We notice that this is exactly $p(\mu)$, which we assumed was negative. Thus, since the leading coefficient of the quadratic is negative, as $d \to \infty$, we have the quadratic, and hence the numerator, and hence the whole derivative is negative for sufficiently large d.

Finally, since the derivative near 0 is positive, and the derivative near 1 is negative, by the intermediate value theorem, there exists a value of $c \in (0, 1)$ such that the derivative value equals 0. Then since the derivative goes from positive to negative, this critical point corresponds to a local maximum. $\square$

F.4 Proof of Theorem 6

Theorem 6 (Peak Location)**.** *If $\mu \in \mathbb{R}_{>0}$ is such that $p(\mu) < 0$, $\sigma_{trn}^2 = N_{trn} = d/c$ and $\sigma_{tst}^2 = N_{tst}$, then the partial derivative with respect to c of the risk $\mathcal{R}(c)$ from Theorem 1 can be written as*

$$\frac{\partial}{\partial c}\mathcal{R}(c, \mu) = \frac{(\mu^2 c + c - 1)P(c, \mu, T(c, \mu), d) + 4d\mu^2 c^2(2\mu^2 c - T(c, \mu))}{Q(c, \mu, T(c, \mu), d)},$$

where $T(c, \mu) = \sqrt{(1 - c + \mu^2 c)^2 + 4\mu^2 c^2}$ and P, Q are polynomials in four variables.

Proof. To begin, we note that the derivative is,

$$\partial_c \mathcal{R}(c,\mu) = \frac{P(c,\mu,d,T)}{Q(c,\mu,d,T)}.$$

Where

$$\begin{aligned} P(c,\mu,d,T) =& -4T^2(-Tc^3d^2\mu^6 - 3Tc^3d^2\mu^4 - 3Tc^3d^2\mu^2 - Tc^3d^2 - Tc^3d\mu^4 \\ & - 5Tc^3d\mu^2 - 4Tc^3d - Tc^2d^2\mu^4 - Tc^2d^2\mu^2 - 2Tc^2d\mu^2 + 5Tc^2d + Tcd^2\mu^2 - Tcd \\ & + Td^2 + c^4d^2\mu^8 + 4c^4d^2\mu^6 + 6c^4d^2\mu^4 + 4c^4d^2\mu^2 + c^4d^2 + c^4d\mu^6 + 2c^4d\mu^4 \\ & + c^4d\mu^2 + 2c^4\mu^2 + 2c^4 + 2c^3d^2\mu^6 + 3c^3d^2\mu^4 - c^3d^2 + 3c^3d\mu^4 + 5c^3d - 2c^3 \\ & + 3c^2d\mu^2 - 6c^2d - 2cd^2\mu^2 + cd^2 + cd - d^2), \end{aligned}$$

$$Q(c,\mu,d,T) = T^7\left(-Td + cd\mu^2 + cd + 2c + d\right)^3,$$

and

$$T = \sqrt{c^2\mu^4 + 2c^2\mu^2 + c^2 + 2c\mu^2 - 2c + 1}.$$

Then if a critical point exists, it must be the case that $P(c,\mu,d,T) = 0$. This happens either if $T^2 = 0$ or $\hat{P} = P/(-4T^2) = 0$. Note we can simplify T^2 as

$$c^2(\mu^2+1)^2 + 2(\mu^2-1)c + 1$$

Then since this is a quadratic, we get that,

$$c = \frac{-2(\mu^2-1) \pm \sqrt{4(\mu^2-1)^2 - 4(\mu^2+1)^2}}{2(\mu^2+1)^2} = \frac{-2(\mu^2-1) \pm \sqrt{-16\mu^4}}{2(\mu^2+1)^2}.$$

Thus, the solutions live in $\mathbb{C}$ and not in $\mathbb{R}$. Since we want to find a root in $(0,1)$, we can discard this factor and focus on $\hat{P}$.

Looking at $\hat{P}$, we see that

$$\hat{P} = \hat{P}_1 + \hat{P}_2 + \hat{P}_3 + \hat{P}_4 + \hat{P}_5,$$

where

$$\begin{aligned} \hat{P}_1 &= -d^2T(\mu^2c + c - 1)(\mu^4c^2 + 2\mu^2c^2 + 2\mu^2c + c^2 + c + 1). \\ \hat{P}_2 &= -dTc(\mu^4c^2 + 5\mu^2c^2 + 2\mu^2c + 4c^2 - 5c + 1). \\ \hat{P}_3 &= d^2(\mu^2c + c - 1)(\mu^2c + c + 1)(\mu^4c^2 + 2\mu^2c^2 + 2\mu^2c + c^2 - c + 1). \\ \hat{P}_4 &= dc(\mu^6c^3 + 2\mu^4c^3 + 3\mu^4c^2 + \mu^2c^3 + 3\mu^2c + 5c^2 - 6c + 1). \\ \hat{P}_5 &= 2c^3(\mu^2c + c - 1). \end{aligned}$$

Here we see that $\mu^2c + c - 1$ is a factor for three of the five polynomials. Hence, the hope is that a multiple of $\mu^2c + c - 1$ can approximate the sum of the other two. Dividing $\hat{P}_2, \hat{P}_4$ by $\mu^2c + c - 1$, we get that

$$\begin{aligned} \hat{P}_2 &= -dTc(\mu^2c + c - 1)(\mu^2c + 4c - 1) - 4dT\mu^2c^2. \\ \hat{P}_4 &= dc(\mu^2c + c - 1)(\mu^4c^2 + \mu^2c^2 + 4\mu^2c - 1) - 3d\mu^2c^3 + 8d\mu^2c^2 + 5dc^3 - 5dC^2. \end{aligned}$$

Now we see that for some $\tilde{P}$

$$\hat{P} = (\mu^2c + c - 1)\tilde{P} - 4dT\mu^2c^2 - 3d\mu^2c^3 + 8d\mu^2c^2 + 5dc^3 - 5dC^2.$$

We further simplify this by dividing the remainder again by $\mu^2c + c - 1$ to get that

$$-4dT\mu^2c^2 - 3d\mu^2c^3 + 8d\mu^2c^2 + 5dc^3 - 5dc^2 = dc^2(\mu^2c + c - 1)(5\mu^2c - 8\mu^2) + 4d\mu^2c^2(2\mu^2c - T).$$

Thus, redefining $\tilde{P}$, we get that

$$\hat{P} = (\mu^2c + c - 1)\tilde{P} + 4d\mu^2c^2(2\mu^2c - T),$$

with

$$\begin{aligned} \tilde{P} =& -Tc^2d^2\mu^4 - 2Tc^2d^2\mu^2 - Tc^2d^2 - Tc^2d\mu^2 - 4Tc^2d - 2Tcd^2\mu^2 - Tcd^2 + Tcd \\ & - Td^2 + c^3d^2\mu^6 + 3c^3d^2\mu^4 + 3c^3d^2\mu^2 + c^3d^2 + c^3d\mu^4 + c^3d\mu^2 + 2c^3 \\ & + 3c^2d^2\mu^4 + 3c^2d^2\mu^2 - 4c^2d\mu^2 + 5c^2d + 3cd^2\mu^2 - cd + d^2. \end{aligned}$$

Thus, we have the needed result.

$\square$

F.5 Proof of Theorem 5

Theorem 5 ($\|W_{opt}\|_F$ Peak)**.** *If $\sigma_{tst} = \sqrt{N_{tst}}$, $\sigma_{trn} = \sqrt{N_{trn}}$ and μ is such that $p(\mu) < 0$, then for N_{trn} large enough and $d = cN_{trn}$, we have that $\|W_{opt}\|_F$ has a local maximum in the under-parameterized regime. Specifically for $c \in ((\mu^2+1)^{-1}, 1)$.*

Proof. Here we note that the expression for the norm of W_{opt} is given by Lemma 21. Differentiating with respect to c, we get that the derivative is given by

$$-\frac{c\sigma_{trn}^4\left(-1+\frac{c\mu^2+c+1}{\sqrt{4c^2\mu^2+(c\mu^2-c+1)^2}}\right)\left(\sigma_{trn}^2+1\right)\left(\mu^2-\frac{4c\mu^2+\frac{\left(2\mu^2-2\right)\left(c\mu^2-c+1\right)}{2}}{\sqrt{4c^2\mu^2+(c\mu^2-c+1)^2}}+1\right)}{2\left(\frac{\sigma_{trn}^2\left(c\mu^2+c-\sqrt{4c^2\mu^2+(c\mu^2-c+1)^2}+1\right)}{2}+1\right)^3}$$
$$+\frac{c\sigma_{trn}^2\left(\sigma_{trn}^2+1\right)\left(\frac{\mu^2+1}{\sqrt{4c^2\mu^2+(c\mu^2-c+1)^2}}+\frac{\left(-4c\mu^2-\frac{\left(2\mu^2-2\right)\left(c\mu^2-c+1\right)}{2}\right)\left(c\mu^2+c+1\right)}{\left(4c^2\mu^2+(c\mu^2-c+1)^2\right)^{\frac{3}{2}}}\right)}{2\left(\frac{\sigma_{trn}^2\left(c\mu^2+c-\sqrt{4c^2\mu^2+(c\mu^2-c+1)^2}+1\right)}{2}+1\right)^2}$$
$$+\frac{\sigma_{trn}^2\left(-1+\frac{c\mu^2+c+1}{\sqrt{4c^2\mu^2+(c\mu^2-c+1)^2}}\right)\left(\sigma_{trn}^2+1\right)}{2\left(\frac{\sigma_{trn}^2\left(c\mu^2+c-\sqrt{4c^2\mu^2+(c\mu^2-c+1)^2}+1\right)}{2}+1\right)^2}.$$

At $c = \frac{1}{\mu^2+1}$, this has value

$$\frac{2\sigma_{trn}^2\left(\mu^2+1\right)^{\frac{3}{2}}\left(-256\mu^7+256\mu^6\sqrt{\mu^2+1}\right)\left(\sigma_{trn}^2+1\right)}{\left(\frac{4\mu^4}{\left(\mu^2+1\right)^2}+\frac{4\mu^2}{\left(\mu^2+1\right)^2}\right)^{\frac{7}{2}}\left(-2\mu\sigma_{trn}^2+2\sigma_{trn}^2\sqrt{\mu^2+1}+2\sqrt{\mu^2+1}\right)^3\left(\mu^6\sqrt{\mu^2+1}+3\mu^4\sqrt{\mu^2+1}+3\mu^2\sqrt{\mu^2+1}+\sqrt{\mu^2+1}\right)}.$$

Then since $\sqrt{\mu^2+1} > \mu$, we have that the derivative is positive at this point. Next, we compute the limit of the derivative as $c \to 1^-$ and see that this is given by

$$\frac{\sigma_{trn}^2\left(\sigma_{trn}^2+1\right)\left(\sigma_{trn}^2 p(\mu)+4\mu^{14}+56\mu^{12}+280\mu^{10}+576\mu^8+384\mu^6-(4\mu^{13}+48\mu^{11}+192\mu^9+256\mu^7)\sqrt{\mu^2+4}\right)}{(\mu^4+4\mu^2)^{\frac{7}{2}}\left(\sigma_{trn}^2\left(\mu^2-\mu\sqrt{\mu^2+4}+2\right)+2\right)^3}.$$

Then we see that the denominator is positive. Hence the sign is determined by the numerator. Again, we assumed $p(\mu) < 0$. Hence the leading coefficient in term of σ_{trn}^2 is negative. Since $\sigma_{trn}^2 = N_{trn}$. If N_{trn} is sufficiently large the derivative is negative near $c = 1$. Thus, we have a peak. □

F.6 Proof of Theorem 7

Theorem 7 (Training Error)**.** *Let τ be as in Theorem 1. The training error for $c < 1$ is given by*

$$\mathbb{E}_{A_{trn}}[\|X_{trn} - W_{opt}(X_{trn} + A_{trn})\|_F^2] = \tau^{-2}\left(\sigma_{trn}^2(1 - c \cdot T_1) + \sigma_{trn}^4 T_2\right) + o(1),$$

where $T_1 = \dfrac{\mu^2}{2}\left(\dfrac{1+c+\mu^2 c}{\sqrt{(1-c+\mu^2 c)^2+4\mu^2c^2}} - 1\right) + \dfrac{1}{2} + \dfrac{1+\mu^2 c - \sqrt{(1-c+\mu^2c)^2+4c^2\mu^2}}{2c}$,

and $T_2 = \dfrac{(\mu^2c+c-1-\sqrt{(1-c+\mu^2c)^2+4c^2\mu^2})^2(\mu^2c+c+1-\sqrt{(1-c+\mu^2c)^2+4c^2\mu^2})}{2\sqrt{(1-c+\mu^2c)^2+4c^2\mu^2}}$.

Proof. Note that we have:

$$\begin{aligned}
\mathbb{E}_{A_{trn}}\left[\frac{\|X_{trn}-W_{opt}Y_{trn}\|_F^2}{N_{trn}}\right] &= \frac{1}{N_{trn}}\mathbb{E}_{A_{trn}}\left[\|X_{trn}-W_{opt}(X_{trn}+A_{trn}))\|_F^2\right] \\
&= \frac{1}{N_{trn}}\mathbb{E}[\|X_{trn}-W_{opt}X_{trn}\|^2]+\frac{1}{N_{trn}}\mathbb{E}[\|W_{opt}A_{trn}\|^2] \\
&+ \frac{2}{N_{trn}}\mathbb{E}\left[\mathrm{Tr}((X_{trn}-W_{opt}X_{trn})^T W_{opt}A_{trn})\right].
\end{aligned}$$

First, by Lemma 2, we have $X_{trn}-W_{opt}X_{trn} = \frac{\hat{\gamma}}{\hat{\tau}}X_{trn}$. Then, $\mathbb{E}[\|X_{trn}-W_{opt}X_{trn}\|^2] = \frac{\hat{\gamma}^2}{\hat{\tau}^2}\mathbb{E}[\|X_{trn}\|^2] = \frac{\hat{\gamma}^2\sigma_{trn}^2}{\hat{\tau}^2}$. Then, let us look at the $\mathbb{E}_{A_{trn}}[\|W_{opt}A_{trn}\|_F^2]$ term.

$$\begin{aligned}
\mathbb{E}_{A_{trn}}[\|W_{opt}A_{trn}\|_F^2] &= \mathbb{E}[\mathrm{Tr}(A_{trn}^T W_{opt}^T W_{opt}A_{trn})] \\
&= \frac{\sigma_{trn}^2\hat{\gamma}^2}{\hat{\tau}^2}\mathbb{E}[\mathrm{Tr}(A_{trn}^T\hat{h}^T u^T u\hat{h}A_{trn})] \\
&+ \frac{\sigma_{trn}^3\hat{\gamma}\|\hat{t}\|^2}{\hat{\tau}^2}\mathbb{E}[\mathrm{Tr}(A_{trn}^T\hat{h}^T u^T u\hat{k}^T\hat{A}_{trn}^\dagger A_{trn})] \\
&+ \frac{\sigma_{trn}^3\hat{\beta}\|\hat{t}\|^2}{\hat{\tau}^2}\mathbb{E}[\mathrm{Tr}(A_{trn}^T(\hat{A}_{trn}^\dagger)^T\hat{k}u^T u\hat{h}A_{trn})] \\
&+ \frac{\sigma_{trn}^4\|\hat{t}\|^4}{\hat{\tau}^2}\mathbb{E}[\mathrm{Tr}(A_{trn}^T(\hat{A}_{trn}^\dagger)^T\hat{k}u^T u\hat{k}^T\hat{A}_{trn}^\dagger A_{trn})] \\
&= \frac{\sigma_{trn}^2\hat{\gamma}^2}{\hat{\tau}^2}\mathbb{E}[\mathrm{Tr}(\hat{h}A_{trn}A_{trn}^T\hat{h}^T)] \\
&+ \frac{\sigma_{trn}^3\hat{\gamma}\|\hat{t}\|^2}{\hat{\tau}^2}\mathbb{E}[\mathrm{Tr}(\hat{k}^T\hat{A}_{trn}^\dagger A_{trn}A_{trn}^T\hat{h}^T)] \\
&+ \frac{\sigma_{trn}^3\hat{\gamma}\|\hat{t}\|^2}{\hat{\tau}^2}\mathbb{E}[\mathrm{Tr}(\hat{h}A_{trn}A_{trn}^T(\hat{A}_{trn}^\dagger)^T\hat{k})] \\
&+ \frac{\sigma_{trn}^4\|\hat{t}\|^4}{\hat{\tau}^2}\mathbb{E}[\mathrm{Tr}(\hat{k}^T\hat{A}_{trn}^\dagger A_{trn}A_{trn}^T(\hat{A}_{trn}^\dagger)^T\hat{k})] \\
&= \frac{\sigma_{trn}^2\hat{\gamma}^2}{\hat{\tau}^2}\mathbb{E}[\mathrm{Tr}(\hat{v}_{trn}^T\hat{A}_{trn}^\dagger A_{trn}A_{trn}^T(\hat{A}_{trn}^\dagger)^T\hat{v}_{trn}^T)] \\
&+ \frac{\sigma_{trn}^3\hat{\gamma}\|\hat{t}\|^2}{\hat{\tau}^2}\mathbb{E}[\mathrm{Tr}(u^T(\hat{A}_{trn}^\dagger)^T\hat{A}_{trn}^\dagger A_{trn}A_{trn}^T(\hat{A}_{trn}^\dagger)^T\hat{v}_{trn}^T)] \\
&+ \frac{\sigma_{trn}^3\hat{\gamma}\|\hat{t}\|^2}{\hat{\tau}^2}\mathbb{E}[\mathrm{Tr}(\hat{v}_{trn}^T\hat{A}_{trn}^\dagger A_{trn}A_{trn}^T(\hat{A}_{trn}^\dagger)^T\hat{A}_{trn}^\dagger u)] \\
&+ \frac{\sigma_{trn}^4\|\hat{t}\|^4}{\hat{\tau}^2}\mathbb{E}[\mathrm{Tr}(u^T(\hat{A}_{trn}^\dagger)^T\hat{A}_{trn}^\dagger A_{trn}A_{trn}^T(\hat{A}_{trn}^\dagger)^T\hat{A}_{trn}^\dagger u)] \\
&= \frac{\sigma_{trn}^2\hat{\gamma}^2}{\hat{\tau}^2}\mathbb{E}[\mathrm{Tr}(\hat{v}_{trn}^T\hat{A}_{trn}^\dagger A_{trn}A_{trn}^T(\hat{A}_{trn}^\dagger)^T\hat{v}_{trn}^T)] \\
&+ \frac{\sigma_{trn}^4\|\hat{t}\|^4}{\hat{\tau}^2}\mathbb{E}[\mathrm{Tr}(u^T(\hat{A}_{trn}^\dagger)^T\hat{A}_{trn}^\dagger A_{trn}A_{trn}^T(\hat{A}_{trn}^\dagger)^T\hat{A}_{trn}^\dagger u)].
\end{aligned}$$

Then, we look at the $\mathrm{Tr}((X_{trn} - W_{opt}X_{trn})^T W_{opt}A_{trn})$ term. By Lemma 2, we have $X_{trn} - W_{opt}X_{trn} = \frac{\hat{\gamma}}{\hat{\tau}}X_{trn}$. Then,

$$\begin{aligned}
\frac{\hat{\gamma}}{\hat{\tau}}\mathrm{Tr}(X_{trn}^T W_{opt}A_{trn}) &= \frac{\hat{\gamma}}{\hat{\tau}}\mathrm{Tr}\left(X_{trn}^T\left(\frac{\sigma_{trn}\hat{\gamma}}{\hat{\tau}}u\hat{h} + \frac{\sigma_{trn}^2\|\hat{t}\|^2}{\hat{\tau}}u\hat{k}^T\hat{A}_{trn}^\dagger\right)A_{trn}\right) \\
&= \frac{\sigma_{trn}\hat{\gamma}^2}{\hat{\tau}^2}\mathrm{Tr}\left(X_{trn}^T u\hat{h}A_{trn}\right) \\
&\quad + \frac{\sigma_{trn}^2\hat{\gamma}\|\hat{t}\|^2}{\hat{\tau}^2}\mathrm{Tr}\left(X_{trn}^T u\hat{k}^T\hat{A}_{trn}^\dagger A_{trn}\right) \\
&= \frac{\sigma_{trn}\hat{\gamma}^2}{\hat{\tau}^2}\mathrm{Tr}\left(\sigma_{trn}v_{trn}\hat{v}_{trn}^T\hat{A}_{trn}^\dagger A_{trn}\right) \\
&\quad + \frac{\sigma_{trn}^2\hat{\gamma}\|\hat{t}\|^2}{\hat{\tau}^2}\mathrm{Tr}\left(\sigma_{trn}v_{trn}u^T(\hat{A}_{trn}^\dagger)^T\hat{A}_{trn}^\dagger A_{trn}\right) \\
&= \frac{\sigma_{trn}^2\hat{\gamma}^2}{\hat{\tau}^2}\mathrm{Tr}\left(\hat{v}_{trn}^T\hat{A}_{trn}^\dagger A_{trn}v_{trn}\right) \\
&\quad + \frac{\sigma_{trn}^3\hat{\gamma}\|\hat{t}\|^2}{\hat{\tau}^2}\mathrm{Tr}\left(u^T(\hat{A}_{trn}^\dagger)^T\hat{A}_{trn}^\dagger A_{trn}v_{trn}\right) \\
&= \frac{\sigma_{trn}^2\hat{\gamma}^2}{\hat{\tau}^2}\mathrm{Tr}\left(\hat{v}_{trn}^T\hat{A}_{trn}^\dagger A_{trn}v_{trn}\right).
\end{aligned}$$

In conclusion, we have the training error:

$$\begin{aligned}
\mathbb{E}_{A_{trn}}\left[\frac{\|X_{trn} - W_{opt}Y_{trn}\|_F^2}{N_{trn}}\right] &= \frac{\hat{\gamma}^2\sigma_{trn}^2}{N_{trn}\hat{\tau}^2} + \frac{\sigma_{trn}^2\hat{\gamma}^2}{N_{trn}\hat{\tau}^2}\mathbb{E}[\mathrm{Tr}(\hat{v}_{trn}^T\hat{A}_{trn}^\dagger A_{trn}A_{trn}^T(\hat{A}_{trn}^\dagger)^T\hat{v}_{trn}^T)] \\
&\quad + \frac{\sigma_{trn}^4\|\hat{t}\|^4}{N_{trn}\hat{\tau}^2}\mathbb{E}[\mathrm{Tr}(u^T(\hat{A}_{trn}^\dagger)^T\hat{A}_{trn}^\dagger A_{trn}A_{trn}^T(\hat{A}_{trn}^\dagger)^T\hat{A}_{trn}^\dagger u)] \\
&\quad + 2\frac{\sigma_{trn}^2\hat{\gamma}^2}{N_{trn}\hat{\tau}^2}\mathbb{E}\left[\mathrm{Tr}\left(\hat{v}_{trn}^T\hat{A}_{trn}^\dagger A_{trn}v_{trn}\right)\right].
\end{aligned}$$

Now we estimate the above terms using random matrix theory. Here we focus on the $c < 1$ case. For $c < 1$, we note that

$$\hat{A}_{trn}^\dagger A_{trn}A_{trn}^T(\hat{A}_{trn}^\dagger)^T = \hat{V}\hat{\Sigma}^{-1}\Sigma\Sigma^T\hat{\Sigma}^{-1}\hat{V}^T.$$

Thus, for $c < 1$

$$\hat{v}_{trn}^T\hat{A}_{trn}^\dagger A_{trn}A_{trn}^T(\hat{A}_{trn}^\dagger)^T\hat{v}_{trn} = \sum_{i=1}^d a_i^2\frac{\sigma_i(A)^4}{(\sigma_i(A)^2 + \mu^2)^2}$$

where $a^T = v_{trn}^T V_{1:d}$. Taking the expectation, and using Lemma 9 we get that

$$\mathbb{E}_{A_{trn}}\left[\hat{v}_{trn}^T\hat{A}_{trn}^\dagger A_{trn}A_{trn}^T(\hat{A}_{trn}^\dagger)^T\hat{v}_{trn}\right] =$$
$$c\left(\frac{1}{2} + \frac{1+\mu^2c - \sqrt{(1-c+\mu^2c)^2 + 4c^2\mu^2}}{2c} + \mu^2\left(\frac{1+c+\mu^2c}{2\sqrt{(1-c+c\mu^2)^2 + 4c^2\mu^2}} - \frac{1}{2}\right)\right) + o(1).$$

Using Lemma 11, we see that the variance is $o(1)$. Similarly, we have that

$$(\hat{A}_{trn}^\dagger)^T\hat{A}_{trn}^\dagger A_{trn}A_{trn}^T(\hat{A}_{trn}^\dagger)^T\hat{A}_{trn}^\dagger = U\hat{\Sigma}^{-2}\Sigma\Sigma^T\hat{\Sigma}^{-2}U^T.$$

Thus, again, using a similar argument, we see that

$$\mathbb{E}_{A_{trn}}\left[\mathrm{Tr}(u^T(\hat{A}_{trn}^\dagger)^T\hat{A}_{trn}^\dagger A_{trn}A_{trn}^T(\hat{A}_{trn}^\dagger)^T\hat{A}_{trn}^\dagger u)\right] = \frac{1+c+\mu^2c}{2\sqrt{(1-c+c\mu^2)^2 + 4c^2\mu^2}} - \frac{1}{2} + o(1)$$

and again using Lemma 11, the variance is $o(1)$. Finally,

$$\hat{A}_{trn}^\dagger A_{trn} = \hat{V}\hat{\Sigma}^{-1}\Sigma V.$$

Thus,

$$\mathrm{Tr}(\hat{v}_{trn}^T \hat{A}_{trn}^\dagger A_{trn} v_{trn} = \sum_{i=1}^d a_i^2 \frac{\sigma_i(A)^2}{\sigma_i(A)^2 + \mu^2}.$$

Thus, using Lemma 9, we get that

$$\mathbb{E}_{A_{trn}}\left[\mathrm{Tr}(\hat{v}_{trn}^T \hat{A}_{trn}^\dagger A_{trn} v_{trn}\right] = \frac{1}{2} + \frac{1 + \mu^2 c - \sqrt{(1 - c + \mu^2 c)^2 + 4c^2\mu^2}}{2c} + o(1)$$

and using Lemma 11, the variance is $o(1)$. Then similar to the proof of Theorem 1, we can simplify the above expression to get the final result. □

F.7 Proof of Proposition 1

Proposition 1 (Optimal σ_{trn})**.** *The optimal value of σ_{trn}^2 for $c < 1$ is given by*

$$\sigma_{trn}^2 = \frac{\sigma_{tst}^2 d[2c(\mu^2+1)^2 - 2T(c\mu^2 + c + 1) + 2(c\mu^2 - 2c + 1)] + N_{tst}(\mu^2 c^2 + c^2 + 1 - T)}{N_{tst}(c^3(\mu^2+1)^2 - T(\mu^2 c^2 + c^2 - 1) - 2c^2 - 1)}.$$

Proof. Let $\sigma := \sigma_{trn}^2$ and

$$F = \tau^{-2}\left(\frac{\sigma_{tst}^2}{N_{tst}} + \frac{1}{d}(\sigma\|\hat{h}\|_2^2 + \sigma^2\|\hat{t}\|_2^4\rho)\right).$$

Notice that only τ is a function of σ, $\|\hat{h}\|_2^2$, $\|\hat{t}\|_2^2$, and $\|\hat{k}\|_2^2$ are all functions of μ. Then

$$\begin{aligned}
\frac{\partial F}{\partial \sigma} &= \tau^{-2}\frac{1}{d}(\|\hat{h}\|_2^2 + 2\sigma\|\hat{t}\|_2^4\rho) - 2\tau^{-3}\frac{\partial \tau}{\partial \sigma}\left(\frac{\sigma_{tst}^2}{N_{tst}} + \frac{1}{d}\left(\sigma\|\hat{h}\|_2^2 + \sigma^2\|\hat{t}\|_2^4\rho\right)\right) \\
&= \tau^{-2}\frac{1}{d}(\|\hat{h}\|_2^2 + 2\sigma\|\hat{t}\|_2^4\rho) - 2\tau^{-3}\|\hat{t}\|_2^2\|\hat{k}\|_2^2\left(\frac{\sigma_{tst}^2}{N_{tst}} + \frac{1}{d}(\sigma\|\hat{h}\|_2^2 + \sigma^2\|\hat{t}\|_2^4\rho)\right) \\
&= \tau^{-2}\left(\frac{1}{d}(\|\hat{h}\|_2^2 + 2\sigma\|\hat{t}\|_2^4\rho) - 2\tau^{-1}\|\hat{t}\|_2^2\|\hat{k}\|_2^2\left(\frac{\sigma_{tst}^2}{N_{tst}} + \frac{1}{d}(\sigma\|\hat{h}\|_2^2 + \sigma^2\|\hat{t}\|_2^4\rho)\right)\right).
\end{aligned}$$

The optimal σ^* satisfies $\frac{\partial F}{\partial \sigma}|_{\sigma=\sigma^*} = 0$. Thus, we can solve the equation

$$\tau^{-2} = 0 \quad \text{or} \quad \frac{1}{d}(\|\hat{h}\|_2^2 + 2\sigma\|\hat{t}\|_2^4\rho) - 2\tau^{-1}\|\hat{t}\|_2^2\|\hat{k}\|_2^2\left(\frac{\sigma_{tst}^2}{N_{tst}} + \frac{1}{d}(\sigma\|\hat{h}\|_2^2 + \sigma^2\|\hat{t}\|_2^4\rho)\right).$$

Let $\alpha := \|\hat{t}\|_2^2\|\hat{k}\|_2^2$, $\delta := d\frac{\sigma_{tst}^2}{N_{tst}}$. Then

$$\tau^{-2} = 0 \implies \sigma = -\frac{1}{\|t\|_2^2\|k\|_2^2}.$$

Notice that $\sigma < 0$ implies σ_{trn} is an imaginary number, something we don't want. Thus, we look at the other expression.

$$\begin{aligned}
0 &= \frac{1}{d}(\|\hat{h}\|_2^2 + 2\sigma\|\hat{t}\|_2^4\rho) - 2\tau^{-1}\|\hat{t}\|_2^2\|k\|_2^2\left(\frac{\sigma_{tst}^2}{N_{tst}} + \frac{1}{d}(\sigma\|\hat{h}\|_2^2 + \sigma^2\|\hat{t}\|_2^4\rho)\right) \\
&= \frac{1}{d}(\|\hat{h}\|_2^2 + 2\sigma\|\hat{t}\|_2^4\rho) - 2\tau^{-1}\alpha\left(\frac{\delta}{d} + \frac{1}{d}(\sigma\|\hat{h}\|_2^2 + \sigma^2\|\hat{t}\|_2^4\rho)\right). \qquad [\alpha = \|\hat{t}\|_2^2\|\hat{k}\|_2^2]
\end{aligned}$$

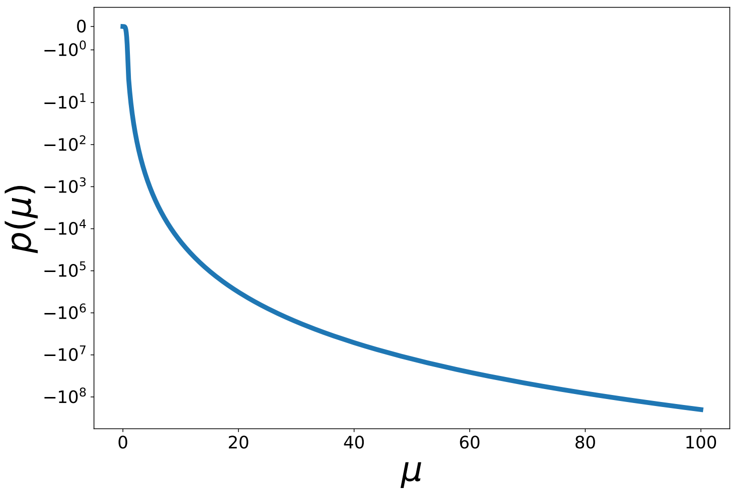


Figure 19: Figure showing the value of $p(\mu)$

Then multiplying through by d and τ

$$\begin{aligned} 0 &= (1+\alpha\sigma)(\|\hat{h}\|_2^2 + 2\sigma\|\hat{t}\|_2^4\rho) - 2\alpha(\delta + \sigma\|\hat{h}\|_2^2 + \sigma^2\|\hat{t}\|_2^4\rho) && [\tau = 1+\alpha\sigma] \\ &= \|\hat{h}\|_2^2 + 2\|\hat{t}\|_2^4\rho\sigma + \alpha\|\hat{h}\|_2^2\sigma + 2\alpha\|\hat{t}\|_2^4\rho\sigma^2 - 2\alpha\delta - 2\alpha\|\hat{h}\|_2^2\sigma - 2\alpha\|\hat{t}\|_2^4\rho\sigma^2 \\ &= \|\hat{h}\|_2^2 + 2\|\hat{t}\|_2^4\rho\sigma + \alpha\|\hat{h}\|_2^2\sigma - 2\alpha\delta - 2\alpha\|\hat{h}\|_2^2\sigma. \end{aligned}$$

Then solving for σ, we get that

$$\sigma = \frac{2\alpha\delta - \|\hat{h}\|^2}{2\|t\|^4\rho - \alpha\|\hat{h}\|^2} = \frac{2d\|\hat{t}\|_2^2\|\hat{k}\|_2^2\sigma_{tst}^2 - \|\hat{h}\|^2 N_{tst}}{N_{tst}(2\|\hat{t}\|_2^4\rho - \|\hat{t}\|_2^2\|\hat{k}\|_2^2\|\hat{h}\|_2^2)}.$$

Then we use the random matrix theory lemmas to estimate this quantity. $\square$

G Experiments

All experiments were conducted using Pytorch and run on Google Colab using an A100 GPU. For each empirical data point, we did at least 100 trials. The maximum number of trials for any experiment was 20000 trials.

For each configuration of the parameters, $N_{trn}, N_{tst}, d, \sigma_{trn}, \sigma_{tst}$, and μ. For each trial, we sampled u, v_{trn}, v_{tst} uniformly at random from the appropriate dimensional sphere. We also sampled new training and test noise for each trial.

For the data scaling regime, we kept $d = 1000$ and for the parameter scaling regime, we kept $N_{trn} = 1000$. For all experiments, $N_{tst} = 1000$.

H Technical Assumption on μ

Notice that we had this assumption that $p(\mu) < 0$. We compute $p(\mu)$ for a million equally spaced points in $(0, 100]$ and see that $p(\mu) < 0$. Here we use Mpmath with a precision of 1000. The result is shown in Figure 19. Hence we see that the assumption is satisfied for $\mu \in (0, 100]$.