Denoising Low-Rank Data Under Distribution Shift: Double Descent and Data Augmentation

Anonymous Author(s)
Affiliation
Address
email

Abstract

Despite the importance of denoising in modern machine learning and ample empirical work on supervised denoising, its theoretical understanding is still relatively scarce. One concern about studying supervised denoising is that one might not always have noiseless training data from the test distribution. It is more reasonable to have access to noiseless training data from a different dataset than the test dataset. Motivated by this, we study supervised denoising and noisy-input regression under distribution shift. We add three considerations to increase the applicability of our theoretical insights to real-life data and modern machine learning. First, while most past theoretical work assumes that the data covariance matrix is full-rank and well-conditioned, empirical studies have shown that real-life data is approximately low-rank. Thus, we assume that our data matrices are low-rank. Second, we drop independence assumptions on our data. Third, the rise in computational power and dimensionality of data have made it important to study non-classical regimes of learning. Thus, we work in the non-classical proportional regime, where data dimension d and number of samples N grow as $d/N = c + o(1)$.

For this setting, we derive general test error expressions for both denoising and noisy-input regression, and study when overfitting the noise is benign, tempered or catastrophic. We show that the test error exhibits double descent under general distribution shift, providing insights for data augmentation and the role of noise as an implicit regularizer. We also perform experiments using real-life data, where we match the theoretical predictions with under 1% MSE error for low-rank data.

1 Introduction

Denoising and noisy-input problems have a rich history in machine learning [1–3]. Aside from its natural application to noisy input data, the idea of noise as a regularizer has led to denoising being tied to many areas of modern machine learning, such as pretraining and feature extraction [4], data-augmentation for representation learning [5], generative modeling [6]. While unsupervised methods like PCA [7] and low rank matrix recovery [8] have been addressed in prior theoretical work [9], *supervised* methods like denoising autoencoders are theoretically less well-understood.

One of the biggest practical qualms to studying a supervised setting is that a learner needs access to noiseless data sampled from the test distribution. However, this is resolved by considering *distribution shift*, which is when the training and test data can come from different distributions. Given this practical motivation, we study supervised denoising and noisy-input regression under distribution shift. It is well understood that non-trivial denoising is made possible by the presence of additional structure in the data (see, for example, Section 3.2 of [1]). One of the most natural such structures is low rank, specifically the idea that the true inputs live in a low dimensional space. In fact, past

Submitted to 37th Conference on Neural Information Processing Systems (NeurIPS 2023). Do not distribute.

work such as [10] has demonstrated that *a lot of real-life data is approximately low-rank* – that is, its covariance matrix only has a few significant eigenvalues.

The classical theory of learning problems would keep the data dimension d fixed and let the number of samples N grow to ∞. These can be theoretically analysed using elementary tools. However, with growing access to computational power and richness of data, it has become important to study *non-classical regimes*. One important and popular example is the proportional regime, where $d \propto N$ and so d is comparable to N [11, 12]. However, there is very little work on learning with *noisy inputs* in non-classical regimes. Our paper takes one of the first steps towards filling this gap.

Additionally, most past theoretical works in non-classical regimes do not test on *real-life data*. As argued above and in [11], a big reason for this issue is that past work assumes that the data covariance matrix is well-conditioned, while real-life data covariance matrices are better modeled by low-rank assumptions. We aim to address this issue by testing our theory for low-rank data on real-life datasets. In real life, one has little control over the independence or even the distribution of the data [13]. There is also a growing need to be robust to adversarially chosen data in machine learning [14]. We would thus like to drop the assumption that the data is IID or even independent. Additionally, explicit structural assumptions made about distribution shift in past work are often quite restrictive, involving requirements like the simultaneous diagonalizability of the train and test covariance matrices [15] or joint distributions of the training data's eigenvalues and certain overlap coefficients [16, 17]. We would like to drop such assumptions and work with general distribution shift, decoupling assumptions on the test and train data. We thus aim to address the following question:

Q.1. Can we derive test error expressions for denoising and noisy-input regression that:
- (a) work with data from a low-dimensional subspace under a non-classical regime,
- (b) make minimal assumptions on the training data, test data and how they are related,
- (c) match experiments that use real-life data distributions?

Q.2. What insights can we obtain from these?

Contributions. Answering our questions, we fill the gap in theoretically studying supervised denoising in a non-classical regime. We drop independence assumptions on data and work with arbitrary test data from our low-dimensional subspace. We also experiment using real-life data, achieving under 1% MSE error.[1] Finally, we provide insights about double descent, overfitting phenomena and data augmentation, all in the context of denoising under general distribution shift.

2 Problem Setup and Notation

Consider training data $X_{trn} \in \mathbb{R}^{d\times N}$, $\beta \in \mathbb{R}^{d\times k}$ with target outputs $Y_{trn} = \beta^T X_{trn}$, and a training noise matrix $A_{trn} \in \mathbb{R}^{d\times N}$. We assume that we have access to Y_{trn} and $X_{trn} + A_{trn}$ while training. The goal is to study the test error of the minimum norm linear function W_{opt} that minimizes the *MSE training error*. MSE error is also one of the most common targets for non-linear auto-encoders [1]. We formalize the definition of W_{opt} below.

$$W_{opt} = \arg\min_W \left\{ \|W\|_F^2 \middle| W \in \arg\min_W \|Y_{trn} - W(X_{trn} + A_{trn})\|_F^2 \right\}$$

Given test data $X_{tst} \in \mathbb{R}^{d\times N_{tst}}$ and $Y_{tst} = \beta^T X_{tst}$, we formally define the <u>*test error*</u> for arbitrary linear functions W by $\mathcal{R}(W, X_{tst})$ below. Since we are not assuming anything about the distribution of the training or test data, we only take the expectation over the training and test noise.

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

We study the test error $\mathcal{R}(W_{opt}, X_{tst})$ of W_{opt} in terms of properties of the data matrices X_{trn} and X_{tst} as well as the noise distributions. For simplicity, we assume access to noiseless outputs Y. Notice that when $\beta = I$, we are studying the linear denoising problem, and when $\beta \in \mathbb{R}^d$, we are studying real-valued regression with noisy inputs. We work in the *proportional regime*, where

[1]The code for the experiments can be found in the following anonymized repository [Link].

$d/N = c + o(1)$ as N grows, for some constant $c > 0$. We discuss the generality of our assumptions in Appendix A, providing a comparison with prior work and justifications for our assumptions.

Assumption 1 (Data). *We have d-dimensional data $X_{trn} \in \mathbb{R}^{d \times N}$ and $X_{tst} \in \mathbb{R}^{d \times N_{tst}}$ so that*

1. *Low-rank: There is a fixed $r > 0$ so that X_{trn} and X_{tst} have data-points lying in an r-dimensional subspace $\mathcal{V} \subset \mathbb{R}^d$, and the column span of X_{trn} is $\mathcal{V}$.*
2. *Data growth: $\|X_{trn}\|_F^2 = O(N)$.*
3. *Low-rank well-conditioning: For the r singular values σ_i of X_{trn}, $\frac{\sigma_j}{\sigma_i} = \Theta(1)$ and $\frac{1}{\sigma_i} = o(1)$ as N grows, for any i, j.*

Assumption 2 (Noise). *Let the train and test noise matrices $A_{trn}, A_{tst} \in \mathbb{R}^{d \times N}$ be sampled from distributions $\mathcal{D}_{trn}$ and $\mathcal{D}_{tst}$ such that A_{trn} satisfies points $1-4$ below and A_{tst} satisfies points $1, 2$.*

1. *For all i, j, $\mathbb{E}_{\mathcal{D}}[A_{ij}] = 0$, and $\mathbb{E}_{\mathcal{D}}[A_{ij}^2] = \eta^2/d$. Here $\eta = \Theta(1)$ as N grows.*
2. *For all $\{i_1, j_1\} \neq \{i_2, j_2\}$, $\mathbb{E}_{\mathcal{D}}[A_{i_1 j_1} A_{i_2 j_2}] = \mathbb{E}_{\mathcal{D}}[A_{i_1 j_1}]\mathbb{E}_{\mathcal{D}}[A_{i_2 j_2}]$.*
3. *$\mathcal{D}$ is a rotationally bi-invariant distribution[2] and $A \sim \mathcal{D}$ is full rank with probability one.*
4. *Suppose $A^{d,N}$ is a sequence of matrices such that with $d/N = c + o(1)$ as N grows, for $c > 0$. Let $\lambda_1^{d,N}, \ldots, \lambda_N^{d,N}$ be the eigenvalues of $(A^{d,N})^T A^{d,N}$. Let $\mu_{d,N} = \sum_i \delta_{\lambda_i^{d,N}}$ be the sum of dirac delta measures for the eigenvalues. Then we shall assume that $\mu_{d,N}$ converges weakly in probability to the Marchenko-Pastur measure with shape c as N grows (see Appendix C).*

Terminology. We now define the overfitting paradigms that we will study. Motivated by past work on benign overfitting, we present a reasonable generalization of overfitting paradigms (benign, tempered and catastrophic, see [18]) to our setting. Consider the minimum norm denoiser that minimizes *expected* MSE training error, similar in spirit to θ^* in [19].

$$W^* = \arg\min_W \left\{ \|W\|_F^2 \middle| W \in \arg\min_W \mathbb{E}_{A_{trn}}[\|Y_{trn} - W(X_{trn} + A_{trn})\|_F^2] \right\}$$

Recall that we obtain W_{opt} by minimizing the MSE error for a *single* noise instance A_{trn}. So, W_{opt} overfits A_{trn} in the overparametrized regime. We would like to see if this overfitting is benign, tempered or catastrophic for test error. Following the definition of overfitting paradigms in [18], we want to take $N \to \infty$. Since we are in the proportional regime, we must let $d \to \infty$ as well, maintaining the relation $d/N = c + o(1)$. For studying overfitting, a natural goal would be to study how the excess error $\mathcal{R}(W_{opt}, X_{tst}) - \mathcal{R}(W^*, X_{tst})$ behaves as $d, N \to \infty$. This is analogous to the excess risk studied in overfitting for noiseless inputs [19]. However, we will see that both errors in our difference individually tend to zero as $d, N \to \infty$, making this a somewhat meaningless criterion. As noted in [20], benign overfitting is traditionally restricted to scenarios where the minimum possible error is non-zero. A natural generalization to consider then is to instead study the limit of *relative excess error* $\frac{\mathcal{R}(W_{opt}, X_{tst}) - \mathcal{R}(W^*, X_{tst})}{\mathcal{R}(W^*, X_{tst})}$ as $d, N \to \infty$ with $d/N = c + o(1)$.

Definition 1. *We say that overfitting is* benign *when this limit is* 0*,* tempered *when it is finite and positive, and* catastrophic *when it is* ∞*.*

3 Theoretical Results

This section presents our main result – Theorem 1. We present the results here and discuss insights at the end of the paper. All proofs are in Appendix F.

Theorem 1 (In-Subspace Test Error). *Let $r < |d - N|$. Let the SVD of X_{trn} be $U\Sigma_{trn}V_{trn}^T$, let $L := U^T X_{tst}$, $\beta_U := U^T \beta$, and $c := d/N$. Under our setup and Assumptions 1 and 2, the test error (Equation 1) is given by the following. If $c < 1$ (under-parameterized regime)*

$$\mathcal{R}(W_{opt}, UL) = \frac{\eta_{trn}^4}{N_{tst}} \left\| \beta_U^T (\Sigma_{trn}^2 c + \eta_{trn}^2 I)^{-1} L \right\|_F^2 + \frac{\eta_{tst}^2}{d} \frac{c^2}{1-c} \operatorname{Tr}\left(\beta_U \beta_U^T \Sigma_{trn}^2 \left(\Sigma_{trn}^2 + \frac{1}{\eta_{trn}^2} I \right) \left(\Sigma_{trn}^2 c + \eta_{trn}^2 I \right)^{-2} \right) + o\left(\frac{1}{N}\right)$$

[2] A distribution over matrices $A \in \mathbb{R}^{m \times n}$ is rotationally bi-invariant if for all orthogonal $U_1 \in \mathbb{R}^{m \times m}$ and all orthogonal $U_2 \in \mathbb{R}^{n \times n}$, $U_1 A U_2$ has the same distribution as A. Another way to phrase rotational bi-invariance is if the SVD of A is given by $A = U_A \Sigma_A V_A^T$, then U_A and V_A are uniformly random orthogonal matrices and are independent of Σ_A and each other.

If $c > 1$ (over-parameterized regime)

$$\mathcal{R}(W_{opt}, UL) = \frac{\eta_{trn}^4}{N_{tst}} \left\| \beta_U^T (\Sigma_{trn}^2 + \eta_{trn}^2 I)^{-1} L \right\|_F^2 + \frac{\eta_{tst}^2}{d} \frac{c}{c-1} \operatorname{Tr}(\beta_U \beta_U^T (I + \eta_{trn}^2 \Sigma_{trn}^{-2})^{-1}) + O\left(\frac{\|\Sigma_{trn}\|^2}{N^2}\right) + o\left(\frac{1}{N}\right)$$

Theorem 1 is significant, non-trivial and can be used to understand OOD and out-of-subspace test error, special cases with IID data, as well as overfitting paradigms. We present consequences for in-subspace distribution shift and overfitting paradigms below, **relegating other results to Appendix E**.

Corollary 1 (Distribution Shift Bound). *Let W_{opt} be tested on test data $X_{tst,1} = UL_1$ and $X_{tst,2} = UL_2$ generated possibly dependently from distributions supported in the span of U with mean $U\mu_i$ and covariance $\Sigma_{U,i} = U\Sigma_i U^T$ respectively. Let $f(c) = c$ for $c < 1$ and $f(c) = 1$. Then, the difference in* generalization errors $\mathcal{G}_i := \mathbb{E}_{X_{tst,i}}[\mathcal{R}(W_{opt}, X_{tst,i})]$ *is bounded for $c < 1$ by*

$$|\mathcal{G}_2 - \mathcal{G}_1| \leq \frac{\sigma_1(\beta)^2 \eta_{trn}^4 r}{(\sigma_r(X_{trn})^2 f(c) + \eta_{trn}^2)^2} \|\Sigma_2 - \Sigma_1 + \mu_2 \mu_2^T - \mu_1 \mu_1^T\|_F + o\left(\frac{1}{N}\right).$$

We add $O(\|\Sigma_{trn}\|_F^2 / N^2)$ to the bound when $c \geq 1$.

Corollary 2 (Relative Excess Error). *Let $\|\Sigma_{trn}\|_F^2 = \Omega(N^{1/2+\epsilon})$. As $d, N \to \infty$ with $d/N \to c$, the relative excess error tends to $\frac{c}{1-c}$ in the underparametrized regime. In the overparametrized regime, when $\|\Sigma_{trn}\|_F^2 = o(N)$, it tends to $\frac{1}{c-1}$ and to $\frac{1}{c-1} + k$ for some constant k when $\|\Sigma_{trn}\|_F^2 = \Theta(N)$.*

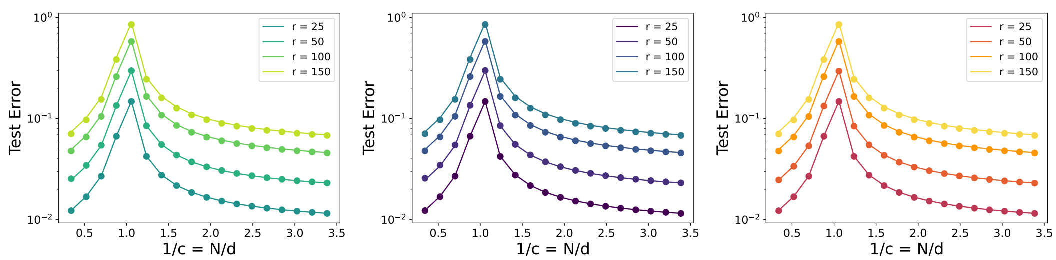


Figure 1: Test error for $\beta = I$ vs $1/c = N/d$. Test error is averaged over 200 trials with fresh A_{tst}. Similar results are obtained for single-variable regression with $\beta \in \mathbb{R}^d$ in Appendix D.2.

Experimental Verification Since d is fixed, we vary c by varying N. Figure 7 shows the empirical performance of W_{opt} trained on CIFAR data and applied to various datasets. We use Principal Component Regression to impose the low-rank condition here, details for which are in Appendix D along with other experiments which use raw real-life data.

Insights. Recall from Corollary 2 that when $\|\Sigma_{trn}\|_F^2 = o(N)$, the relative excess error is given by $\frac{1}{c-1}$ when $c > 1$ and by $\frac{c}{1-c}$ when $c < 1$. This means that we experience catastrophic overfitting when $c = 1$, tempered overfitting for $c \neq 1$, and approach benign overfitting only as c becomes arbitrarily large or arbitrarily small (the latter is essentially the classical regime). If $\|\Sigma_{trn}\|_F^2 = \Theta(N)$, the relative excess error may increase by a constant. We expand on this in Appendix B, also providing insights on double descent and data augmentation under distribution shift.

4 Conclusion

We studied the problem of denoising low-dimensional input data perturbed with high-dimensional noise. Under very general assumptions, we provided estimates test error in terms of the specific instantiations of the training data and test data. This result is significant, as there is scarce prior work in the area of generalization for noisy inputs as well as generalization for low-rank data. Further, we tested our results using *real data* and achieve a relative MSE of 1%. Finally, the data-dependent estimate lets us provide many insights that would be harder to get with results on generalization error, such as showing double descent for arbitrary test data in our low-dimensional subspace, theoretically understanding data augmentation and provably demonstrating as well as explaining the lack of benign overfitting. Our work opens the door for the analysis of non-linear denoising in a similar setting.

References

[1] Pascal Vincent, Hugo Larochelle, Isabelle Lajoie, Yoshua Bengio, and Pierre-Antoine Manzagol. Stacked denoising autoencoders: learning useful representations in a deep network with a local denoising criterion. *Journal of Machine Learning Research*, 11(110):3371–3408, 2010. URL: `http://jmlr.org/papers/v11/vincent10a.html` (cited on pages 1, 2).

[2] Chunwei Tian, Lunke Fei, Wenxian Zheng, Yong Xu, Wangmeng Zuo, and Chia-Wen Lin. Deep learning on image denoising: an overview. *Neural Networks*, 131:251–275, 2020. ISSN: 0893-6080. DOI: `https://doi.org/10.1016/j.neunet.2020.07.025`. URL: `https://www.sciencedirect.com/science/article/pii/S0893608020302665` (cited on page 1).

[3] Michael Elad, Bahjat Kawar, and Gregory Vaksman. Image denoising: the deep learning revolution and beyond—a survey paper. *SIAM Journal on Imaging Sciences*, 16(3):1594–1654, 2023. DOI: `10.1137/23M1545859`. eprint: `https://doi.org/10.1137/23M1545859`. URL: `https://doi.org/10.1137/23M1545859` (cited on page 1).

[4] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet Classification with Deep Convolutional Neural Networks. *Communications of the ACM*, 2012 (cited on page 1).

[5] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A Simple Framework for Contrastive Learning of Visual Representations. In *Proceedings of the 37th International Conference on Machine Learning*, Proceedings of Machine Learning Research, pages 1597–1607. PMLR, 2020 (cited on page 1).

[6] Robin Rombach, A. Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-Resolution Image Synthesis with Latent Diffusion Models. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022 (cited on page 1).

[7] Karl Pearson F.R.S. Liii. on lines and planes of closest fit to systems of points in space. *Philosophical Magazine Series 1*, 2:559–572, 1901. URL: `https://api.semanticscholar.org/CorpusID:125037489` (cited on page 1).

[8] Mark A. Davenport and Justin K. Romberg. An overview of low-rank matrix recovery from incomplete observations. *CoRR*, abs/1601.06422, 2016. arXiv: `1601.06422`. URL: `http://arxiv.org/abs/1601.06422` (cited on page 1).

[9] Pierre Baldi and Kurt Hornik. Neural networks and principal component analysis: learning from examples without local minima. *Neural Networks*, 2:53–58, 1989. URL: `https://api.semanticscholar.org/CorpusID:14333248` (cited on page 1).

[10] Madeleine Udell and Alex Townsend. Why Are Big Data Matrices Approximately Low Rank? *SIAM Journal on Mathematics of Data Science*, 2019 (cited on page 2).

[11] Chen Cheng and Andrea Montanari. Dimension Free Ridge Regression. *arXiv preprint arXiv:2210.08571*, 2022 (cited on pages 2, 10).

[12] Mojtaba Sahraee-Ardakan, Melikasadat Emami, Parthe Pandit, Sundeep Rangan, and Alyson K. Fletcher. Kernel methods and multi-layer perceptrons learn linear models in high dimensions. *ArXiv*, abs/2201.08082, 2022. URL: `https://api.semanticscholar.org/CorpusID:246064069` (cited on page 2).

[13] Matthias Kirchler, Christoph Lippert, and Marius Kloft. Training normalizing flows from dependent data. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 17105–17121. PMLR, 23–29 Jul 2023. URL: `https://proceedings.mlr.press/v202/kirchler23a.html` (cited on page 2).

[14] Shashank Kotyan. A reading survey on adversarial machine learning: adversarial attacks and their understanding, 2023. arXiv: `2308.03363 [cs.LG]` (cited on page 2).

[15] Daniel LeJeune, Jiayu Liu, and Reinhard Heckel. Monotonic Risk Relationships under Distribution Shifts for Regularized Risk Minimization. *arXiv preprint arXiv:2210.11589*, 2022 (cited on pages 2, 16).

[16] Nilesh Tripuraneni, Ben Adlam, and Jeffrey Pennington. Covariate Shift in High-Dimensional Random Feature Regression. *ArXiv*, 2021 (cited on pages 2, 10, 12).

[17] Nilesh Tripuraneni, Ben Adlam, and Jeffrey Pennington. Overparameterization Improves Robustness to Covariate Shift in High Dimensions. In *Advances in Neural Information Processing Systems*, 2021 (cited on pages 2, 10).

[18] Neil Mallinar, James B. Simon, Amirhesam Abedsoltan, Parthe Pandit, Mikhail Belkin, and Preetum Nakkiran. Benign, tempered, or catastrophic: a taxonomy of overfitting, 2022. URL: https://arxiv.org/abs/2207.06569 (cited on page 3).

[19] Peter Bartlett, Philip M. Long, Gabor Lugosi, and Alexander Tsigler. Benign Overfitting in Linear Regression. *Proceedings of the National Academy of Sciences*, 2020 (cited on pages 3, 10).

[20] Ohad Shamir. The Implicit Bias of Benign Overfitting. In *Proceedings of 35th Conference on Learning Theory*, 2022 (cited on page 3).

[21] Edgar Dobriban and Stefan Wager. High-dimensional asymptotics of prediction: ridge regression and classification. *The Annals of Statistics*, 2018 (cited on page 10).

[22] 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 10).

[23] 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 page 10).

[24] 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*, 2021 (cited on page 10).

[25] 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 pages 10, 13).

[26] 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 pages 10, 11, 13).

[27] Denny Wu and Ji Xu. On the Optimal Weighted ℓ_2 Regularization in Overparameterized Linear Regression. *Advances in Neural Information Processing Systems*, 2020 (cited on page 10).

[28] 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 10, 11).

[29] 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 10, 11, 13, 16, 22–25, 31–34, 36).

[30] Hugo Cui and Lenka Zdeborová. High-dimensional asymptotics of denoising autoencoders. *arXiv preprint arXiv:2305.11041*, 2023 (cited on page 11).

[31] Arnu Pretorius, Steve Kroon, and Herman Kamper. Learning Dynamics of Linear Denoising Autoencoders. In *Proceedings for the 35th International Conference on Machine Learning*, 2018 (cited on page 11).

[32] 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 page 11).

[33] 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 page 12).

[34] 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 12).

[35] Jeffrey Pennington and Pratik Worah. Nonlinear Random Matrix Theory for Deep Learning. In *Advances in Neural Information Processing Systems*, 2017 (cited on page 12).

[36] Lucas Benigni and Sandrine Péché. Eigenvalue distribution of some nonlinear models of random matrices. *Electronic Journal of Probability*, 26:1–37, 2021 (cited on page 12).

[37] Sandrine Péché. A Note on the Pennington-Worah Distribution. *Electronic Communications in Probability* (cited on page 12).

[38] Fatih Furkan Yilmaz and Reinhard Heckel. Regularization-Wise Double Descent: Why it Occurs and How to Eliminate it. In *IEEE International Symposium on Information Theory*, 2022 (cited on page 13).

[39] Alex Krizhevsky. Learning Multiple Layers of Features from Tiny Images. In 2009 (cited on pages 13, 45).

[40] Adam Coates, A. Ng, and Honglak Lee. An Analysis of Single-Layer Networks in Unsupervised Feature Learning. In *Proceedings of the 14th International Conference on Artificial Intelligence and Statistics*, 2011 (cited on pages 13, 45).

[41] Yuval Netzer, Tao Wang, Adam Coates, A. Bissacco, Bo Wu, and A. Ng. Reading Digits in Natural Images with Unsupervised Feature Learning. In 2011 (cited on pages 13, 45).

[42] Yimin Wei. The Weighted Moore–Penrose Inverse of Modified Matrices. *Applied Mathematics and Computation*, 2001 (cited on pages 21, 30).

[43] George W. Bohrnstedt and Arthur S. Goldberger. On the exact covariance of products of random variables. *Journal of the American Statistical Association*, 64(328):1439–1442, 1969. ISSN: 01621459. URL: `http://www.jstor.org/stable/2286081` (visited on 05/24/2023) (cited on pages 25, 28, 36).

[44] Rishi Sonthalia, Xinyue Li, and Bochao Gu. Under-Parameterized Double Descent for Ridge Regularized Least Squares Denoising of Data on a Line. In 2023 (cited on pages 41, 44).

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

[46] Friedrich Götze and A. Tikhomirov. On the Rate of Convergence to the Marchenko–Pastur Distribution. *arXiv: Probability*, 2011 (cited on page 44).

[47] Friedrich Götze and Alexander Tikhomirov. Rate of Convergence to the Semi-Circular Law. *Probability Theory and Related Fields*, 2003 (cited on page 44).

[48] Friedrich Götze and Alexander Tikhomirov. Rate of Convergence in Probability to the Marchenko-Pastur Law. *Bernoulli*, 2004 (cited on page 44).

[49] 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 44).

[50] 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 44).

Contents

A Discussion of Assumptions

Assumptions about the data. We recall the assumptions below. Note that they formalize three natural requirements on the data – (1) that it lies in a low-dimensional subspace as argued above; (2) that the norm of the training data does not grow too much faster than the norm of the training noise, otherwise there will not be enough noise to train on; (3) that the training data "sees enough" of the subspace containing the data.

Assumption 1 (Data). *We have d-dimensional data $X_{trn} \in \mathbb{R}^{d \times N}$ and $X_{tst} \in \mathbb{R}^{d \times N_{tst}}$ so that*

1. *Low-rank: There is a fixed $r > 0$ so that X_{trn} and X_{tst} have data-points lying in an r-dimensional subspace $\mathcal{V} \subset \mathbb{R}^d$, and the column span of X_{trn} is $\mathcal{V}$.*
2. *Data growth: $\|X_{trn}\|_F^2 = O(N)$.*
3. *Low-rank well-conditioning: For the r singular values σ_i of X_{trn}, $\frac{\sigma_j}{\sigma_i} = \Theta(1)$ and $\frac{1}{\sigma_i} = o(1)$ as N grows, for any i, j.*

Notice that we don't assume that X_{trn} is IID or even independent, and X_{tst} is completely arbitrary, besides lying in the subspace $\mathcal{V}$. In our results, we will characterize the dependence of the error on X_{trn} and X_{tst} using their singular values. These intuitively measure "how much each direction is sampled," and don't depend on the distribution of the data. Finally, let $X_{trn} = U\Sigma_{trn}V_{trn}^T$ be the SVD of X_{trn} with $U \in \mathbb{R}^{d\times r}$, $\Sigma_{trn} \in \mathbb{R}^{r\times r}$ and $V_{trn}^T \in \mathbb{R}^{r\times N}$. Note that the columns of U span $\mathcal{V}$. Then there exists a matrix L such that $X_{tst} = UL$. For Theorem 3, we will relax our assumption on X_{tst} to say that there exists L and $\alpha > 0$ so that $\|X_{tst} - UL\| < \alpha$.

Comparison with assumptions in prior work. Most prior work assumes that the data comes from a Gaussian or Gaussian-like distribution. Specifically, [16, 17, 21–26] assume that $x \sim \mathcal{N}(0, \Sigma)$. Most real data cannot be modeled as Gaussian data. Another common assumption is that $x = \Sigma^{1/2}z$ where the coordinates of z are independent, centered, and have a variance of 1. This setting is a little bit more general than the previous setting. The independence of data is still a limiting assumption that prevents it from modeling real-life data well. In addition, as the dimension increases, due to the (Lyapunov's) central limit theorem, the data's higher moments tend towards those of a Gaussian distribution again. This makes this assumption nearly as limiting as the first one. Papers with this (or very similar) assumption include [11, 19, 27, 28].

In conclusion, we provide results on test error in a very different low-rank setting inspired by real-life data, and drop many restrictive assumptions. A small number of papers [25, 26, 29] that do assume a low-rank structure. However, the first two *further* assume that the data is low-rank Gaussian, while the third only provides results for one-dimensional data. Notice that our assumptions completely subsume both of these.

Assumptions about the training noise. Our assumptions on noise are fairly natural and general. We recall them below. Informally, we require the training noise to (1) have finite second moments, (2) be uncorrelated across entries, (3) be isotropic, and (4) follow a natural limit theorem. On the other hand, the test noise only needs (1) finite second moments and (2) uncorrelated entries. Our assumptions include a broad class of noise distributions (see Proposition 1 of [29]). One of the many examples of noise distributions satisfying these is Gaussian noise, with each coordinate having variance $1/d$. We recall our noise assumptions.

Assumption 2 (Noise). *Let the train and test noise matrices $A_{trn}, A_{tst} \in \mathbb{R}^{d\times N}$ be sampled from distributions $\mathcal{D}_{trn}$ and $\mathcal{D}_{tst}$ such that A_{trn} satisfies points $1-4$ below and A_{tst} satisfies points $1, 2$.*

1. *For all i, j, $\mathbb{E}_{\mathcal{D}}[A_{ij}] = 0$, and $\mathbb{E}_{\mathcal{D}}[A_{ij}^2] = \eta^2/d$. Here $\eta = \Theta(1)$ as N grows.*
2. *For all $\{i_1, j_1\} \neq \{i_2, j_2\}$, $\mathbb{E}_{\mathcal{D}}[A_{i_1j_1}A_{i_2j_2}] = \mathbb{E}_{\mathcal{D}}[A_{i_1j_1}]\mathbb{E}_{\mathcal{D}}[A_{i_2j_2}]$.*
3. *$\mathcal{D}$ is a rotationally bi-invariant distribution[3] and $A \sim \mathcal{D}$ is full rank with probability one.*
4. *Suppose $A^{d,N}$ is a sequence of matrices such that with $d/N = c + o(1)$ as N grows, for $c > 0$. Let $\lambda_1^{d,N}, \ldots, \lambda_N^{d,N}$ be the eigenvalues of $(A^{d,N})^T A^{d,N}$. Let $\mu_{d,N} = \sum_i \delta_{\lambda_i^{d,N}}$ be the sum of dirac delta measures for the eigenvalues. Then we shall assume that $\mu_{d,N}$ converges weakly in probability to the Marchenko-Pastur measure with shape c as N grows (see Appendix C).*

[3] A distribution over matrices $A \in \mathbb{R}^{m\times n}$ is rotationally bi-invariant if for all orthogonal $U_1 \in \mathbb{R}^{m\times m}$ and all orthogonal $U_2 \in \mathbb{R}^{n\times n}$, U_1AU_2 has the same distribution as A. Another way to phrase rotational bi-invariance is if the SVD of A is given by $A = U_A\Sigma_AV_A^T$, then U_A and V_A are uniformly random orthogonal matrices and are independent of Σ_A and each other.

Comparison with assumptions in prior work. There are three papers in denoising to compare to, namely [29–31]. Our assumptions on noise are strictly more general than the first two. [31] has the same assumptions as ours, except that they do not require rotational invariance of noise. In contrast to our general closed form results, they analyse learning dynamics for denoising by choosing a specific orthogonal initialization for the coupled ODE that they derive.

B Other Important Insights for Denoising

Double Descent under Distribution Shift Notice that all our curves plotting test error against $1/c$ have a similar shape – they rise when c approaches 1 from either side, and there is a peak at $c = 1$. This matches our theoretical results and establishes that denoising test error curves exhibit double descent, even for arbitrary test data in $\mathcal{V}$. To understand why this is happening, consider the denoising target, given by the MSE error below.

$$\mathbb{E}_{A_{tst}}[\|Y_{trn}-W(X_{trn}+A_{trn})\|_F^2] = \|Y_{trn}-WX_{trn}\|_F^2+2Tr(Y_{trn}-WX_{trn})^T A_{trn})+\|WA_{trn}\|_F^2.$$

The noise is regularizing $\|W\|_F$ through the variance term $Tr(W^TWA_{trn}A_{trn}^T)$. This is the implicit regularization of W due to noise. However, the strength of regularization due to the noise instance A_{trn} is not the same across different values of c. When c is close to 1, the distribution of the spectrum of $A_{trn}A_{trn}^T$ (the Marchenko-Pastur distribution) has support very close to zero. On the other hand, for c far from 1, the non-zero eigenvalues of $A_{trn}A_{trn}^T$ are all bounded away from zero. This establishes that the effect of regularization weakens most near $c = 1$,[4] leading to a spike in the test error coming from the large norm of the learnt W_{opt}. This explanation is similar in spirit to the explanations for double descent in [26] and others, but crucially adapts to implicit regularization due to noise.

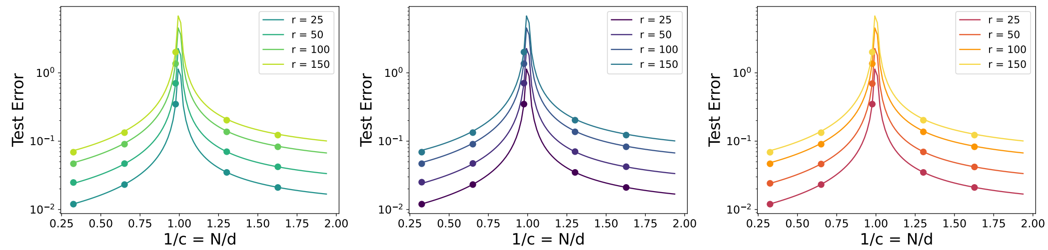
Data Augmentation to Reduce Test Error. In contrast with [32], but similar to [29], optimally picking the noise parameter will not remove the peak in the test error (see Appendix C). Instead, we use data augmentation and increase N to try to move away from the peak, studying Theorem 1 to understand how this will affect test error. We take two approaches to data augmentation that individually exploit the absence of the IID assumptions. Since *the data does not have to be independent*, we can take the same training data and add fresh noise to increase N. Alternatively, since *the data does not have to be sampled from a specific distribution*, we can combine two different datasets into a larger training dataset to increase N. When $c < 1$, applying data augmentation increases N, thus decreasing c further away from the peak at 1 and decreasing test error. When $c > 1$, applying data augmentation increases N, decreasing c towards the peak at 1 and increasing test error.[5] Of course, the latter phenomenon could be mitigated by adding other regularizers or by further augmenting the data. Figures 2 and 3 empirically verify the validity of Theorem 1 for the training data obtained from data augmentation. We also see that increasing the number of in-distribution training data points reduces the out-of-distribution test error.

Benign Overfitting through the Lens of Data Augmentation. Notice we don't observe benign overfitting except in the limit of arbitrarily large or arbitrarily small c. We make sense of this phenomenon using the following argument. Recall that W^* is the minimum-norm optimizer for the *expectation* of the MSE error over noise. Taking the expectation over noise in the training target is in spirit like augmenting the data with "infinitely many" copies of itself, each with fresh noise. So, obtaining W^* is intuitively like training W_{opt} over a dataset with c replaced with a vanishingly small value while keeping $\Sigma_{trn}^2/N = \Sigma_{trn}^2 c/d$ constant. We can compute the effect of this change in c on the test error using Theorem 1, computationally justifying our overfitting phenomena. For intuition, we relate this change in c to the explanation behind double descent. The implicit regularization due to noise is much more unstable for c close to 1. This means that replacing c with a vanishingly small value while keeping the signal-to-noise ratio $\Sigma_{trn}^2/(\eta_{trn}^2 N)$ constant will greatly reduce test error, if we start with c close to 1. On the other hand, the effect of this change in c on the regularization

[4] The eigenvalues that are exactly zero do not contribute to weakening of the regularization. This is because we are choosing the minimum-norm optimizer W^* for expected MSE error, and more zero eigenvalues increases flexibility, creating a larger set of optimizers to minimize the norm over. This helps decrease the components of W^* by spreading them into more dimensions. This is identical in spirit to arguments about variance in overparametrized regimes in section 1.1 of [28].

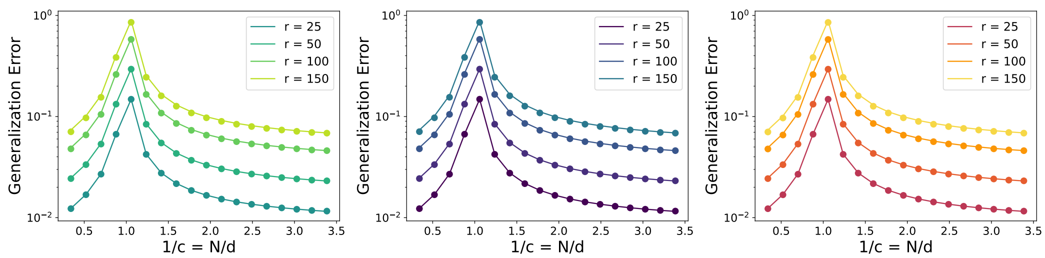
[5] One technicaly also has to account for the effect of data augmentation on Σ_{trn}, but $\Sigma_{trn}^2 c$ can be thought of as constant in this process.

due to noise will be much smaller if we start with an arbitrarily small or arbitrarily large c. So the performance of W^* and W_{opt} is much closer in this case but not when we start with c close to 1. This intuitively explains our overfitting phenomena.



(a) CIFAR Dataset (b) STL10 Dataset (c) SVHN Dataset

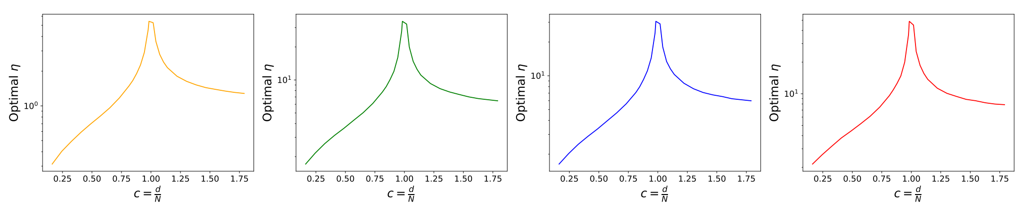
Figure 2: Data augmentation exploiting non-independence. For different N_{trn} the training data is formed by repeating the same 1000 images from the CIFAR dataset.



(a) CIFAR Dataset (b) STL10 Dataset (c) SVHN Dataset

Figure 3: Data augmentation exploiting non-identicality of the distribution. The training data is formed by mixing CIFAR train split with STL10 train split dataset.

C Additional Remarks and Definitions

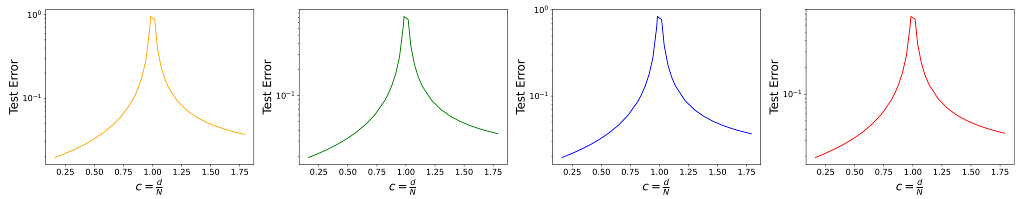


(a) Training Data with New Noise (b) CIFAR Dataset (c) STL10 Dataset (d) SVHN Dataset

Figure 4: Optimal η_{trn} that minimizes the test error given in Theorem 1 versus $c = d/N_{trn}$.

C.1 Extension to non-linear models.

Many prior works [16, 33, 34] study non-linear models using what is known as the Gaussian Equivalence Principle. This is a fact that comes from the Pennington-Worah distribution [35–37] and states the following. Suppose $X \in \mathbb{R}^{d \times N}$ with I.I.D. elements with mean 0 and variance 1 is our data matrix and $W \in \mathbb{R}^{m \times d}$ is a weight matrix with I.I.D. entries with mean zero and variance 1. Let f be any real analytic activation function and let $Y = \frac{1}{\sqrt{N}} f\left(\frac{1}{\sqrt{d}} WX\right)$, then the limiting distribution (as $N, d, m \to \infty, d/n \to \phi, d/m \to \psi$) of the eignevalues of YY^T is the same as the



(a) Training Data with New Noise | (b) CIFAR Dataset | (c) STL10 Dataset | (d) SVHN Dataset

Figure 5: Test Error using Theorem 1 versus $1/c$ with optimal η_{trn}.

limiting distribution of the eigenvalues of

$$\frac{1}{N}\left(\sqrt{\kappa_2(f)}\frac{WX}{\sqrt{d}}+\sqrt{\kappa_1(f)-\kappa_2(f)}Z\right)\left(\sqrt{\kappa_2(f)}\frac{WX}{\sqrt{d}}+\sqrt{\kappa_1(f)-\kappa_2(f)}Z\right)^T.$$

Here Z is a matrix with I.I.D standard normal entries. If we consider the case when $k > d$, we can imagine d being the rank of the data. Then is similar to our case, except that we consider the case when the rank is fixed, whereas here we need the rank to go to infinity proportionally to the number of data points.

C.2 Marchenko-Pastur Distribution

We recall the definition of the Marchenko-Pastur distribution with shape c, for completeness.

Definition 2. *Let $c \in (0, \infty)$ be a shape paramter. Then the Marchenko-Pastur distribution with shape c is the measure μ_c supported on $[c_-, c_+]$, where $c_\pm = (1 \pm \sqrt{c})^2$ is such that*

$$\mu_c = \begin{cases} \left(1-\frac{1}{c}\right)\delta_0 + \nu & c > 1 \\ \nu & c \leq 1 \end{cases}$$

where ν has density

$$d\nu(x) = \frac{1}{2\pi xc}\sqrt{(c_+ - x)(x - c_-)}.$$

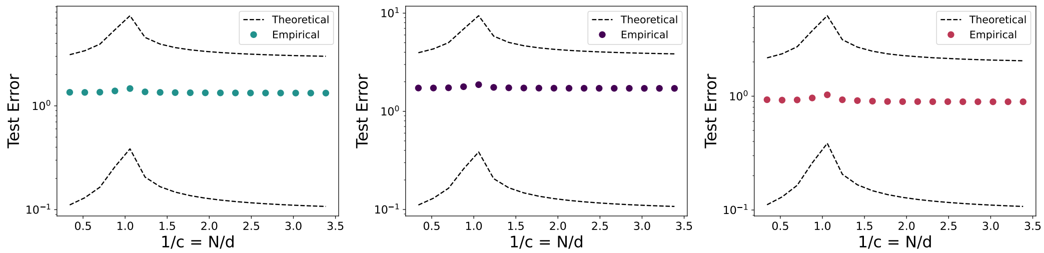
C.3 Amount of Training noise

It was highlighted in [29] that optimally picking the training noise level does not mitigate the double-descent phenomena observed in the generalization error for a linear model. In this section, we support this claim using our result from Theorem 1. Figure 4 shows the double descent curve of η_{trn} and figure 5 shows the generalization error when using the optimal amount of training noise. As in other works such as [29, 38], we see double descent in the regularization strength. As we can see, increasing r decreases α, which improves our bounds.

D Additional Experimental Results

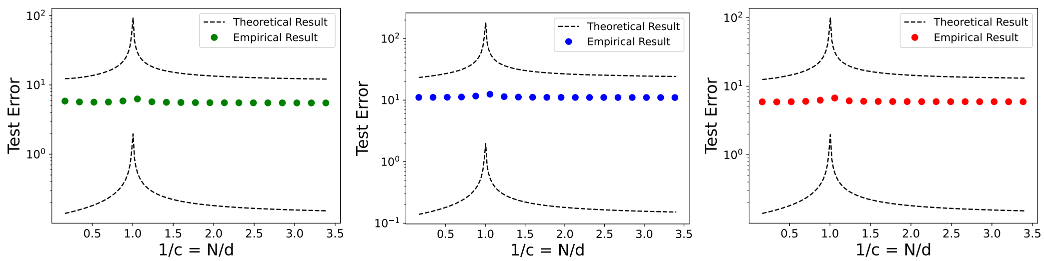
D.1 Detailed Experiments when $\beta = I$

To experimentally verify our test error predictions using real-life data with distribution shift, we train a linear function W_{opt} on CIFAR [39] and test on CIFAR, STL10 [40], and SVHN [41]. For computing test error, we simply compute W_{opt} and plot the empirical average of $\frac{1}{N_{tst}}\|X_{tst} - W_{opt}(X_{tst} + A_{tst})\|_F^2$ over 200 trials. We run three main kinds of experiments. (a) First, to enforce the low-rank assumption to isolate the effect of distribution shift, we use principal component regression or PCR [25, 26]. In PCR, instead of working with the true (and approximately low-rank) training data matrices X_{tst}, we find the best low-rank approximation $\hat{X}_{trn}$ of the training data by projecting it to an embedded subspace of the highest principal components. When testing, we project the test datasets to the same subspace to enforce the low-rank assumption before computing the empirical test error. (b) Second, to explicitly control the amount of deviation α from the low-rank subspace,



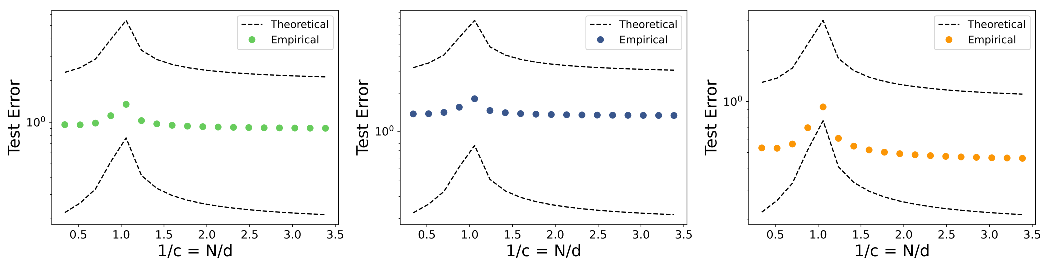
(a) CIFAR Dataset (b) STL10 Dataset (c) SVHN Dataset

(d) r = 25; We find that α is approximately 66, 85 and 44 for (a)-(c) respectively.



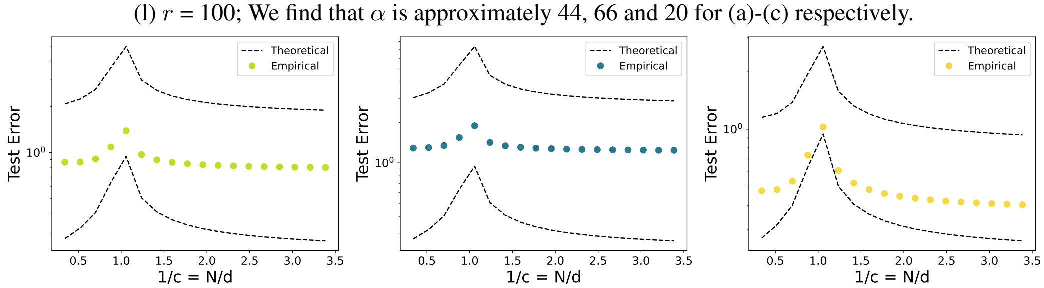
(e) CIFAR Dataset (f) STL10 Dataset (g) SVHN Dataset

(h) r = 50; We find that α is approximately 54, 75 and 31 for (a)-(c) respectively.



(i) CIFAR Dataset (j) STL10 Dataset (k) SVHN Dataset

(l) r = 100; We find that α is approximately 44, 66 and 20 for (a)-(c) respectively.

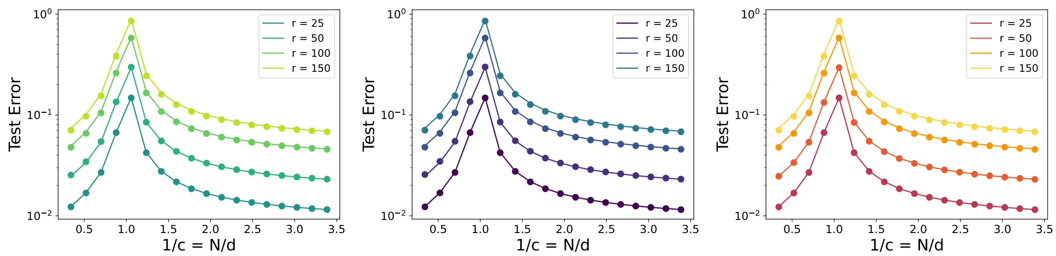


(m) CIFAR Dataset (n) STL10 Dataset (o) SVHN Dataset

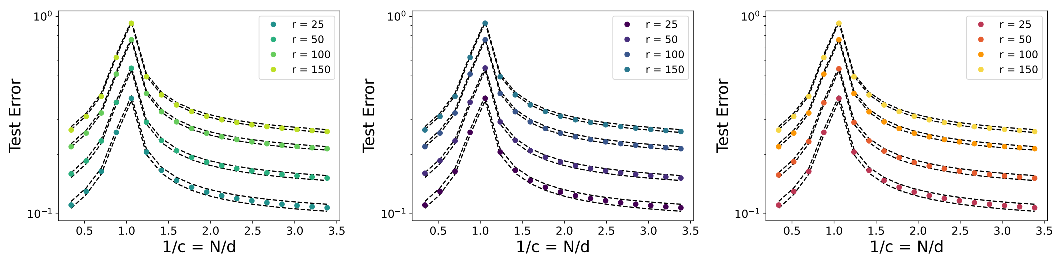
(p) r = 150; We find that α is approximately 37, 60 and 15 for (a)-(c) respectively.

Figure 6: Figure showing the test error vs $1/c$ when the test datasets retain their high dimensions. The training data is projected onto its first r principal components. The markers denote the square root of test error obtained from empirical experiments. The dashed black lines, which act as the upper bounds for the empirical results, are given by $\sqrt{\mathcal{R}(UL)} + \alpha\sigma_1(W_{opt} + I)$ where $\mathcal{R}(UL)$ is the theoretical generalization error (refer Theorem 3). The dashed black lines, which act as the lower bounds, are given by $\sqrt{\mathcal{R}(UL)}$.

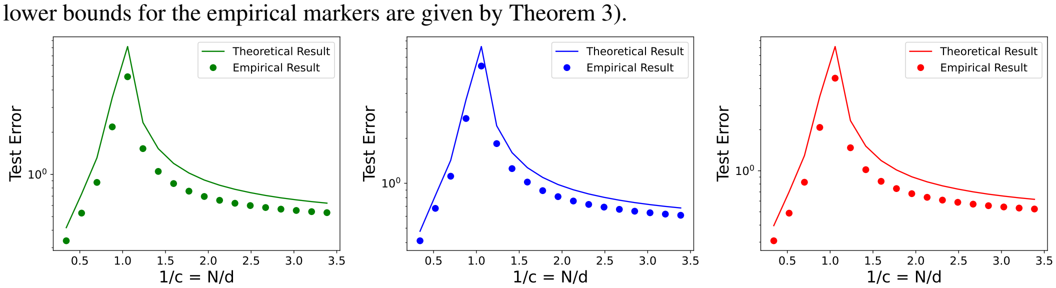
we perturb the low-rank testing data from setting (a) and test using $\tilde{X}_{tst} := \hat{X}_{tst} + K_{tst}$, where K_{tst} is Gaussian noise with covariance designed to control α. (c) Third, we rely on the approximate low-rank nature of real-life data, and report the test error for the matrices X_{tst} themselves. Since d is fixed, we vary c by varying N. Figure 7 shows that the theoretical curves and the empirical results align perfectly for experimental setup (a) and that we have tight bounds for experimental setup (b). Numerically, we find that the relative error between the generalization error estimate and the average empirical error in experimental setup (a) is under 1% on average. For setup (c), since real-life data is only *approximately* low rank, we see a non-negligible error. However, the predictions align well with the empirical results.



(a) In-subspace test error.



(b) For the out-of-subspace curves, we add full-dimensional Gaussian noise such that $\alpha = 0.1$. The upper and lower bounds for the empirical markers are given by Theorem 3).



(c) Test error estimated without projecting data, relying on the approximate low-rank structure of real-life data.

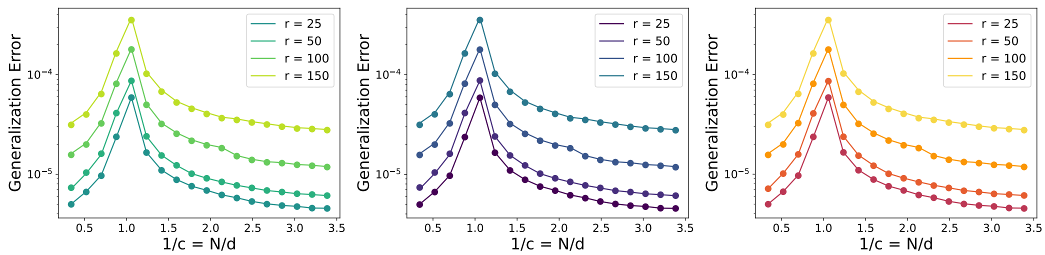
Figure 7: Figures showing the test error for $\beta = I$ vs $1/c = N/d$. In (a) and (b), training data from the CIFAR dataset is projected onto its first r principal components for $r = 25, 50, 100, 150$. 2500 test data points from CIFAR (Green, Left col.), STL10 (Blue, Middle col.), and SVHN (Red, Right col.) datasets are projected onto the same low-dimensional subspace. (a) is in-subspace test error and (b) is out-of-subspace test error. In (c), we don't project the test data and report the standard test error, relying on the approximate low-rank structure in data instead of imposing it. For empirical data points, shown by markers, we report the mean test error over at least 200 trials. Similar results are obtained for single-variable regression with $\beta \in \mathbb{R}^d$ (see Appendix D.2)

D.2 Single-variable Regression

We present analogues for figures in the main paper. See Figure 8.

D.3 Out of subspace PCR for large α

As mentioned in Section 3, we numerically verify Theorem 3 in two out-of-distribution setups namely small α and large α. The application of our result to the small α case was already presented in the main paper; see Figure 6. Here, we present the additional numerical results when the value of α is relatively large. We do not project the test datasets onto the low-dimensional subspace for this. The



(a) CIFAR Dataset (b) STL10 Dataset (c) SVHN Dataset

Figure 8: Figures showing the test error for Linear Regression vs $1/c = N/d$. Training data from the CIFAR dataset is projected onto its first r principal components for $r = 25, 50, 100, 150$. 2500 test data points from CIFAR, STL10, and SVHN datasets are projected onto the same low-dimensional subspace. For empirical data points, shown by markers, we report the mean test error over at least 200 trials.

training dataset from the CIFAR train split is projected onto its first r principal components where $r = 25, 50, 100$ and 150. Figure 6 shows the theoretical bounds on the generalization error from Theorem 3. Unfortunately, for the large α case, the proposed lower bound in Theorem 3 is negative. However, we conjecture that $\mathcal{R}(UL)$ is a lower bound instead. The results for the large α case, shown in Figure 6, suggest the same. However, these bounds do not tell us anything about the shape of the generalization error curve.

E Additional Theoretical Results

E.1 Test Error and Generalization Error

Recall from the introduction that the work of [15] requires the simultaneous diagonalizability of the covariance matrices of training and test data. In a similar spirit, if we assume that the training and test data have the same left singular vectors, we recover the conjectured formula in [29] as an immediate consequence of Theorem 1.

Corollary 3 (Conjecture of [29])**.** *Let the SVD of X_{tst} be $U_{tst}\Sigma_{tst}V_{tst}^T$. In Theorem 1, if we further assume that $U^T U_{tst} = I$, then we can replace L with Σ_{tst} in the expression for the test error.*

Additionally, we can use Theorem 1 to give an expression for generalization error when the test data points are drawn from a distribution, possibly dependently.

Corollary 4 (Generalization Error)**.** *In the setting of Theorem 1, if we further assume that the data X_{tst} is generated possibly dependently from distributions supported in the span of U with mean $U\mu$ and covariance $\Sigma_U = U\Sigma U^T$, then we can remove the $\frac{1}{N_{tst}}$ and replace L with $(\Sigma + \mu\mu^T)^{1/2}$ in the expression for test error to get the generalization error $\mathbb{E}_{X_{tst}}[\mathcal{R}(W_{opt}, X_{tst})]$.*

E.2 Out-of-Distribution Generalization

Consider the following theorem bounding the difference in generalization error in terms of the change in the test set. Our main distribution shift result is a corollary of its proof.

Theorem 2 (Test Set Shift Bound)**.** *Under the assumptions of Theorem 1, consider a linear regressor W_{opt} trained on training data $X_{trn} = U\Sigma_{trn}V_{trn}^T$ with Σ_{trn} such that $\sigma_r(X_{trn}) > M$, and tested on test data $X_{tst,1} = UL_1$ and $X_{tst,2} = UL_2$ with noise $A_{tst,1}, A_{tst,2}$ with the same variance η_{tst^2}/d. Then, the generalization errors $\mathcal{R}_1$ and $\mathcal{R}_2$ differ for $c < 1$ by*

$$|\mathcal{R}_2 - \mathcal{R}_1| \leq \frac{\sigma_1(\beta)^2}{N_{tst}} \frac{\eta_{trn}^4 r}{(\sigma_r(X_{trn})^2 f(c) + \eta_{trn}^2)^2} \|L_2 L_2^T - L_1 L_1^T\|_F + o\left(\frac{1}{N}\right)$$

where $f(c) = c$ for $c < 1$ and $f(c) = 1$ for $c \geq 1$. We add $O(\|\Sigma_{trn}\|_F^2/N^2)$ to the bound when $c > 1$.

E.3 Out-of-Subspace Generalization

Theorem 3 (Out-of-Subspace Shift Bound). *If we have the same training data and solution W_{opt} assumptions as in Theorem 1. Then, for* **any** *X_{tst} for which there exists an L and an $\alpha > 0$ such that $\|X_{tst} - UL\|_F \leq \alpha$, and A_{tst} that satisfies 1,2 from Assumption 2, we have that the generalization error $\mathcal{R}(W_{opt}, X_{tst})$ satisfies*

$$|\mathcal{R}(W_{opt}, X_{tst}) - \mathcal{R}(W_{opt}, UL)| \leq \alpha^2 \sigma_1(W_{opt} + I)^2.$$

The following corollary follows immediately from Theorem 3 and Theorem 2.

Corollary 5. *If $X_{tst,1}$ and $X_{tst,2}$ are two different test datasets and $X_{trn} = U\Sigma_{trn}V_{trn}^T$ is the training data such that there exists L_i with $\alpha_i = \|X_{tst,i} - UL_i\|_F$, then for $\mathcal{R}_i := \mathcal{R}(W_{opt}, X_{tst,i})$*

$$|\mathcal{R}_2 - \mathcal{R}_1| \leq (\alpha_1^2 + \alpha_2^2)\sigma_1(W_{opt} + I)^2 + \frac{\sigma_1(\beta)^2}{N_{tst}} \frac{\eta_{trn}^4 r}{(\sigma_r(X_{trn})^2 f(c) + \eta_{trn}^2)^2} \|L_2 L_2^T - L_1 L_1^T\|_F + o\left(\frac{1}{N}\right)$$

E.4 Overfitting Paradigms

The following theorem and its proof are used to prove Corollary 2. The proofs are in Appendix F.5

Theorem 4 (Test Error for W^*). *In the same setting as Theorem 1, we have that $W^* = \beta_U^T \left(I + \frac{\eta_{trn}^2}{c}\Sigma_{trn}^{-2}\right)^{-1} U^T$ and*

$$\mathcal{R}(W^*, UL) = \frac{\eta_{trn}^4 N^2}{d^2} \left\| \beta_U^T \left(\Sigma_{trn}^2 + \frac{\eta_{trn}^2 N}{d} I\right)^{-1} L \right\|_F^2 + \frac{\eta_{tst}^2}{d} Tr\left(\beta_U \beta_U^T \left(I + \frac{\eta_{trn}^2 N}{d}\Sigma_{trn}^{-2}\right)^{-2}\right).$$

E.5 Independent Identical Test data

Let us assume that the test data is identically and independently drawn from some distribution $\mathcal{D}_{tst}$ with mean zero and covariance Σ. Then the generalization error is given by the following corollary.

Corollary 6 (IID Test Data). *In the setting of Theorem 1, if we further assume that the columns of L are drawn IID from a distribution with mean zero and Covariance Σ, then we can remove the $\frac{1}{N_{tst}}$ and replace L with $\Sigma^{1/2}$ in the expression for the generalization risk.*

Remark 1. *Given any distribution on $\mathcal{V}$, we can consider the diffeomorphism that changes the basis to U. Hence, making assumptions on the distribution of L versus the distribution of X_{tst} does not cost us any generality.*

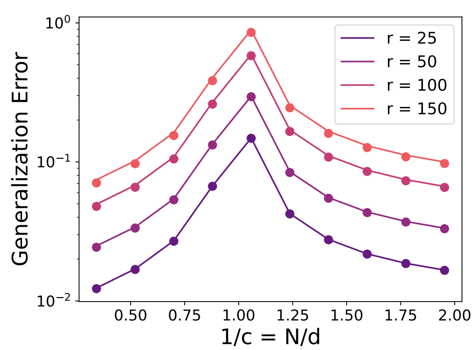
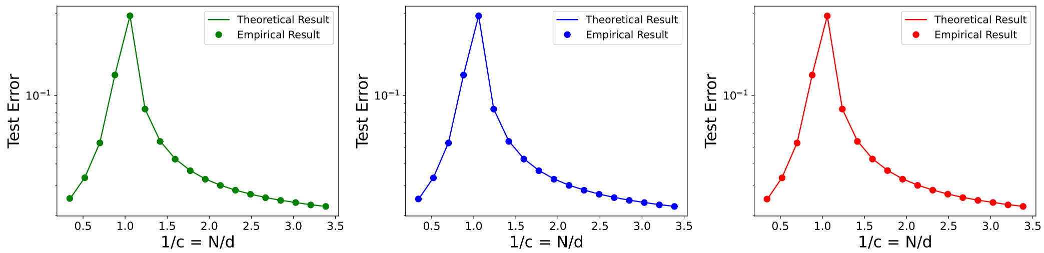


Figure 9: Figure showing the generalization error vs 1/c obtained for IID test data for $r = 25, 50, 100, 150$. The theoretical solid line curve is given by Corollary 6. We report the mean generalization error over at least 200 trials for empirical data points, shown by markers.

Figure 9, shows that the theoretical error aligns perfectly with the empirical result. The model is trained on the CIFAR dataset and tested on data drawn from an anisotropic Gaussian. The case of IID training data is presented in Appendix E.6.

E.6 Independent Isotropic Identical Training Data

Next, we consider the case of I.I.D training data. Let $U \in \mathbb{R}^{d\times r}$ be a matrix whose columns form an orthonormal basis for an r-dimensional space $\mathcal{V}$. Suppose the data is of the form Uz for $z \in \mathbb{R}^r$ such that the coordinates of z are sampled independently, have mean 0, variance $1/r$, and have bounded forth moments. Hence, in this case, we get the following theorem. Proof in Section F.7.



(a) CIFAR Dataset (b) STL10 Dataset (c) SVHN Dataset

Figure 10: Figure showing the test error vs 1/c for I.I.D. training data. The theoretical solid curves are obtained from the formula in Theorem 5. We report the mean test error over at least 200 trials for empirical data points, shown by markers.

Theorem 5 (I.I.D. Training Data With Isotropic Covariance)**.** *Let $c = d/N$ and $c_r = r/N$. Then if $c < 1$*

$$\mathbb{E}_{X_{trn}}[\mathcal{R}] = \frac{\eta_{trn}^4}{N_{tst}} \|(\Sigma_{trn}^2 c + \eta_{trn}^2 I)^{-1} L\|_F^2 + \eta_{tst}^2 \frac{r}{d}\frac{1}{1-c}\left(T_1(c_r, \eta_{trn}^2/c) + \frac{1}{\eta_{trn}^2} T_2(c_r, \eta_{trn}^2/c)\right) + o\left(\frac{1}{N}\right)$$

and if $c > 1$

$$\mathbb{E}_{X_{trn}}[\mathcal{R}] = \frac{\eta_{trn}^4}{N_{tst}} \|(\Sigma_{trn}^2 + \eta_{trn}^2 I)^{-1} L\|_F^2 + \eta_{tst}^2 \frac{r}{d}\frac{c}{c-1} T_3(c_r, \eta_{trn}^2) + O\left(\frac{1}{N}\right)$$

where $T_1(c_r, z) = T_3(cr, z) - zT_2(cr, z)$, and

$$T_2(c_r, z) = \frac{1 + c_r + zc_r}{2\sqrt{(1 - c_r + c_r z)^2 + 4c_r^2 z}} - \frac{1}{2},\ T_3(c_r, z) = \frac{1}{2} + \frac{1 + zc_r - \sqrt{(1 - c_r + zc_r)^2 + 4c_r^2 z}}{2c_r}.$$

Figure 10 shows that the theoretical curves align perfectly with the empirical results where the training data is I.I.D. from a Gaussian with dimension 50. The test datasets from CIFAR, STL10, and SVHN datasets are also projected onto the low-dimensional subspace.

I.I.D Test and Training Data We can combine the two cases where training and test data are I.I.D.. Specifically, for the case when X_{tst} has κI as the covariance and X_{trn} is as in the previous instantiation Section. Then the generalization error is given by the following corollary.

Corollary 7 (I.I.D. Train and Tests Data With Isotropic Covariance)**.** *Let $c = d/N$ and $c_r = r/N$. Then if $c < 1$*

$$\mathbb{E}_{X_{trn}}[\mathcal{R}] = \eta_{trn}^4 \cdot r \cdot \kappa \cdot T_4(c_r, \eta_{trn}^2/c) + \frac{r}{d}\frac{1}{1-c}\left(T_1(c_r, \eta_{trn}^2/c) + \frac{1}{\eta_{trn}^2} T_2(c_r, \eta_{trn}^2/c)\right) + o\left(\frac{1}{N}\right)$$

and if $c > 1$

$$\mathbb{E}_{X_{trn}}[\mathcal{R}] = \eta_{trn}^4 \cdot r \cdot \kappa \cdot T_4(c_r, \eta_{trn}^2) + \frac{r}{d}\frac{c}{c-1} T_3(c_r, \eta_{trn}^2) + O\left(\frac{1}{N}\right)$$

where $T_1(c_r, z) = T_3(c_r, z) - zT_2(c_r, z)$, and

$$T_2(c_r, z) = \frac{1 + c_r + zc_r}{2\sqrt{(1 - c_r + c_r z)^2 + 4c_r^2 z}} - \frac{1}{2},\ T_3(c_r, z) = \frac{1}{2} + \frac{1 + zc_r - \sqrt{(1 - c_r + zc_r)^2 + 4c_r^2 z}}{2c_r},$$

$$T_4(c_r, z) = \frac{zc_r^2 + c_r^2 + zc_r - 2c_r + 1}{2z^2 c_r \sqrt{(1 - c_r + c_r z)^2 + 4c_r^2 z}} - \frac{1}{2z^2}\left(1 - \frac{1}{c_r}\right).$$

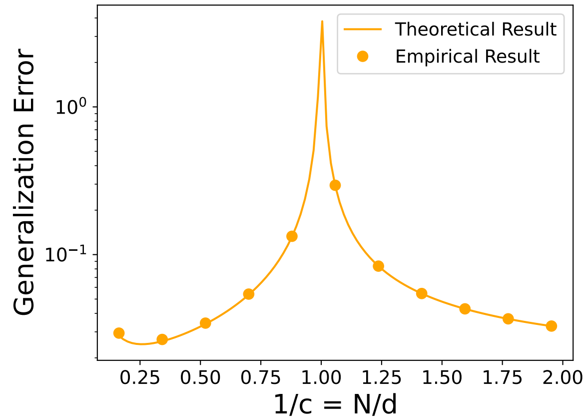


Figure 11: Figure showing the generalization error vs 1/c where training and test datasets are both I.I.D. The theoretical solid curve is obtained from Corollary 8. The empirical generalization error, shown by markers, is averaged over 50 trials.

Figure 11 shows that the theoretical error aligns perfectly with the empirical result.

Similar to the denoising case, we have the following versions for single-variable regression.

Theorem 6 (I.I.D. Training Data With Isotropic Covariance)**.** *Let $c = d/N$ and $c_r = r/N$. Let $\|\beta_{opt}\| = 1$. Then if $c < 1$*

$$\mathbb{E}_{X_{trn}}[\mathcal{R}] = \frac{\eta_{trn}^4}{N_{tst}}\|\hat{\beta}^T(\Sigma_{trn}^2 c + \eta_{trn}^2 I)^{-1}L\|_F^2 + \eta_{tst}^2 \frac{r}{d}\frac{1}{1-c}\left(T_1(c_r, \eta_{trn}^2/c) + \frac{1}{\eta_{trn}^2}T_2(c_r, \eta_{trn}^2/c)\right) + o\left(\frac{1}{N}\right)$$

and if $c > 1$

$$\mathbb{E}_{X_{trn}}[\mathcal{R}] = \frac{\eta_{trn}^4}{N_{tst}}\|\hat{\beta}^T(\Sigma_{trn}^2 + \eta_{trn}^2 I)^{-1}L\|_F^2 + \eta_{tst}^2 \frac{r}{d}\frac{c}{c-1}T_3(c_r, \eta_{trn}^2) + O\left(\frac{1}{N}\right)$$

where $T_1(c_r, z) = T_3(cr, z) - zT_2(cr, z)$, and

$$T_2(c_r, z) = \frac{1 + c_r + zc_r}{2\sqrt{(1 - c_r + c_r z)^2 + 4c_r^2 z}} - \frac{1}{2},\ T_3(c_r, z) = \frac{1}{2} + \frac{1 + zc_r - \sqrt{(1 - c_r + zc_r)^2 + 4c_r^2 z}}{2c_r}.$$

Corollary 8 (I.I.D. Train and Tests Data With Isotropic Covariance)**.** *Let $c = d/N$ and $c_r = r/N$. Let $\|\beta_{opt}\| = 1$. Then if $c < 1$*

$$\mathbb{E}_{X_{trn}}[\mathcal{R}] = \eta_{trn}^4 r\kappa T_4(c_r, \eta_{trn}^2/c) + \frac{r}{d}\frac{1}{1-c}\left(T_1(c_r, \eta_{trn}^2/c) + \frac{1}{\eta_{trn}^2}T_2(c_r, \eta_{trn}^2/c)\right) + o\left(\frac{1}{N}\right)$$

and if $c > 1$

$$\mathbb{E}_{X_{trn}}[\mathcal{R}] = \eta_{trn}^4 r\kappa T_4(c_r, \eta_{trn}^2) + \frac{r}{d}\frac{c}{c-1}T_3(c_r, \eta_{trn}^2) + O\left(\frac{1}{N}\right)$$

where $T_1(c_r, z) = T_3(c_r, z) - zT_2(c_r, z)$, and

$$T_2(c_r, z) = \frac{1 + c_r + zc_r}{2\sqrt{(1 - c_r + c_r z)^2 + 4c_r^2 z}} - \frac{1}{2},\ T_3(c_r, z) = \frac{1}{2} + \frac{1 + zc_r - \sqrt{(1 - c_r + zc_r)^2 + 4c_r^2 z}}{2c_r},$$

$$T_4(c_r, z) = \frac{zc_r^2 + c_r^2 + zc_r - 2c_r + 1}{2z^2 c_r\sqrt{(1 - c_r + c_r z)^2 + 4c_r^2 z}} - \frac{1}{2z^2}\left(1 - \frac{1}{c_r}\right).$$

F Proofs

In all proofs, WLOG we assume $d/N = c$ since even though $d/N = c + o(1)$, the *relative* error we will accumulate from this assumption be $o(1)$. For instance, this means that the absolute error from this assumption in Theorem 1 will be $o(1/N)$, which can be absorbed into the $o(1/N)$ estimation error in the theorem.

F.1 Proof for Theorem 1, Test Error

One useful piece of notation for the following proof is that of big O in probability.

Definition 3. *Let χ_k be a sequence of random variables. Then we say that χ_k is $O_P(a_k)$ as $k \to \infty$, if for all $\epsilon > 0$, we have there exists an M and K such that for all $k > K$, we have that*

$$\Pr\left[\left|\frac{\chi_k}{a_k}\right| > M\right] < \epsilon.$$

Definition 4. *Let χ_k be a sequence of random variables. Then we say that χ_k is $o_P(a_k)$ as $k \to \infty$, if for all $\epsilon > 0$, we have that*

$$\lim_{k\to\infty} \Pr\left[\left|\frac{\chi_k}{a_k}\right| \geq \epsilon\right] = 0.$$

Note that big-O_P behaves a lot like big-O. Specifically, if $\alpha_n = O_P(a_n)$ and $\beta_n = O_P(b_n)$. Then $\alpha_n\beta_n = O_P(a_n b_n)$ and $\alpha_n + \beta_n = O_P(a_n + b_n)$. Further, it is easy to see that mean zero random variables are big-O_P of the square root of the variance (using Chebyshev's inequality).

F.1.1 The Overparametrized Regime, $d > N$

We derive test error bounds for $\beta = I$ in our problem setting. We also denote W_{opt} by W in this subsection, for ease of notation.

Theorem 7. *For rank r data and $d > N + r$, with $c = \frac{d}{N}$ the following is true.*

1. *For the $\beta = I$ case, we denote the minimum norm linear denoiser W_{opt} by just W in this subsection. It is given by*

$$W = U\Sigma_{trn}(P^T P)^{-1} Z^T K_1^{-1} H - U\Sigma_{trn} Z^{-1} H H^T K_1^{-1} Z P^\dagger$$

2. *The test error when $X_{tst} = UL$ is given by*

$$\mathbb{E}_{A_{trn}}\left[\frac{1}{N_{tst}}\|U\Sigma_{trn}(P^T P)^{-1} Z^T K_1^{-1}\Sigma_{trn}^{-1} L\|_F^2 + \frac{\eta_{tst}^2}{d}\|W\|_F^2\right],$$

where $P = -(I - A_{trn}A_{trn}^\dagger)U\Sigma_{trn}$, $H = V_{trn}^T A_{trn}^\dagger$, $Z = I + V_{trn}^T A_{trn}^\dagger U\Sigma_{trn}$, $K_1 = HH^T + Z(P^T P)^{-1} Z^T$.

The sizes of the matrices:

1. U is $d \times r$ with $U^T U = I_{r\times r}$.
2. Σ_{trn} is $r \times r$, with rank r.
3. A_{trn} is $d \times N$ with rank N.
4. $A_{trn}A_{trn}^\dagger$ is $d \times d$
5. H is $r \times d$, with rank r.
6. Z is $r \times r$, with rank r.
7. K_1 is $r \times r$, with rank r.
8. $A_{trn} = \eta_{trn}\tilde{U}\tilde{\Sigma}\tilde{V}^T$.
9. $\tilde{U}$ is $d \times d$ unitary.
10. $\tilde{\Sigma}$ is $d \times N$.

Proof. Part 1 follows from Lemma 1. For part 2, note that the test error is given by $\mathcal{R}(W, X_{tst}) = \mathbb{E}_{A_{trn}, A_{tst}}\left[\frac{1}{N_{tst}}\|X_{tst} - W(X_{tst} + A_{tst})\|_F^2\right]$, which is the same as the folllowing.

$$
\begin{aligned}
\mathcal{R}(W, X_{tst}) &= \frac{1}{N_{tst}}\mathbb{E}_{A_{trn}, A_{tst}}\left[\|X_{tst} - WX_{tst}\|_F^2\right] + \frac{2}{N_{tst}}\mathbb{E}_{A_{trn}, A_{tst}}[Tr((X_{tst} - WX_{tst})A_{tst}) \\
&\quad + \frac{1}{N_{tst}}\mathbb{E}_{A_{trn}, A_{tst}}\left[\|WA_{tst}\|_F^2\right] \\
&= \frac{1}{N_{tst}}\mathbb{E}_{A_{trn}}\left[\|X_{tst} - WX_{tst}\|_F^2\right] + 0 + \frac{1}{N_{tst}}\mathbb{E}_{A_{trn}, A_{tst}}\left[Tr(W^TWA_{tst}A_{tst}^T)\right] \\
&= \frac{1}{N_{tst}}\mathbb{E}_{A_{trn}}\left[\|X_{tst} - WX_{tst}\|_F^2\right] + 0 + \frac{1}{N_{tst}}\mathbb{E}_{A_{trn}}\left[Tr(W^TW\mathbb{E}_{A_{tst}}\left[A_{tst}A_{tst}^T\right])\right] \\
&= \frac{1}{N_{tst}}\mathbb{E}_{A_{trn}}\left[\|X_{tst} - WX_{tst}\|_F^2\right] + 0 + \frac{\eta_{tst}^2 N_{tst}}{dN_{tst}}\mathbb{E}_{A_{trn}}[Tr(W^TW)] \\
&= \mathbb{E}_{A_{trn}}\left[\frac{1}{N_{tst}}\|U\Sigma_{trn}(P^TP)^{-1}Z^TK_1^{-1}\Sigma_{trn}^{-1}L\|_F^2 + \frac{\eta_{tst}^2}{d}\|W\|_F^2\right].
\end{aligned}
$$

□

We will henceforth drop the subscript A_{trn} in the expectation $\mathbb{E}_{A_{trn}}$.

Lemma 1. *Let $P = -(I - A_{trn}A_{trn}^\dagger)U\Sigma_{trn},\ H = V_{trn}^TA_{trn}^\dagger,\ Z = I + V_{trn}^TA_{trn}^\dagger U\Sigma_{trn},\ K_1 = HH^T + Z(P^TP)^{-1}Z^T$. If $d > N$ and A_{trn} has full column rank, then*

$$W = U\Sigma_{trn}(P^TP)^{-1}Z^TK_1^{-1}H - U\Sigma_{trn}Z^{-1}HH^TK_1^{-1}ZP^\dagger. \tag{2}$$

Proof. Note that P has full column rank and A_{trn} has rank N. Thus, we can use corollary 2.2 from Wei [42] to obtain

$$(A_{trn}+U\Sigma_{trn}V_{trn}^T)^\dagger = A_{trn}^\dagger+A_{trn}^\dagger U\Sigma_{trn}P^\dagger-(A_{trn}^\dagger H^T+A_{trn}^\dagger U\Sigma_{trn}(P^TP)^{-1}Z^T)K_1^{-1}(H+ZP^\dagger).$$

We are interested in simiplifying the expression for $W = (U\Sigma_{trn}V_{trn}^T)(A_{trn} + U\Sigma_{trn}V_{trn}^T)^\dagger$. Multiplying this through, we obtain

$$
\begin{aligned}
W = U\Sigma_{trn}V_{trn}^TA_{trn}^\dagger + U\Sigma_{trn}V_{trn}^TA_{trn}^\dagger U\Sigma_{trn}P^\dagger \\
- U\Sigma_{trn}V_{trn}^T(A_{trn}^\dagger H^T + A_{trn}^\dagger U\Sigma_{trn}(P^TP)^{-1}Z^T)K_1^{-1}(H + ZP^\dagger).
\end{aligned}
$$

Replacing $V_{trn}^TA_{trn} = H$,

$$
\begin{aligned}
W = U\Sigma_{trn}H + U\Sigma_{trn}HU\Sigma_{trn}P^\dagger - U\Sigma_{trn}V_{trn}^T(A_{trn}^\dagger H^TK_1^{-1}H + A_{trn}^\dagger H^TK_1^{-1}ZP^\dagger \\
+ A_{trn}^\dagger U\Sigma_{trn}(P^TP)^{-1}Z^TK_1^{-1}H + A_{trn}^\dagger U\Sigma_{trn}(P^TP)^{-1}Z^TK_1^{-1}ZP^\dagger).
\end{aligned}
$$

Through further simplification, we obtain

$$
\begin{aligned}
W = U\Sigma_{trn}H + U\Sigma_{trn}HU\Sigma_{trn}P^\dagger - U\Sigma_{trn}HH^TK_1^{-1}H - U\Sigma_{trn}HH^TK_1^{-1}ZP^\dagger \\
- U\Sigma_{trn}HU\Sigma_{trn}(P^TP)^{-1}Z^TK_1^{-1}H - U\Sigma_{trn}HU\Sigma_{trn}(P^TP)^{-1}Z^TK_1^{-1}ZP^\dagger.
\end{aligned}
$$

Setting $HU\Sigma_{trn} = Z - I$ yields

$$
\begin{aligned}
W = U\Sigma_{trn}H + U\Sigma_{trn}ZP^\dagger - U\Sigma_{trn}P^\dagger - U\Sigma_{trn}HH^TK_1^{-1}H - U\Sigma_{trn}HH^TK_1^{-1}ZP^\dagger \\
- U\Sigma_{trn}Z(P^TP)^{-1}Z^TK_1^{-1}H + U\Sigma_{trn}(P^TP)^{-1}Z^TK_1^{-1}H \\
- U\Sigma_{trn}Z(P^TP)^{-1}Z^TK_1^{-1}ZP^\dagger + U\Sigma_{trn}(P^TP)^{-1}Z^TK_1^{-1}ZP^\dagger.
\end{aligned}
$$

Combining terms and replacing $HH^T + Z(P^TP)^{-1}Z^T = K_1$, we prove

$$
\begin{aligned}
W &= -U\Sigma_{trn}P^\dagger + U\Sigma_{trn}(P^TP)^{-1}Z^TK_1^{-1}H + U\Sigma_{trn}(P^TP)^{-1}Z^TK_1^{-1}ZP^\dagger, \\
&= U\Sigma_{trn}(P^TP)^{-1}Z^TK_1^{-1}H - U\Sigma_{trn}Z^{-1}(K_1 - Z(P^TP)^{-1}Z^T)K_1^{-1}ZP^\dagger, \\
&= U\Sigma_{trn}(P^TP)^{-1}Z^TK_1^{-1}H - U\Sigma_{trn}Z^{-1}HH^TK_1^{-1}ZP^\dagger.
\end{aligned}
$$

□

Lemma 2. *For $d > N + r$, $X_{tst} - WX_{tst} = U\Sigma_{trn}(P^TP)^{-1}Z^TK_1^{-1}\Sigma_{trn}^{-1}L$.*

Proof. Here, $X_{tst} = UL$ and W is given by equation 2. Substituting this, we get

$$X_{tst} - WX_{tst} = UL - U\Sigma_{trn}(P^TP)^{-1}Z^TK_1^{-1}HUL + U\Sigma_{trn}Z^{-1}HH^TK_1^{-1}ZP^{\dagger}UL.$$

Note that $P^{\dagger}U = -\Sigma_{trn}^{-1}$ and $HU\Sigma_{tst} = V_{trn}^TA_{trn}^{\dagger}U\Sigma_{trn}\Sigma_{trn}^{-1}\Sigma_{tst} = (Z-I)\Sigma_{trn}^{-1}\Sigma_{tst}$ which yields

$$\begin{aligned}
X_{tst} - WX_{tst} &= UL - U\Sigma_{trn}(P^TP)^{-1}Z^TK_1^{-1}(Z-I)\Sigma_{trn}^{-1}L - U\Sigma_{trn}Z^{-1}HH^TK_1^{-1}Z\Sigma_{trn}^{-1}L,\\
&= U\Sigma_{trn}Z^{-1}(Z - Z(P^TP)^{-1}Z^TK_1^{-1}(Z-I) - HH^TK_1^{-1}Z)\Sigma_{trn}^{-1}L,\\
&= U\Sigma_{trn}Z^{-1}(Z - (Z-I) + HH^TK_1^{-1}(Z-I) - HH^TK_1^{-1}Z)\Sigma_{trn}^{-1}L,\\
&= U\Sigma_{trn}Z^{-1}(K_1 - HH^T)K_1^{-1}\Sigma_{trn}^{-1}L,\\
&= U\Sigma_{trn}(P^TP)^{-1}Z^TK_1^{-1}\Sigma_{trn}^{-1}L.
\end{aligned}$$

□

Lemma 3. *For $c > 1$, we have that*

$$\mathbb{E}[HH^T] = \frac{c}{\eta_{trn}^2(c-1)}I_r + o(1)$$

and the variance of each entry is $O(1/(\eta_{trn}^4N))$. For $c < 1$, we have that

$$\mathbb{E}[HH^T] = \frac{c^2}{\eta_{trn}^2(1-c)}I_r + o(1)$$

and the variance is $O(1/(\eta_{trn}^4d))$.

Proof. Here we see that

$$HH^T = V_{trn}^TA_{trn}^{\dagger}(A_{trn}^{\dagger})^TV_{trn} = V_{trn}^T(A_{trn}^TA_{trn})^{\dagger}V_{trn}.$$

Thus, if $V_{trn} = [v_1 \cdots v_r]$. Then we see that HH^T is an $r \times r$ matrix such that

$$(HH^T)_{ij} = v_i^T(A_{trn}^TA_{trn})^{\dagger}v_j.$$

Using ideas from [29], we see that if $i \neq j$, then we see that the expectation is 0. On the other hand if $i = j$, then using Lemma 6 from [29], with $p = N$, $q = d$ and $A = \frac{1}{\eta_{trn}}A_{trn}$, we get that for $c > 1$

$$\mathbb{E}[v_i^T(A_{trn}^TA_{trn})^{\dagger}v_i] = \frac{c}{\eta_{trn}^2(c-1)} + o(1).$$

while for $c < 1$

$$\mathbb{E}[v_i^T(A_{trn}^TA_{trn})^{\dagger}v_i] = \frac{c^2}{\eta_{trn}^2(1-c)} + o(1).$$

For the variance, let $A_{trn} = \eta_{trn}\tilde{U}\tilde{\Sigma}\tilde{V}^T$, then we have that

$$\begin{aligned}
v_i^T(A_{trn}^TA_{trn})^{\dagger}v_j &= \frac{1}{\eta_{trn}^2}v_i^T\tilde{V}\tilde{\Sigma}^2\tilde{V}^Tv_j\\
&= \frac{1}{\eta_{trn}^2}a^T\tilde{\Sigma}^2b\\
&= \sum_{i=1}^{N}\frac{1}{\eta_{trn}^2}\frac{1}{\tilde{\sigma}_i^2}a_ib_i.
\end{aligned}$$

Where a, b are orthogonal vectors (when $i \neq j$). Then for computing the variance when $c > 1$,

$$\begin{aligned}
\mathbb{E}\left[\left(v_i^T (A_{trn}^T A_{trn})^\dagger v_j\right)^2\right] &= \mathbb{E}\left[\left(\frac{1}{\eta_{trn}^2}\sum_{i=1}^N \frac{1}{\tilde{\sigma}_i^2} a_i b_i\right)^2\right] \\
&= \frac{1}{\eta_{trn}^4}\mathbb{E}\left[\sum_{i=1}^N \sum_{j=1}^N \frac{1}{\tilde{\sigma}_i^2 \tilde{\sigma}_j^2} a_i b_i a_j b_j\right] \\
&= \left(\frac{c^2}{\eta_{trn}^4 (c-1)^2} + o(1)\right)\mathbb{E}\left[\left(\sum_{i=1}^N a_i b_i\right)\left(\sum_{j=1}^N a_j b_j\right)\right] \\
&\quad + \left(\frac{c^3}{\eta_{trn}^4 (c-1)^3} - \frac{c^2}{\eta_{trn}^4 (c-1)^2} + o(1)\right)\sum_{i=1}^N \mathbb{E}[a_i^2 b_i^2] \\
&= 0 + \left(\frac{c^2}{\eta_{trn}^4 (c-1)^3} + o(1)\right)\sum_{i=1}^N \frac{1}{N^2} + o\left(\frac{1}{N}\right) \\
&= \frac{c^2}{\eta_{trn}^4 (c-1)^3}\frac{1}{N} + o\left(\frac{1}{N}\right).
\end{aligned}$$

Here even though a, b are not independent, because of the smaller variance in the entries, the error is absorbed in the $o\left(\frac{1}{N}\right)$ term.

When $i = j$, we use the same proof [29], to see that the variance is at most

$$\frac{c^2(2c-1)}{\eta_{trn}^4 (c-1)^3}\frac{1}{N} + o\left(\frac{1}{N}\right).$$

A very similar computation follows for the variance when $c < 1$. □

We prove a general result on inverses of matrices that whose expected norms are $\Omega(1)$.

Lemma 4. *If* $\|\mathbb{E}[X_N]\| = \Omega(1)$ *as* N *grows and* $\mathrm{Var}((X_N)_{ij}) = s_N$*, then* $\mathbb{E}[X_N^{-1}] = \mathbb{E}[X_N]^{-1} + O(s_N)$*. Additionally, if* $\mathrm{Var}((X_N - \mathbb{E}[X_N])_{ij}^2) = O(t_N)$*, then* $\mathrm{Var}((X_N^{-1})_{ij}) = O(s_N + t_N)$*.*

Proof. Let $\delta X_N = X_N - \mathbb{E}[X_n]$. Notice that $\delta X_N = O_P(s_N)$ and $\mathbb{E}[\delta X_N] = 0$. Additionally, by the Taylor expansion $(Y + \delta Y)^{-1} = Y^{-1} + Y^{-1}\delta Y Y^{-1} + O(\delta Y^2)$ we have that

$$X_N^{-1} = \mathbb{E}[X_N]^{-1} + \mathbb{E}[X_N]^{-1}\delta X_N \mathbb{E}[X_N]^{-1} + O(\delta X_N^2).$$

In particular, since $\mathbb{E}[X_n]^{-1} = O(1)$, we have

$$O(\mathbb{E}[X_N^{-1}] = \mathbb{E}[X_N]^{-1} + O(\mathrm{Var}((X_N)_{ij})) = \mathbb{E}[X_N]^{-1} + O(s_N).$$

Finally, note that $\mathrm{Var}((\delta X_N^2)_{ij}) = O(t_N)$ by assumption. So,

$$\mathrm{Var}((X_N^{-1})_{ij}) = \mathrm{Var}((\mathbb{E}[X_N]^{-1}\delta X_N \mathbb{E}[X_N]^{-1})_{ij}) + O(\mathrm{Var}((\delta X_N^2)_{ij})) = O(s_N + t_N)$$

since $\mathbb{E}[X_N]^{-1} = O(1)$. □

Lemma 5. *For* $c > 1$*, we claim that* $\mathbb{E}[\Sigma_{trn}^{-1} P^T P \Sigma_{trn}^{-1}] = \left(1 - \frac{1}{c}\right) I_r$*, each entry has variance* $O\left(\frac{1}{d}\right)$*, and*

$$\mathbb{E}[\Sigma_{trn}(P^T P)^{-1}\Sigma_{trn}] = \frac{c}{c-1} I_r + O\left(\frac{1}{d}\right).$$

with element-wise variance $O(1/d)$*.*

Proof. Recall that $P = -(I - A_{trn}A_{trn}^\dagger)U\Sigma_{trn}$ Thus, we have that

$$\begin{aligned}P^TP &= \Sigma_{trn}^T U^T(I - A_{trn}A_{trn}^\dagger)U\Sigma_{trn}.\\ &= \Sigma_{trn}^T\Sigma_{trn} - \Sigma_{trn}^T U^T A_{trn}A_{trn}^\dagger U\Sigma_{trn}\\ &= \Sigma_{trn}^T\Sigma_{trn} - \Sigma_{trn}^T U^T \tilde{U}\tilde{\Sigma}\tilde{\Sigma}^\dagger\tilde{U}^T U\Sigma_{trn}\\ &= \Sigma_{trn}^T\Sigma_{trn} - \Sigma_{trn}^T R\begin{bmatrix} I_N & 0\\ 0 & 0_{d-N}\end{bmatrix} R^T\Sigma_{trn}.\end{aligned}$$

Where R is a uniformly random $r \times d$ unitary matrix. Then by symmetry (of the sign of rows of R), we have that

$$\mathbb{E}[P^TP] = \Sigma_{trn}^2 - \Sigma_{trn}^T\left(\frac{1}{c}I_r\right)\Sigma_{trn} = \left(1 - \frac{1}{c}\right)\Sigma_{trn}^2.$$

So, we have that

$$\mathbb{E}\left[\Sigma_{trn}^{-1}P^TP\Sigma_{trn}^{-1}\right] = U^T\left(I - \mathbb{E}\left[A_{trn}A_{trn}^\dagger\right]\right)U = \left(1 - \frac{1}{c}\right)I_r.$$

Thus to compute the variance, we first compute the variance of $(A_{trn}A_{trn}^\dagger)_{ij}$. For this, we first note that

$$\begin{bmatrix}\frac{1}{c}I_N & 0\\ 0 & 0\end{bmatrix} = \mathbb{E}\left[\tilde{U}\tilde{\Sigma}\tilde{\Sigma}^\dagger\tilde{U}^T\right] = \mathbb{E}\left[A_{trn}A_{trn}^\dagger\right] = \mathbb{E}\left[A_{trn}A_{trn}^\dagger A_{trn}A_{trn}^\dagger\right].$$

The first equality follows from the symmetry of the signs of the rows of $\tilde{U}$. Then we can see that

$$\sum_k^d (A_{trn}A_{trn}^\dagger)_{ik}^2 = \begin{cases}\frac{1}{c} & i \le N\\ 0 & i > N\end{cases}.$$

From Lemma 14 in [29], we have that $\mathbb{E}[(A_{trn}A_{trn}^\dagger)_{ii}^2] = \frac{1}{c^2} + \frac{2}{cd} + o(1)$. Then combining this with the computation above and using symmetry, we have that for $i \neq j$ and $\min(i,j) \le N$

$$\mathbb{E}[(A_{trn}A_{trn}^\dagger)_{ij}^2] = \frac{1}{N-1}\left(\frac{1}{c} - \frac{1}{c^2} + \frac{2}{cd} + o(1)\right).$$

Now consider the other (full) SVD of X_{trn} given by $\hat{U}_{d\times d}\hat{\Sigma}_{d\times N}\hat{V}_{N\times N}^T$. Note that the top left $r \times r$ block of $\hat{\Sigma}$ is Σ_{trn}, and we can choose $\hat{U}$ so that the first r columns of $\hat{U}$ give U. Note that since $\hat{U}^T\tilde{U}$ is still uniformly random, the symmetry argument above follows for $\hat{U}^T A_{trn}A_{trn}^\dagger\hat{U}$. Additionally, for $i, j \le r$, $(\hat{U}^T A_{trn}A_{trn}^\dagger\hat{U})_{ij} = (U^T A_{trn}A_{trn}^\dagger U)_{ij}$ Thus, we see that for $i, j \le r$

$$\mathbb{E}\left[(U^T A_{trn}A_{trn}^\dagger U)_{ij}^2\right] = \frac{1}{N-1}\left(\frac{1}{c} - \frac{1}{c^2} + \frac{2}{cd} + o(1)\right),$$

while for $i = j$, we get that it is $O\left(\frac{1}{N}\right)$ by Lemma 14 of Sonthalia and Nadakuditi [29]. Thus, finally, we have that arranged as a matrix

$$\mathbb{E}\left[(\Sigma_{trn}^{-1}P^TP\Sigma_{trn}^{-1}) \odot (\Sigma_{trn}^{-1}P^TP\Sigma_{trn}^{-1})\right] = O\left(\frac{1}{d}\right).$$

By an analogous symmetry argument, since $(A_{trn}A_{trn}^\dagger)^i = A_{trn}A_{trn}^\dagger$ for any i, we can show that

$$\mathrm{Var}\left((U^T A_{trn}A_{trn}^\dagger U)_{ij}^2\right) = O\left(\frac{1}{d}\right).$$

We can in principle show a faster decay for this with a more involved argument, but this is enough for our purposes. We can now apply Lemma 4 with $X_N = I - (U^T A_{trn}A_{trn}^\dagger U)$ to see that

$$\mathbb{E}[\Sigma_{trn}(P^TP)^{-1}\Sigma_{trn}] = \frac{c}{c-1}I_r + O\left(\frac{1}{d}\right)$$

and has element-wise variance $O(1/d)$. $\square$

Lemma 6. *We have that*

$$\mathbb{E}[Z] = I \text{ and } \mathrm{Var}(Z_{ij}) = O\left(\frac{\|\Sigma_{trn}\|^2}{\eta_{trn}^2 d}\right).$$

Further, $E[Z\Sigma_{trn}^{-1}] = E[\Sigma_{trn}^{-1}Z] = \Sigma_{trn}^{-1}$ *and each element has variance* $O\left(\frac{1}{d}\right)$. *Finally,*

$$\mathbb{E}[Z^{-1}] = I + O\left(\frac{\|\Sigma_{trn}\|^2}{d}\right) \text{ with } \mathrm{Var}((Z^{-1})_{ij}) = O\left(\frac{\|\Sigma_{trn}\|^2}{d} + \frac{\|\Sigma_{trn}\|^4}{d^2}\right).$$

Proof. The element-wise variance and expectation of Z can be computed exactly as in the proof of Lemma 11 in Sonthalia and Nadakuditi [29]. Specifically, by considering the row u_j of U and the row v_i of V, treating Z_{ij} as β, and replacing θ_{trn} by σ_j. The expressions for the element-wise expectation and variance of $Z\Sigma_{trn}^{-1}$ and $\Sigma_{trn}^{-1}Z$ immediately follow from those of Z and the fact that $\sigma_i/\sigma_j = \Theta(1)$ by Assumption 1.

For Z^{-1}, we continue the computation using $Z_{ij} = 1 + T_{ij}$ with

$$T_{ij} = \sigma_j \sum_{k=1}^{\min(d,N)} \frac{1}{\lambda_k} a_k b_k$$

with a and b obtained using v_j and u_i respectively, and λ_k a singular value of A_{trn}. It is easy to check that

$$\mathrm{Var}(T_{ij}^2) = O\left(\frac{\|\Sigma_{trn}\|^4}{N^2}\right)$$

using a symmetry argument for a_k and b_k and the fact that $\mathbb{E}[1/\lambda_k^4] = O(1)$ by Lemma 5 of [29]. Now we can use Lemma 4 to conclude that

$$\mathbb{E}[Z^{-1}] = I + O\left(\frac{\|\Sigma_{trn}\|^2}{d}\right) \text{ with } \mathrm{Var}((Z^{-1})_{ij}) = O\left(\frac{\|\Sigma_{trn}\|^2}{d} + \frac{\|\Sigma_{trn}\|^4}{d^2}\right).$$

□

Lemma 7. *For* $c > 1$, $\mathbb{E}[K_1] = \frac{1}{\eta_{trn}^2}\frac{c}{c-1}I_r + \frac{c}{c-1}\Sigma_{trn}^{-2} + o(1)$ *with element-wise variance* $O(1/d)$. *Further,*

$$\mathbb{E}[K_1^{-1}] = \eta_{trn}^2\left(1 - \frac{1}{c}\right)\left(\eta_{trn}^2\Sigma_{trn}^{-2} + I_r\right)^{-1} + o(1)$$

with element-wise variance $O(1/d)$.

Proof. From Lemma 5, we have that

$$\mathbb{E}[\Sigma_{trn}(P^TP)^{-1}\Sigma_{trn}] = \frac{c}{c-1}I_r + O\left(\frac{1}{d}\right).$$

Recall that

$$K_1 = HH^T + Z(P^TP)^{-1}Z^T = HH^T + Z\Sigma_{trn}^{-1}(\Sigma_{trn}(P^TP)^{-1}\Sigma_{trn})\Sigma_{trn}^{-1}Z^T.$$

Then recall from Lemma 3 that

$$\mathbb{E}[HH^T] = \frac{1}{\eta_{trn}^2}\frac{c}{c-1}I_r + o(1).$$

For the second term in the expression for K_1, we want to use Lemmas 5 and 6, but they give expectations of each term separately. Note that

$$|\mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]| = |Cov(X,Y)| \leq \sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)}$$

and also note the following fact, from [43].

$$\begin{aligned} Cov(XY, WZ) = {} & \mathbb{E}X\mathbb{E}W Cov(Y,Z) + \mathbb{E}Y\mathbb{E}Z Cov(X,W) + \mathbb{E}X\mathbb{E}Z Cov(Y,W) + \\ & \mathbb{E}Y\mathbb{E}W Cov(X,Z) + Cov(X,W)Cov(Y,Z) + Cov(Y,W)Cov(X,Z) \end{aligned}$$

We use the facts above along with Lemmas 5 and 6 to compute the expectation. Specifically, the second term in K_1 is the product of three terms $Z\Sigma_{trn}^{-1}$, $(\Sigma_{trn}(P^TP)^{-1}\Sigma_{trn})$, and $\Sigma_{trn}^{-1}Z^T$. Hence we need the first fact to replace the expectation of the product of two terms with the product of the expectation of the two terms. To use this again, we would need to bound the variance of the product. Hence we need the second fact. Doing this computation, we get that

$$\mathbb{E}[K_1] = \frac{1}{\eta_{trn}^2}\frac{c}{c-1}I_r + \frac{c}{c-1}\Sigma_{trn}^{-2} + O\left(\frac{1}{d}\right) + o(1)$$

For the element-wise variance, consider $\delta K_1 = K_1 - \mathbb{E}[K_1]$. We cover the $i \neq j$ case. The $i = j$ case is analogous. From the proofs of Lemmas 3, 5, and 6, we have $Z_{ij} = I + T_{ij}$ and $(\Sigma_{trn}(P^TP)^{-1}\Sigma_{trn})_{ij} = U^TA_{trn}A_{trn}^\dagger U)_{ij}$. The expanding the product, we get that

$$\begin{aligned}(\delta K_1)_{ij} = &\left(v_i(A_{trn}^TA_{trn})^\dagger v_j\right) + O\left((U^TA_{trn}A_{trn}^\dagger U)_{ij}\right) + O\left((U^TA_{trn}A_{trn}^\dagger U)_{ij}^2\right) + O(T_{ij}) \\ &+ O\left(\sum_{k=1}^N T_{ik}(U^TA_{trn}A_{trn}^\dagger U)_{kj}\right) + O\left(\sum_{k=1}^N T_{ik}(U^TA_{trn}A_{trn}^\dagger U)_{kj}^2\right) + O\left(\sum_{k=1}^N T_{ik}T_{kj}\right) \\ &+ O\left(\sum_{k,l=1}^d T_{ik}(U^TA_{trn}A_{trn}^\dagger U)_{kl}T_{lj}\right) + O\left(\sum_{k,l=1}^d T_{ik}(U^TA_{trn}A_{trn}^\dagger U)_{kl}^2T_{lj}\right)\end{aligned}$$

Then since

$$\mathrm{Var}(XY) = Cov(X^2, Y^2) + (\mathrm{Var}(X) + (\mathbb{E}X)^2)(\mathrm{Var}(Y) + (\mathbb{E}Y)^2) - (Cov(X,Y) + \mathbb{E}X\mathbb{E}Y)^2$$

using this for terms five through nine, we get that

$$\mathrm{Var}\left((\delta K_1)_{ij}\right) = O\left(\frac{1}{d}\right).$$

For the inverse, we cover the $i \neq j$ case again. The $i = j$ case is analogous. We can perform an analogous computation to the one in the proof of Lemma 3 to get that

$$\mathrm{Var}\left((v_i(A_{trn}^TA_{trn})^\dagger v_j)^2\right) = O\left(\frac{1}{N}\right),$$

using the fact that $\mathbb{E}\left[\frac{1}{\lambda^4}\right] = O(1)$ for a random eigenvalue λ_k of A_{trn}. We also use the fact that $(A_{trn}A_{trn}^\dagger)^p = A_{trn}A_{trn}^\dagger$ for any p and a symmetry argument analogous to the one in the proof of Lemma 5 to note that

$$\mathbb{E}\left[(U^TA_{trn}A_{trn}^\dagger U)_{ij}^p\right] = O\left(\frac{1}{d}\right) \qquad p = 2, \ldots, 8.$$

One can also check by the arguments in the proof of Lemma 6 that

$$\mathbb{E}\left[T_{ij}^{2p}\right] = O\left(\frac{\sigma_i^p\sigma_j^p}{d^p}\right) = O(1).$$

These together with the facts about $\mathrm{Var}(XY)$ and $Cov(XY, ZW)$ above establish after a tedious but straightforward computation that

$$\mathrm{Var}((\delta K_1)_{ij}^2) = O\left(\frac{1}{d}\right).$$

We can now use Lemma 4 to establish that

$$\begin{aligned}\mathbb{E}[K_1^{-1}] &= \eta_{trn}^2\left(1 - \frac{1}{c}\right)\left(\eta_{trn}^2\Sigma_{trn}^{-2} + I_r\right)^{-1} + O\left(\frac{1}{d}\right) + o(1) \\ &= \eta_{trn}^2\left(1 - \frac{1}{c}\right)\left(\eta_{trn}^2\Sigma_{trn}^{-2} + I_r\right)^{-1} + o(1)\end{aligned}$$

and

$$\mathrm{Var}((K_1^{-1})_{ij}) = O\left(\frac{1}{d}\right).$$

□

Lemma 8. *When $c > 1$, we have for $W = W_{opt}$ that*

$$\mathbb{E}[\|W\|_F^2] = \frac{c}{c-1}\operatorname{Tr}(\Sigma_{trn}^2(\Sigma_{trn}^2 + \eta_{trn}^2 I)^{-1}) + O\left(\frac{\|\Sigma_{trn}\|^2}{d}\right) + o(1).$$

Proof. We first use the estimates for the expectations from Lemmas 3, 5, 6, and 7 to get an estimate for the expectation of $\|W\|_F^2$. We get this estimate by treating various matrices in the product as independent. We then bound the deviation of the true expectation from this estimate using the variance estimates above. We begin the calculation as

$$\|W\|_F^2 = \operatorname{Tr}(W^T W)$$

Using Lemma 1, we see that the trace has three terms. The first term is

$$\operatorname{Tr}\left(H^T(K_1^{-1})^T Z((P^TP)^{-1})^T\Sigma_{trn}^T U^T U \Sigma_{trn}(P^TP)^{-1}Z^T K_1^{-1} H\right).$$

Here we have that U is $d \times r$ with orthonormal columns. Hence we get that $U^TU = I$. Then since the trace is invariant under cyclic permutations, we get the following term

$$\operatorname{Tr}\left((\Sigma_{trn}(P^TP)^{-1}\Sigma_{trn})(\Sigma_{trn}^{-1}Z^T)K_1^{-1}HH^T(K_1^{-1})^T(Z\Sigma_{trn}^{-1})(\Sigma_{trn}(P^TP)^{-1}\Sigma_{trn})^T\right).$$

Now we use our random matrix theory estimates for various terms in the product. From Lemma 6, we have that $\mathbb{E}_{A_{trn}}[Z\Sigma_{trn}^{-1}] = \Sigma_{trn}^{-1}$. Thus, that first term's expectation can be estimated by

$$\operatorname{Tr}\left((\Sigma_{trn}(P^TP)^{-1}\Sigma_{trn})\Sigma_{trn}^{-1}K_1^{-1}HH^T(K_1^{-1})^T\Sigma_{trn}^{-1}(\Sigma_{trn}(P^TP)^{-1}\Sigma_{trn})^T\right).$$

Then using Lemma 3, we can further estimate this by

$$\frac{1}{\eta_{trn}^2}\frac{c}{c-1}\operatorname{Tr}\left((\Sigma_{trn}(P^TP)^{-1}\Sigma_{trn})\Sigma_{trn}^{-1}K_1^{-1}(K_1^{-1})^T\Sigma_{trn}^{-1}(\Sigma_{trn}(P^TP)^{-1}\Sigma_{trn})^T\right) + o(1).$$

Here, the error contribution of the $o(1)$ error from Lemma 3 is still $o(1)$ since we will see that all the other estimates are $O(1)$. Then we use Lemma 5, to replace $\Sigma_{trn}(P^TP)^{-1}\Sigma_{trn}$ to get

$$\frac{1}{\eta_{trn}^2}\frac{c}{c-1}\left(1-\frac{1}{c}\right)^{-2}\operatorname{Tr}\left(\Sigma_{trn}^{-1}K_1^{-1}(K_1^{-1})^T(\Sigma_{trn}^T)^{-1}\right) + o(1).$$

Finally, we use Lemma 7 to replace the last term and get

$$\frac{1}{\eta_{trn}^2}\frac{c}{c-1}\left(\frac{c}{c-1}\right)^2\operatorname{Tr}\left(\Sigma_{trn}^{-2}\eta_{trn}^4\left(1-\frac{1}{c}\right)^2\left(I_r + \eta_{trn}^2\Sigma_{trn}^{-2}\right)^{-2}\right) + o(1).$$

This immediately simplifies to

$$\eta_{trn}^2\frac{c}{c-1}\operatorname{Tr}\left(\Sigma_{trn}^2(\Sigma_{trn}^2 + \eta_{trn}^2 I_r)^{-2}\right) + o(1). \tag{3}$$

The second term in $\operatorname{Tr}(W^TW)$ is

$$-2\operatorname{Tr}\left(H^T(K_1^{-1})^T Z^T((P^TP)^{-1})^T\Sigma_{trn}^T U^T U\Sigma_{trn}Z^{-1}HH^TZP^\dagger\right).$$

We can rearrange this using cyclic invariance to

$$-2\operatorname{Tr}\left((K_1^{-1})^T Z^T\Sigma_{trn}^{-1}(\Sigma_{trn}(P^TP)^{-1}\Sigma_{trn})^T\Sigma_{trn}Z^{-1}HH^TZP^\dagger H^T\right).$$

Let us focus on the $P^\dagger H^T$ term. Since P^TP is invertible, we have that P has full column rank. Hence we have that

$$P^\dagger = (P^TP)^{-1}P^T.$$

Further, since $P = -(I - A_{trn}A_{trn}^\dagger)U\Sigma_{trn}$ and $H = V_{trn}^T A_{trn}^\dagger$, we have that

$$P^\dagger H^T = (P^TP)^{-1}\Sigma_{trn}^T U^T(I - A_{trn}A_{trn}^\dagger)(A_{trn}^\dagger)^T V_{trn}.$$

Finally, we notice that

$$A_{trn}A_{trn}^\dagger(A_{trn}^\dagger)^T = (A_{trn}^\dagger)^T.$$

Thus, we have that

$$P^\dagger H^T = (P^TP)^{-1}\Sigma_{trn}^T U^T (I - A_{trn}A_{trn}^\dagger)(A_{trn}^\dagger)^T V_{trn} = 0. \tag{4}$$

Finally, the last term in $\mathrm{Tr}(W^TW)$ is

$$\mathrm{Tr}\left((P^\dagger)^T Z^T (K_1^{-1})^T HH^T (Z^{-1})^T \Sigma_{trn}^T U^T U \Sigma_{trn} Z^{-1} HH^T K_1^{-1} Z P^\dagger\right).$$

We note that

$$P^\dagger (P^\dagger)^T = (P^TP)^\dagger = (P^TP)^{-1}.$$

We use this observation along with cyclic invariance to get that the last term is the same as

$$\mathrm{Tr}\left((K_1^{-1})^T HH^T \Sigma_{trn}^2 Z^{-1} HH^T K_1^{-1} Z \Sigma_{trn}^{-1} (\Sigma_{trn}(P^TP)^{-1}\Sigma_{trn})\Sigma_{trn}^{-1} Z^T\right).$$

We again use Lemmas 3 and 6 to get that its expectation is estimated by

$$\frac{1}{\eta_{trn}^4}\left(\frac{c}{c-1}\right)^2 \mathrm{Tr}\left((K_1^{-1})^T \Sigma_{trn}^2 K_1^{-1} \Sigma_{trn}^{-1}(\Sigma_{trn}(P^TP)^{-1}\Sigma_{trn})\Sigma_{trn}^{-1}\right) + O\left(\frac{\|\Sigma_{trn}\|^2}{d}\right) + o(1).$$

The contribution of the $O\left(\frac{\|\Sigma_{trn}\|^2}{d}\right)$ error from Lemma 6 is still $O\left(\frac{\|\Sigma_{trn}\|^2}{d}\right)$ since the estimate for the expectation is $O(1)$. We now use Lemma 5, and 7 to see that the final term's expectation can be estimated by

$$\begin{aligned}
&\frac{1}{\eta_{trn}^4}\left(\frac{c}{c-1}\right)^3 \eta_{trn}^4 \left(\frac{c-1}{c}\right)^{-2} (I_r + \eta_{trn}\Sigma_{trn}^{-2})^{-2} + O\left(\frac{\|\Sigma_{trn}\|^2}{d}\right) + o(1) \\
=&\frac{c}{c-1}\mathrm{Tr}(\Sigma_{trn}^4(\Sigma_{trn}^2 + \eta_{trn}^2 I_r)^{-2}) + O\left(\frac{\|\Sigma_{trn}\|^2}{d}\right) + o(1).
\end{aligned} \tag{5}$$

Finally, to bound the deviation from this estimate, note that for real valued random variables X, Y we have that $|\mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]| = |Cov(X,Y)| \le \sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)}$ and for real valued random variables X, Y, Z, W, we have the following fact, from [43].

$$\begin{aligned}
Cov(XY, WZ) = \mathbb{E}X\mathbb{E}W Cov(Y,Z) + \mathbb{E}Y\mathbb{E}Z Cov(X,W) + \mathbb{E}X\mathbb{E}Z Cov(Y,W) + \\
\mathbb{E}Y\mathbb{E}W Cov(X,Z) + Cov(X,W)Cov(Y,Z) + Cov(Y,W)Cov(X,Z)
\end{aligned}$$

We repeatedly apply these two to upper bound the deviation between the product of the expectations in the estimates above and the expectation of the product. It is then straightforward to see that since all variances are $O(1/d)$ except for those of Z^{-1} and Z, which are both $O(1)$ whenever $\Sigma_{trn} = O(\sqrt{d})$, the estimation error is $O(1/\sqrt{d}) = o(1)$.

So, we can conclude that each of the estimates in equations 3, 4 and 5 have error $o(1)$. Combining the terms together, we get from equations 3, 4 and 5 that

$$\begin{aligned}
\|W\|_F^2 &= \frac{c}{c-1}\mathrm{Tr}\left(\Sigma_{trn}^2(\Sigma_{trn}^2 + \eta_{trn}^2 I_r)(\Sigma_{trn} + \eta_{trn}^2 I_r)^{-2}\right) + O\left(\frac{\|\Sigma_{trn}\|^2}{d}\right) + o(1) \\
&= \frac{c}{c-1}\mathrm{Tr}(\Sigma_{trn}^2(\Sigma_{trn}^2 + \eta_{trn}^2 I)^{-1}) + O\left(\frac{\|\Sigma_{trn}\|^2}{d}\right) + o(1).
\end{aligned}$$

□

Theorem 8. *When $d > N + r$ and $\beta = I$, then the test error $\mathcal{R}(W, X_{tst})$ for $W = W_{opt}$ is given by*

$$\frac{\eta_{trn}^4}{N_{tst}}\|\left(\Sigma_{trn}^2 + \eta_{trn}^2 I\right)^{-1} L\|_F^2 + \frac{\eta_{tst}^2}{d}\frac{c}{c-1}\mathrm{Tr}(\Sigma_{trn}^2(\Sigma_{trn}^2 + \eta_{trn}^2 I)^{-1}) + O\left(\frac{\|\Sigma_{trn}\|^2}{d^2}\right) + o\left(\frac{1}{d}\right).$$

Proof. Recall from theorem 7 that

$$\mathcal{R}(W, X_{tst}) = \mathbb{E}\left[\frac{1}{N_{tst}}\|U\Sigma_{trn}(P^TP)^{-1}Z^TK_1^{-1}\Sigma_{trn}^{-1}L\|_F^2 + \frac{\eta_{tst}^2}{d}\|W\|_F^2\right]$$

To compute the expectation of the first term, we observe that it is given by

$$\frac{1}{N_{tst}} Tr(U\Sigma_{trn}(P^TP)^{-1}Z^TK_1^{-1}\Sigma_{trn}^{-1}LL^T\Sigma_{trn}^{-1}K_1^{-1}Z(P^TP)^{-1}\Sigma_{trn}U^T).$$

We apply cyclic invariance to get that it is the same as

$$\frac{1}{N_{tst}} Tr(\Sigma_{trn}^{-1}K_1^{-1}Z\Sigma_{trn}^{-1}(\Sigma_{trn}(P^TP)^{-1}\Sigma_{trn})(\Sigma_{trn}(P^TP)^{-1}\Sigma_{trn})\Sigma_{trn}^{-1}Z^TK_1^{-1}\Sigma_{trn}^{-1}LL^T).$$

We finally use Lemmas 5, 6, and 7 to estimate it by

$$\begin{aligned}
&\frac{1}{N_{tst}} Tr\left(\Sigma_{trn}^{-2}\left(\frac{c}{c-1}\right)^2\left(\frac{c-1}{c}\right)^2\left(\Sigma_{trn}^{-2}+\frac{1}{\eta_{trn}^2}I\right)^{-2}\Sigma_{trn}^{-2}LL^T\right)+o\left(\frac{1}{d}\right)\\
&=\frac{\eta_{trn}^4}{N_{tst}} Tr\left((\Sigma_{trn}^2+\eta_{trn}^2I)^{-2}LL^T\right)+o\left(\frac{1}{d}\right)\\
&=\frac{\eta_{trn}^4}{N_{tst}}\|\left(\Sigma_{trn}^2+\eta_{trn}^2I\right)^{-1}L\|_F^2+o\left(\frac{1}{d}\right)
\end{aligned}$$

Since test and train data are decoupled, we can treat LL^T/N_{tst} as a constant as N grows, noting that due the Σ_{trn}^{-2}, the final estimate is $o(1)$. So, repeating the deviation argument at the end of the proof of Lemma 8 above, we then have that the deviation from this estimate is $o\left(\frac{1}{d}\right)$.

Combining this with Lemma 8, we get that

$$\frac{\eta_{trn}^4}{N_{tst}}\|\left(\Sigma_{trn}^2+\eta_{trn}^2I\right)^{-1}L\|_F^2+\frac{\eta_{tst}^2}{d}\frac{c}{c-1}\operatorname{Tr}(\Sigma_{trn}^2(\Sigma_{trn}^2+\eta_{trn}^2I)^{-1})+O\left(\frac{\|\Sigma_{trn}^2\|}{d^2}\right)+o\left(\frac{1}{d}\right).$$

$\square$

F.1.2 The Underparametrized Regime, $d < N$

We derive test error bounds for $\beta = I$ in our problem setting. We also denote W_{opt} by W in this subsection, for ease of notation.

Theorem 9. *For rank r data and $d < N - r$, with $c = \frac{d}{N}$, the following is true.*

1. *For the $\beta = I$ case, we denote the minimum norm linear denoiser W_{opt} by just W in this subsection. It is given by*

$$W = -U\Sigma_{trn}H_1^{-1}K^TA_{trn}^{\dagger} + U\Sigma_{trn}H_1^{-1}Z^T(QQ^T)^{-1}H$$

2. *The test error when $X_{tst} = UL$ is given by*

$$\mathbb{E}_{A_{trn}}\left[\frac{1}{N_{tst}}\|U\Sigma_{trn}H_1^{-1}Z^T(QQ^T)^{-1}\Sigma_{trn}^{-1}L\|_F^2+\frac{\eta_{tst}^2}{d}\|W\|_F^2\right],$$

where $Q = V^T(I - A_{trn}^{\dagger}A_{trn})$, $H = V_{trn}^TA_{trn}^{\dagger}$, $K = -A_{trn}^{\dagger}U\Sigma_{trn}$, $Z = I + V_{trn}^TA_{trn}^{\dagger}U\Sigma_{trn}$, $H_1 = K^TK + Z^T(QQ^T)^{-1}Z$.

The sizes of the matrices:

1. U is $d \times r$ with $U^TU = I_{r\times r}$.
2. Σ_{trn} is $r \times r$, with rank r.
3. A_{trn} is $d \times N$ with rank d.
4. $A_{trn}^{\dagger}A_{trn}$ is $N \times N$
5. H is $r \times d$, with rank r.
6. K is $N \times r$, with rank r.
7. Z is $r \times r$, with rank r.

8. H_1 is $r \times r$, with rank r.
9. $A_{trn} = \eta_{trn}\tilde{U}\tilde{\Sigma}\tilde{V}^T$.
10. $\tilde{U}$ is $d \times d$ unitary.
11. $\tilde{\Sigma}$ is $d \times N$.

Proof. Part 1 follows from Lemma 1. For part 2, note that the test error is given by $\mathcal{R}(W, X_{tst}) = \mathbb{E}_{A_{trn},A_{tst}}\left[\frac{1}{N_{tst}}\|X_{tst} - W(X_{tst} + A_{tst})\|_F^2\right]$, which is the same as the following.

$$\begin{aligned}
\mathcal{R}(W, X_{tst}) &= \frac{1}{N_{tst}}\mathbb{E}_{A_{trn},A_{tst}}\left[\|X_{tst} - WX_{tst}\|_F^2\right] + \frac{2}{N_{tst}}\mathbb{E}_{A_{trn},A_{tst}}[Tr((X_{tst} - WX_{tst})A_{tst}) \\
&\quad + \frac{1}{N_{tst}}\mathbb{E}_{A_{trn},A_{tst}}\left[\|WA_{tst}\|_F^2\right] \\
&= \frac{1}{N_{tst}}\mathbb{E}_{A_{trn}}\left[\|X_{tst} - WX_{tst}\|_F^2\right] + 0 + \frac{1}{N_{tst}}\mathbb{E}_{A_{trn},A_{tst}}\left[Tr(W^TWA_{tst}A_{tst}^T)\right] \\
&= \frac{1}{N_{tst}}\mathbb{E}_{A_{trn}}\left[\|X_{tst} - WX_{tst}\|_F^2\right] + 0 + \frac{1}{N_{tst}}\mathbb{E}_{A_{trn}}\left[Tr(W^TW\mathbb{E}_{A_{tst}}\left[A_{tst}A_{tst}^T\right])\right] \\
&= \frac{1}{N_{tst}}\mathbb{E}_{A_{trn}}\left[\|X_{tst} - WX_{tst}\|_F^2\right] + 0 + \frac{\eta_{tst}^2 N_{tst}}{dN_{tst}}\mathbb{E}_{A_{trn}}[Tr(W^TW)] \\
&= \mathbb{E}_{A_{trn}}\left[\frac{1}{N_{tst}}\|U\Sigma_{trn}H_1^{-1}Z^T(QQ^T)^{-1}\Sigma_{trn}^{-1}L\|_F^2 + \frac{\eta_{tst}^2}{d}\|W\|_F^2\right]
\end{aligned}$$

□

We will henceforth drop the subscript A_{trn} in the expectation $\mathbb{E}_{A_{trn}}$.

Lemma 9. *When $d < N - r$, for $Q = V^T(I - A_{trn}^\dagger A_{trn})$, $K = -A_{trn}^\dagger \Sigma_{trn} U$, $H_1 = K^TK + Z^T(QQ^T)^{-1}Z$ and other notation as in previous lemmas, we have that*

$$W = -U\Sigma_{trn}H_1^{-1}K^TA_{trn}^\dagger + U\Sigma_{trn}H_1^{-1}Z^T(QQ^T)^{-1}H.$$

Proof. We know that $W = X(X + A_{trn})^\dagger$. By Corollary 2.3 of Wei [42], setting $X = -CB$ with $C = -U\Sigma_{trn}$ and $B = V^T$, we have that

$$(X + A_{trn})^\dagger = A_{trn}^\dagger - Q^\dagger H - (K + Q^\dagger Z)H_1^{-1}(K^TA_{trn}^\dagger - Z^T(QQ^T)^{-1}H).$$

So, using the facts that $X = U\Sigma_{trn}V^T$, $K = -A_{trn}^\dagger U\Sigma_{trn}$, we have that

$$\begin{aligned}
W &= X(X + A_{trn}^\dagger) \\
&= U\Sigma_{trn}V^TA_{trn}^\dagger - U\Sigma_{trn}Q^\dagger H + U\Sigma_{trn}V^TA_{trn}^\dagger U\Sigma_{trn}H_1^{-1}K^TA_{trn}^\dagger \\
&\quad - U\Sigma_{trn}V^TQ^\dagger ZH_1^{-1}K^TA_{trn}^\dagger - U\Sigma_{trn}V^TA_{trn}^\dagger U\Sigma_{trn}H_1^{-1}Z^T(QQ^T)^{-1}H \\
&\quad + U\Sigma_{trn}V^TQ^\dagger ZH_1^{-1}Z^T(QQ^T)^{-1}H.
\end{aligned}$$

Using the fact that $H = V^TA_{trn}^\dagger$, we get that

$$\begin{aligned}
W &= U\Sigma_{trn}H - U\Sigma_{trn}Q^\dagger H + U\Sigma_{trn}HU\Sigma_{trn}H_1^{-1}K^TA_{trn}^\dagger - U\Sigma_{trn}V^TQ^\dagger ZH_1^{-1}K^TA_{trn}^\dagger \\
&\quad - U\Sigma_{trn}HU\Sigma_{trn}H_1^{-1}Z^T(QQ^T)^{-1}ZZ^{-1}H + U\Sigma_{trn}V^TQ^\dagger ZH_1^{-1}Z^T(QQ^T)^{-1}ZZ^{-1}H.
\end{aligned}$$

Using the fact that $Z = I + V^TA_{trn}^\dagger U\Sigma_{trn} = I + HU\Sigma_{trn}$, we get that

$$\begin{aligned}
W &= U\Sigma_{trn}H - U\Sigma_{trn}Q^\dagger H + U\Sigma_{trn}(Z - I)H_1^{-1}K^TA_{trn}^\dagger - U\Sigma_{trn}V^TQ^\dagger ZH_1^{-1}K^TA_{trn}^\dagger \\
&\quad - U\Sigma_{trn}(Z - I)H_1^{-1}Z^T(QQ^T)^{-1}ZZ^{-1}H + U\Sigma_{trn}V^TQ^\dagger ZH_1^{-1}Z^T(QQ^T)^{-1}ZZ^{-1}H.
\end{aligned}$$

Using the fact that $H_1 = K^TK + Z^T(QQ^T)^{-1}Z$, we get that

$$\begin{aligned}
W &= U\Sigma_{trn}H - U\Sigma_{trn}Q^\dagger H + U\Sigma_{trn}ZH_1^{-1}K^TA_{trn}^\dagger - U\Sigma_{trn}H_1^{-1}K^TA_{trn}^\dagger \\
&\qquad - U\Sigma_{trn}V^TQ^\dagger ZH_1^{-1}K^TA_{trn}^\dagger - U\Sigma_{trn}ZH_1^{-1}(H_1 - K^TK)Z^{-1}H \\
&\qquad + U\Sigma_{trn}H_1^{-1}Z^T(QQ^T)^{-1}H + U\Sigma_{trn}V^TQ^\dagger ZH_1^{-1}(H_1 - K^TK)Z^{-1}H \\
&= U\Sigma_{trn}H - U\Sigma_{trn}Q^\dagger H + U\Sigma_{trn}ZH_1^{-1}K^TA_{trn}^\dagger - U\Sigma_{trn}H_1^{-1}K^TA_{trn}^\dagger \\
&\qquad - U\Sigma_{trn}V^TQ^\dagger ZH_1^{-1}K^TA_{trn}^\dagger - U\Sigma_{trn}H + U\Sigma_{trn}ZH_1^{-1}K^TKZ^{-1}H \\
&\qquad + U\Sigma_{trn}H_1^{-1}Z^T(QQ^T)^{-1}H + U\Sigma_{trn}V^TQ^\dagger H - U\Sigma_{trn}V^TQ^\dagger ZH_1^{-1}K^TKZ^{-1}H.
\end{aligned}$$

Cancelling terms, we get that

$$\begin{aligned}
W &= U\Sigma_{trn}ZH_1^{-1}K^TA_{trn}^\dagger - U\Sigma_{trn}H_1^{-1}K^TA_{trn}^\dagger - U\Sigma_{trn}V^TQ^\dagger ZH_1^{-1}K^TA_{trn}^\dagger \\
&\qquad + U\Sigma_{trn}ZH_1^{-1}K^TKZ^{-1}H + U\Sigma_{trn}H_1^{-1}Z^T(QQ^T)^{-1}H \\
&\qquad - U\Sigma_{trn}V^TQ^\dagger ZH_1^{-1}K^TKZ^{-1}H.
\end{aligned}$$

And we rearrange to get that

$$\begin{aligned}
W &= -U\Sigma_{trn}H_1^{-1}K^TA_{trn}^\dagger + U\Sigma_{trn}H_1^{-1}Z^T(QQ^T)^{-1}H + U\Sigma_{trn}(I - V^TQ^\dagger)ZH_1^{-1}K^TA_{trn}^\dagger \\
&\qquad + U\Sigma_{trn}(I - V^TQ^\dagger)ZH_1^{-1}K^TKZ^{-1}H \\
&= -U\Sigma_{trn}H_1^{-1}K^TA_{trn}^\dagger + U\Sigma_{trn}H_1^{-1}Z^T(QQ^T)^{-1}H,
\end{aligned}$$

where the last equality is because $Q = V^T(I - A_{trn}^\dagger A_{trn})$ has full rank, so $Q^\dagger = Q^T(QQ^T)^{-1}$, so $V^TQ^\dagger = V^T(I - A_{trn}^\dagger A_{trn})V(V^T(I - A_{trn}^\dagger A_{trn})V)^{-1} = I$. □

Lemma 10. *For $d < N - r$, with notation as in Lemma 9 have that*

$$X_{tst} - WX_{tst} = U\Sigma_{trn}H_1^{-1}Z^T(QQ^T)^{-1}\Sigma_{trn}^{-1}L.$$

Proof. Note that

$$X_{tst} - WX_{tst} = UL - U\Sigma_{trn}H_1^{-1}K^TA_{trn}^\dagger UL - U\Sigma_{trn}H_1^{-1}Z^T(QQ^T)^{-1}HUL.$$

Remember that $K = -A_{trn}U\Sigma$, so $A_{trn}U\Sigma_{tst} = -K\Sigma_{trn}^{-1}\Sigma_{tst}$ and $HU\Sigma_{tst} = (HU\Sigma)\Sigma_{trn}^{-1}\Sigma_{tst} = (Z - I)\Sigma_{trn}^{-1}\Sigma_{tst}$ This gives us the following equality.

$$\begin{aligned}
X_{tst} - WX_{tst} &= UL - U\Sigma_{trn}H_1^{-1}K^TK\Sigma_{trn}^{-1}L - U\Sigma_{trn}H_1^{-1}Z^T(QQ^T)^{-1}Z\Sigma_{trn}^{-1}L \\
&\qquad + U\Sigma_{trn}H_1^{-1}Z^T(QQ^T)^{-1}\Sigma_{trn}^{-1}L \\
&= U(I - \Sigma_{trn}H_1^{-1}(K^TK + Z^T(QQ^T)^{-1}Z)\Sigma_{trn}^{-1} + \Sigma_{trn}H_1^{-1}Z^T(QQ^T)^{-1}\Sigma_{trn}^{-1})L.
\end{aligned}$$

Using the fact that $H_1 = K^TK + Z^T(QQ^T)^{-1}Z$, we get that

$$\begin{aligned}
X_{tst} - WX_{tst} &= UL - U\Sigma_{trn}H_1^{-1}H_1\Sigma_{trn}^{-1}L + U\Sigma_{trn}H_1^{-1}Z^T(QQ^T)^{-1}\Sigma_{trn}^{-1}L \\
&= U\Sigma_{trn}H_1^{-1}Z^T(QQ^T)^{-1}\Sigma_{trn}^{-1}L.
\end{aligned}$$

□

Lemma 11. *For $c < 1$, we have that*

$$\mathbb{E}[\Sigma_{trn}^{-1}K^TK\Sigma_{trn}^{-1}] = \frac{1}{\eta_{trn}^2}\frac{c}{1-c} + o(1)$$

and the variance of the ij^{th} entry is $O\left(\frac{1}{N}\right)$.

Proof. Note that $K^TK = \Sigma_{trn}U^T(A_{trn}A_{trn}^T)^\dagger U\Sigma_{trn}$. So, $(K^TK)_{ij} = \sigma_i u_i^T(A_{trn}A_{trn}^T)^\dagger u_j\sigma_j$. Using ideas from Sonthalia and Nadakuditi [29], we see that if $i \neq j$, then the expectation is 0. On the other hand if $i=j$, then using Lemma 6 from [29], with $p=N$, $q=d$, $A=\frac{1}{\eta_{trn}}A_{trn}^T$, we get that

$$\mathbb{E}[(\Sigma_{trn}^{-1}K^TK\Sigma_{trn}^{-1})_{ii}] = \frac{1}{\eta_{trn}^2}\frac{c}{1-c} + o(1).$$

The result on the expectation follows immediately from this.

For the variance, pick arbitrary $i \neq j$ and fix them. Consider $a = \tilde{U}^* u_i$ and $b = \tilde{U}^* u_j$. They are uniformly random orthogonal unit vectors, not necessarily independent. Now note that

$$\begin{aligned}
(\Sigma_{trn}^{-1}(K^TK)\Sigma_{trn}^{-1})_{ij} &= \sigma_i u_i^T (A_{trn}A_{trn}^T)^\dagger u_j \sigma_j \\
&= u_i^T (\tilde{U}\tilde{\Sigma}\tilde{\Sigma}^*\tilde{U}^*)^\dagger u_j \\
&= u_i^T \tilde{U}(\tilde{\Sigma}\tilde{\Sigma}^*)^\dagger \tilde{U}^* u_j \\
&= a^T(\tilde{\Sigma}\tilde{\Sigma}^*)^\dagger b \\
&= \sum_{k=1}^d \frac{1}{\tilde{\sigma}_k^2} a_k b_k.
\end{aligned}$$

So, we get that

$$\begin{aligned}
\mathbb{E}[((\Sigma_{trn}^{-1}(K^TK)\Sigma_{trn}^{-1})_{ij})^2] &= \mathbb{E}\left[\left(\sum_{k=1}^d \frac{1}{\tilde{\sigma}_k^2} a_k b_k\right)^2\right] \\
&= \mathbb{E}\left[\sum_{k=1}^d \sum_{l=1}^d \frac{1}{\tilde{\sigma}_k^2\tilde{\sigma}_l^2} a_k b_k a_l b_l\right] \\
&= \left(\frac{c^2}{(1-c)^2} + o(1)\right)\mathbb{E}\left[\left(\sum_{k=1}^d a_k b_k\right)^2\right] \\
&\quad + \left(\frac{c^2}{(1-c)^3} - \frac{c^2}{(1-c)^2} + o(1)\right)\mathbb{E}\left[\sum_{k=1}^d a_k^2 b_k^2\right] \\
&= \left(\frac{c^2}{(1-c)^3} - \frac{c^2}{(1-c)^2} + o(1)\right)\mathbb{E}\left[\sum_{k=1}^d a_k^2 b_k^2\right] \\
&= \left(\frac{c^3}{(1-c)^3} + o(1)\right)\mathbb{E}\left[\sum_{k=1}^d a_k^2 b_k^2\right] \\
&= \frac{c^3}{(1-c)^3}\sum_{k=1}^d \mathbb{E}[a_k^2]\mathbb{E}[b_k^2] + o\left(\frac{1}{d}\right),
\end{aligned}$$

where the last line holds due to the following reasoning, even though a and b are not independent. We then use the fact that

$$\mathbb{E}[a_k^2 b_k^2] - \mathbb{E}[a_k^2]\mathbb{E}[b_k^2] \leq \sqrt{\mathrm{Var}(a_k^2)\mathrm{Var}(b_k^2)}$$

and Lemma 13 of [29], to get that

$$\mathrm{Var}\left(\sum_{k=1}^d a_k^2\right) = O\left(\frac{1}{d}\right).$$

So, by symmetry of coordinates,

$$\mathrm{Var}(a_k^2) = O\left(\frac{1}{d^2}\right).$$

The same holds for b_k, giving us that

$$\left|\mathbb{E}[a_k^2 b_k^2] - \mathbb{E}[a_k^2]\mathbb{E}[b_k^2]\right| \leq O\left(\frac{1}{d^2}\right).$$

This gives us that

$$\mathrm{Var}\left((\Sigma_{trn}^{-1}(K^TK)\Sigma_{trn}^{-1})_{ij}^2\right) = \frac{c^3}{d(1-c)^3} + o\left(\frac{1}{d}\right) \qquad i \neq j.$$

For $i = j$, we use Sonthalia and Nadakuditi [29] to see that the variance is $O\left(\frac{1}{d}\right) = O\left(\frac{1}{N}\right)$ since $d = cN$. □

Lemma 12. *For $c < 1$, we have that*

$$\mathbb{E}[\Sigma_{trn}^{-1}K^TA_{trn}^{\dagger}(A_{trn}^{\dagger})^TK\Sigma_{trn}^{-1}] = \frac{1}{\eta_{trn}^2}\frac{c^2}{(1-c)^3} + o(1)$$

and the variance of the ij^{th} entry is $O\left(\frac{1}{N}\right)$.

Proof. Let $M := \Sigma_{trn}^{-1}K^TA_{trn}^{\dagger}(A_{trn}^{\dagger})^TK\Sigma_{trn}^{-1}$ and note that

$$\Sigma_{trn}^{-1}K^TA_{trn}^{\dagger}(A_{trn}^{\dagger})^TK\Sigma_{trn}^{-1} = \Sigma_{trn}U^T(A_{trn}A_{trn}^T)^{\dagger}(A_{trn}A_{trn}^T)^{\dagger}U\Sigma_{trn}.$$

So,

$$M_{ij} = \sigma_i u_i^T(A_{trn}A_{trn}^T)^{\dagger}(A_{trn}A_{trn}^T)^{\dagger}u_j\sigma_j.$$

Using ideas from [29], we see that if $i \neq j$, then the expectation is 0. On the other hand if $i = j$, then using Lemma 6 from [29], with $p = N$, $q = d$, we get that

$$\mathbb{E}[M_{ii}] = \frac{\sigma_i^2}{\eta_{trn}^2}\frac{c^2}{(1-c)^3} + o(1).$$

For the variance, pick arbitrary $i \neq j$ and fix them. Consider $a = \tilde{U}^*u_i$ and $b = \tilde{U}^*u_j$. They are uniformly random orthogonal unit vectors, not necessarily independent. Now note that

$$\begin{aligned}
M_{ij} &= u_i^T(A_{trn}A_{trn}^T)^{\dagger}(A_{trn}A_{trn}^T)^{\dagger}u_j \\
&= u_i^T(\tilde{U}\tilde{\Sigma}\tilde{\Sigma}^*\tilde{\Sigma}\tilde{\Sigma}^*\tilde{U}^*)^{\dagger}u_j \\
&= u_i^T\tilde{U}(\tilde{\Sigma}\tilde{\Sigma}^*\tilde{\Sigma}\tilde{\Sigma}^*)^{\dagger}\tilde{U}^*u_j \\
&= a^T(\tilde{\Sigma}\tilde{\Sigma}^*\tilde{\Sigma}\tilde{\Sigma}^*)^{\dagger}b \\
&= \sum_{k=1}^{d}\frac{1}{\tilde{\sigma}_k^4}a_kb_k.
\end{aligned}$$

So, we get that

$$\begin{aligned}
\mathbb{E}[M_{ij}^2] &= \mathbb{E}\left[\left(\sum_{k=1}^{d}\frac{1}{\tilde{\sigma}_k^4}a_kb_k\right)^2\right] \\
&= \mathbb{E}\left[\sum_{k=1}^{d}\sum_{l=1}^{d}\frac{1}{\sigma_k^4\sigma_l^4}a_kb_ka_lb_l\right] \\
&= \left(\frac{c^4(c^2+22/6c+1)}{(1-c)^7} + o(1)\right)\mathbb{E}\left[\left(\sum_{k=1}^{d}a_kb_k\right)^2\right] + (\chi(c)+o(1))\mathbb{E}\left[\sum_{k=1}^{d}a_k^2b_k^2\right] \\
&= (\chi(c)+o(1))\mathbb{E}\left[\sum_{k=1}^{d}a_k^2b_k^2\right] \\
&= (\chi(c)+o(1))\mathbb{E}\left[\sum_{k=1}^{d}a_k^2b_k^2\right] \\
&= \chi(c)\sum_{k=1}^{d}\mathbb{E}[a_k^2]\mathbb{E}[b_k^2] + o\left(\frac{1}{d}\right),
\end{aligned}$$

where the last line holds due to the argument in the proof of Lemma 11. Here $\chi(c)$ is some function of c. This gives us that $\text{Var}[M_{ij}] = \frac{1}{d}\chi(c) + o\left(\frac{1}{d}\right)$ for $i \neq j$. For $i = j$, we use Sonthalia and Nadakuditi [29] to see that the variance is $O\left(\frac{1}{d}\right)$. □

Lemma 13. *For $c < 1$, we have that $\mathbb{E}[QQ^T] = (1-c)I_r$ and the variance of each entry is $O\left(\frac{1}{d}\right)$. Further,*

$$\mathbb{E}[(QQ^T)^{-1}] = \frac{1}{1-c}I_r + O\left(\frac{1}{d}\right).$$

and each element has variance $O(1/d)$

Proof. Recall that $Q = V^T(I - A_{trn}A_{trn}^\dagger)$. We thus have that

$$\begin{aligned}
P^TP &= V^T(I - A_{trn}^\dagger A_{trn})V. \\
&= V^TV - V^TA_{trn}^\dagger A_{trn}V \\
&= I_r - V^T\tilde{V}\tilde{\Sigma}^\dagger\tilde{\Sigma}\tilde{V}^TV \\
&= I_r - R\begin{bmatrix} I_d & 0 \\ 0 & 0_{N-d}\end{bmatrix}R^T.
\end{aligned}$$

Where R is a uniformly random $r \times N$ unitary matrix. Then by symmetry (of the sign of rows of R), we have that

$$\mathbb{E}[QQ^T] = I_r - cI_r = (1-c)\, I_r.$$

Next notice that

$$\mathbb{E}[QQ^T] = V^T(I - \mathbb{E}[A_{trn}^\dagger A_{trn}])V,$$

thus to compute the variance, we first compute the variance of $(A_{trn}^\dagger A_{trn})_{ij}$. For this, we first note that

$$\begin{bmatrix} cI_d & 0 \\ 0 & 0\end{bmatrix} = \mathbb{E}[A_{trn}^\dagger A_{trn}] = \mathbb{E}[A_{trn}^\dagger A_{trn}A_{trn}^\dagger A_{trn}].$$

Since $A_{trn}^\dagger A_{trn}$ is symmetric, we can see that

$$\sum_k^d ((A_{trn}^\dagger A_{trn})_{ik})^2 = \begin{cases} c & i \leq d \\ 0 & i > d\end{cases}.$$

From Lemma 15 in [29], we have that $\mathbb{E}[((A_{trn}^\dagger A_{trn})_{ii})^2] = c^2 + \frac{2c}{N} + o(1)$. Then combining this with the computation above and using symmetry, we have that for $i \neq j$ and $\min(i,j) \leq d$

$$\mathbb{E}[(A_{trn}^\dagger A_{trn})_{ij}^2] = \frac{1}{d-1}\left(\frac{1}{c} - \frac{1}{c^2} + \frac{3}{cd} + o(1)\right).$$

Now consider the other (full) SVD of X_{trn} given by $\hat{U}_{d\times d}\hat{\Sigma}_{d\times N}\hat{V}_{N\times N}^T$. Note that the top left $r \times r$ block of $\hat{\Sigma}$ is Σ_{trn}, and the first r rows of $\hat{V}$ give V. Note that since $\hat{V}^T\tilde{V}$ is still uniformly random, the variance argument above follows for $\hat{V}^TA_{trn}^\dagger A_{trn}\hat{V}$. Additionally, for $i, j \leq r$, $(\hat{V}^TA_{trn}^\dagger A_{trn}\hat{V})_{ij} = (V^TA_{trn}^\dagger A_{trn}V)_{ij}$ Thus, we see that for $i, j \leq r$,

$$\mathbb{E}[((V^TA_{trn}^\dagger A_{trn}V)_{ij})^2] = \frac{1}{d-1}\left(c - c^2 + \frac{2}{cd} + o(1)\right).$$

Thus, finally, we have that arranged as a matrix

$$\mathbb{E}[QQ^T \odot QQ^T] = O\left(\frac{1}{d}\right).$$

By an analogous symmetry argument, we can show that

$$\text{Var}\left((V^TA_{trn}^\dagger A_{trn}V)_{ij}^2\right) = O\left(\frac{1}{d}\right).$$

In principle, one can get a faster decay bound with a more sophisticated argument, but this is sufficient for our purposes. Now, by Lemma 4, we get that

$$\mathbb{E}[(QQ^T)^{-1}] = \frac{1}{1-c} I_r + O\left(\frac{1}{d}\right).$$

and each element has variance $O(1/d)$.

$\square$

Lemma 14. *For $c < 1$,*

$$\mathbb{E}\left[\Sigma_{trn}^{-1} H_1 \Sigma_{trn}^{-1}\right] = \frac{1}{1-c}\Sigma_{trn}^{-2} + \frac{1}{\eta_{trn}^2}\frac{c}{1-c} I_r + o(1)$$

and the variance of each element is $O\left(\frac{1}{d}\right)$. Additionally

$$\mathbb{E}\left[\Sigma_{trn} H_1^{-1} \Sigma_{trn}\right] = (1-c)\eta_{trn}^2(\eta_{trn}^2\Sigma_{trn}^{-2} + cI_r)^{-1} + o(1),$$

and the variance of each term is $O\left(\frac{1}{d}\right)$

Proof. Recall that

$$H_1 = K^T K + Z^T (QQ^T)^{-1} Z = K^T K + Z^T \Sigma_{trn}^{-1} (\Sigma_{trn} (P^T P)^{-1} \Sigma_{trn}) \Sigma_{trn}^{-1} Z.$$

Using Lemmas 6, 11 and 13 along with an argument analogous to the one in Lemma 7, we get that

$$\mathbb{E}[\Sigma_{trn}^{-1} H_1 \Sigma_{trn}^{-1}] = \frac{1}{1-c}\Sigma_{trn}^{-2} + \frac{1}{\eta_{trn}^2}\frac{c}{1-c} I_r + O\left(\frac{1}{d}\right) + o(1)$$

and the variance of each element is $O\left(\frac{1}{d}\right)$.

For the inverse, we define $\delta H_1 := H_1 - \mathbb{E}[H_1]$ and by an argument analogous to the one in the proof of Lemma 7, we get that

$$\mathbb{E}\left[\Sigma_{trn} H_1^{-1} \Sigma_{trn}\right] = (1-c)\eta_{trn}^2(\eta_{trn}^2\Sigma_{trn}^{-2} + cI_r)^{-1} + o(1)$$

and the variance of each term is $O\left(\frac{1}{d}\right)$. $\square$

Lemma 15. *When $c < 1$, we have for $W = W_{opt}$ that*

$$\mathbb{E}[\|W\|_F^2] = \frac{c^2}{1-c}\operatorname{Tr}\left(\Sigma_{trn}^2\left(\Sigma_{trn}^2 + \frac{1}{\eta_{trn}^2} I_r\right)(\Sigma_{trn}^2 c + \eta_{trn}^2 I_r)^{-2}\right) + o(1).$$

Proof. Again, like in Lemma 8, we first use the estimates for the expectations from the lemmas above to get an estimate for the expectation of $\|W\|_F^2$, and then bound the deviation from it using the variance estimates in this section. We see that the first term in $\operatorname{Tr}(W^T W)$ is

$$\operatorname{Tr}((A_{trn}^\dagger)^T K (H_1^{-1})^T \Sigma_{trn}^2 H_1^{-1} K^T A_{trn}^\dagger) = \operatorname{Tr}(K^T A_{trn}^\dagger (A_{trn}^\dagger)^T K (H_1^{-1})^T \Sigma_{trn}^2 H_1^{-1}).$$

Then using Lemma 12 along with cyclic invariance of traces, we see that this is estimated by

$$\frac{1}{\eta_{trn}^2}\frac{c^2}{(1-c)^3}\operatorname{Tr}(\Sigma_{trn}(H_1^{-1})^T \Sigma_{trn}^2 H_1^{-1} \Sigma_{trn}) + o(1).$$

Then using Lemma 14, we get that this is estimated by

$$\begin{aligned}
&\eta_{trn}^2 \frac{c^2}{(1-c)^3}(1-c)^2 (cI_r + \eta_{trn}^2 \Sigma_{trn}^{-2})^{-2} + o(1) \\
&= \eta_{trn}^2 \frac{c^2}{1-c}\operatorname{Tr}\left(\Sigma_{trn}^4 (\Sigma_{trn}^2 c + \eta_{trn}^2 I_r)^{-2}\right) + o(1).
\end{aligned}$$

The second term is

$$\operatorname{Tr}(((QQ^T)^{-1})^T Z (H_1^{-1})^T \Sigma_{trn}^2 H_1^{-1} Z^T (QQ^T)^{-1} H H^T).$$

We can rewrite this as

$$\operatorname{Tr}(((QQ^T)^{-1})^T Z\Sigma_{trn}^{-1}(\Sigma_{trn}(H_1^{-1})^T\Sigma_{trn})(\Sigma_{trn}H_1^{-1}\Sigma_{trn})\Sigma_{trn}^{-1}Z^T(QQ^T)^{-1}HH^T).$$

Using Lemmas 3 and 6, we can estimate its expectation by

$$\frac{1}{\eta_{trn}^2}\frac{c^2}{1-c}\operatorname{Tr}\left(((QQ^T)^{-1})^T\Sigma_{trn}^{-1}(\Sigma_{trn}(H_1^{-1})^T\Sigma_{trn})(\Sigma_{trn}H_1^{-1}\Sigma_{trn})\Sigma_{trn}^{-1}(QQ^T)^{-1}\right)+o(1).$$

Then using Lemma 13 and the fact that $H_1^T = H_1$, we get that this be further estimated by

$$\frac{1}{\eta_{trn}^2}\frac{c^2}{(1-c)^3}\operatorname{Tr}(\Sigma_{trn}^{-1}(\Sigma_{trn}(H_1^{-1})\Sigma_{trn})^2\Sigma_{trn}^{-1})+o(1).$$

Then using Lemma 14, we can simplify this estimate to

$$\begin{aligned}\frac{1}{\eta_{trn}^2}\frac{c^2}{(1-c)^3}(1-c)^2\eta_{trn}^4(cI_r+\eta_{trn}^2\Sigma_{trn}^{-2})^{-2}+o(1)\\ =\eta_{trn}^2\frac{c^2}{1-c}\operatorname{Tr}\left(\Sigma_{trn}^2(\Sigma_{trn}^2c+\eta_{trn}^2I_r)^{-2}\right)+o(1).\end{aligned}$$

The cross term in $\operatorname{Tr}(W^TW)$ is

$$-2\operatorname{Tr}((A_{trn}^\dagger)^TK(H_1^{-1})^T\Sigma_{trn}^2H_1^{-1}Z^T(QQ^T)^{-1}H).$$

Here the term (after cyclically permuting) that we should focus on is

$$\operatorname{Tr}(H(A_{trn}^\dagger)^TK) = -\operatorname{Tr}(V_{trn}^TA_{trn}^\dagger(A_{trn}^\dagger)^TA_{trn}^\dagger\Sigma_{trn}U).$$

Here since $A_{trn} = \eta_{trn}\tilde{U}\tilde{\Sigma}\tilde{V}^T$ and $\tilde{U},\tilde{V}$ are independent of each other, we see that using ideas from Lemma 8 in [29] and extending them to rank r as before, the expectation of this term is 0 with $O(1/d)$ variance. Thus, the whole cross-term has an expectation equal to 0.

Again, to bound the deviation from this estimate, note that for real valued random variables X,Y we have that $|\mathbb{E}[XY]-\mathbb{E}[X]\mathbb{E}[Y]| = |Cov(X,Y)| \le \sqrt{\operatorname{Var}(X)\operatorname{Var}(Y)}$. For real valued random variables X,Y,Z,W, we have the following fact, from [43].

$$\begin{aligned}Cov(XY,WZ) = \mathbb{E}X\mathbb{E}WCov(Y,Z)+\mathbb{E}Y\mathbb{E}ZCov(X,W)+\mathbb{E}X\mathbb{E}ZCov(Y,W)+\\ \mathbb{E}Y\mathbb{E}WCov(X,Z)+Cov(X,W)Cov(Y,Z)+Cov(Y,W)Cov(X,Z).\end{aligned}$$

We repeatedly apply these two to upper bound the deviation between the product of the expectations in the estimates above and the expectation of the product. It is then straightforward to see that since all variances are $O(1/d)$, the estimation error is $O(1/d) = o(1)$.

Finally, combining the terms, we get that

$$\mathbb{E}[\|W\|_F^2] = \frac{c^2}{1-c}\operatorname{Tr}\left(\Sigma_{trn}^2\left(\Sigma_{trn}^2+\frac{1}{\eta_{trn}^2}I_r\right)(\Sigma_{trn}^2c+\eta_{trn}^2I_r)^{-2}\right)+o(1).$$

□

Theorem 10. *When $d < N-r$ and $\beta = I$, then the test error $\mathcal{R}(W,X_{tst})$ for $W = W_{opt}$ is given by*

$$\begin{aligned}\frac{\eta_{trn}^4}{N_{tst}}\|\left(\Sigma_{trn}^2c+\eta_{trn}^2I\right)^{-1}L\|_F^2\\ +\frac{\eta_{tst}^2}{d}\frac{c^2}{1-c}\operatorname{Tr}\left(\Sigma_{trn}^2\left(\Sigma_{trn}^2+\frac{1}{\eta_{trn}^2}I_r\right)(\Sigma_{trn}^2c+\eta_{trn}^2I_r)^{-2}\right)+o\left(\frac{1}{d}\right).\end{aligned}$$

Proof. Note from theorem 9 that $\mathcal{R}(W,X_{tst}) = \frac{1}{N_{tst}}\|U\Sigma_{trn}H_1^{-1}Z^T(QQ^T)^{-1}\Sigma_{trn}^{-1}L\|_F^2 + \frac{\eta_{tst}^2}{d}\|W\|_F^2$.

To compute the first term, we observe that it is given by

$$\frac{1}{N_{tst}}Tr(U\Sigma_{trn}H_1^{-1}Z^T(QQ^T)^{-1}\Sigma_{trn}^{-1}LL^T\Sigma_{trn}^{-1}(QQ^T)^{-1}ZH_1^{-1}\Sigma_{trn}U^T).$$

This can be rewritten using cyclic invariance as

$$\frac{1}{N_{tst}}Tr(U^TU\Sigma_{trn}H_1^{-1}Z^T\Sigma_{trn}^{-1}\Sigma_{trn}(QQ^T)^{-1}\Sigma_{trn}^{-1}LL^T\Sigma_{trn}^{-1}(QQ^T)^{-1}\Sigma_{trn}\Sigma_{trn}^{-1}ZH_1^{-1}\Sigma_{trn}).$$

We apply Lemmas 13, 14 and 6 to get that its expectation can be estimated by

$$\begin{aligned}
&\frac{1}{N_{tst}}Tr\left(\left((c-1)\eta_{trn}^2(\eta_{trn}^2 I + c\Sigma_{trn}^2)^{-1}\right)^2\left(\frac{1}{1-c}\right)^2 LL^T\right) + o(1/d)\\
&= \frac{\eta_{trn}^4}{N_{tst}}Tr\left(\left(\Sigma_{trn}^2 c + \eta_{trn}^2 I\right)^{-2} LL^T\right) + o(1/d)\\
&= \frac{\eta_{trn}^4}{N_{tst}}\|\left(\Sigma_{trn}^2 c + \eta_{trn}^2 I\right)^{-1} L\|_F^2 + o(1/d).
\end{aligned}$$

We get $o\left(\frac{1}{d}\right)$ due to the Σ_{trn}^{-2} term. Again, we can argue as in the proof of Lemma 15 to bound the deviation of the true expectation from this estimate by $o(1/d)$, noting that since train and test data assumptions are decoupled, LL^T/N_{tst} can be treated as constant as N grows.

Combining this with Lemma 8, we get that

$$\begin{aligned}
&\frac{\eta_{trn}^4}{N_{tst}}\|\left(\Sigma_{trn}^2 c + \eta_{trn}^2 I\right)^{-1} L\|_F^2\\
&\qquad + \frac{\eta_{tst}^2}{d}\frac{c^2}{1-c}\operatorname{Tr}\left(\Sigma_{trn}^2\left(\Sigma_{trn}^2 + \frac{1}{\eta_{trn}^2}I_r\right)(\Sigma_{trn}^2 c + \eta_{trn}^2 I_r)^{-2}\right) + o\left(\frac{1}{d}\right).
\end{aligned}$$

□

Theorem 1 (In-Subspace Test Error)**.** *Let $r < |d-N|$. Let the SVD of X_{trn} be $U\Sigma_{trn}V_{trn}^T$, let $L := U^TX_{tst}$, $\beta_U := U^T\beta$, and $c := d/N$. Under our setup and Assumptions 1 and 2, the test error (Equation 1) is given by the following. If $c < 1$ (under-parameterized regime)*

$$\begin{aligned}
\mathcal{R}(W_{opt}, UL) &= \frac{\eta_{trn}^4}{N_{tst}}\left\|\beta_U^T(\Sigma_{trn}^2 c + \eta_{trn}^2 I)^{-1}L\right\|_F^2\\
&\quad + \frac{\eta_{tst}^2}{d}\frac{c^2}{1-c}\operatorname{Tr}\left(\beta_U\beta_U^T\Sigma_{trn}^2\left(\Sigma_{trn}^2 + \frac{1}{\eta_{trn}^2}I\right)\left(\Sigma_{trn}^2 c + \eta_{trn}^2 I\right)^{-2}\right) + o\left(\frac{1}{N}\right)
\end{aligned}$$

If $c > 1$ (over-parameterized regime)

$$\begin{aligned}
\mathcal{R}(W_{opt}, UL) &= \frac{\eta_{trn}^4}{N_{tst}}\left\|\beta_U^T(\Sigma_{trn}^2 + \eta_{trn}^2 I)^{-1}L\right\|_F^2\\
&\quad + \frac{\eta_{tst}^2}{d}\frac{c}{c-1}\operatorname{Tr}(\beta_U\beta_U^T(I + \eta_{trn}^2\Sigma_{trn}^{-2})^{-1}) + O\left(\frac{\|\Sigma_{trn}\|^2}{N^2}\right) + o\left(\frac{1}{N}\right)
\end{aligned}$$

Proof. The version for $\beta = I$ follows immediately from Theorem 8 and Theorem 10.

We now demonstrate how the the general version is a straightforward repetition of the proofs of the two theorems. First denote by Z_{opt} the minimum norm solution to the denoising problem (where $\beta = I$). Then $Z_{opt} = X_{trn}(X_{trn} + A_{trn})^\dagger$ and note that

$$W_{opt} = Y_{trn}(X_{trn} + A_{trn})^\dagger = \beta^T X_{trn}(X_{trn} + A_{trn})^\dagger = \beta^T Z_{opt}$$

We present the adaptation of Lemma 8, the other lemmas can be adapted accordingly.

We first use the estimates for the expectations from the lemmas to get an estimate for $\|W_{opt}\|_F^2 = \|\beta^T Z_{opt}\|_F^2$, and then bound the deviation from it using the variance estimates above. We begin the calculation as

$$\|\beta^T Z_{opt}\|_F^2 = \operatorname{Tr}(Z_{opt}^T\beta\beta^T Z_{opt})$$

Using Lemma 1, we see that the trace has three terms. The first term is

$$\mathrm{Tr}(H^T(K_1^{-1})^T Z((P^TP)^{-1})^T\Sigma_{trn}^T U^T\beta\beta^T U\Sigma_{trn}(P^TP)^{-1}Z^TK_1^{-1}H)$$

Using $\beta_U^T = \beta_{opt}^T U$ Then since the trace is invariant under cyclic permutations, we get the following term

$$\mathrm{Tr}(\beta_U^T\Sigma_{trn}(P^TP)^{-1}Z^TK_1^{-1}HH^T(K_1^{-1})^TZ((P^TP)^{-1})^T\Sigma_{trn}^T\beta_U)$$

The rest of the proof for this term is the same as Lemma 8.

The second term in $\mathrm{Tr}(W^T\beta\beta^TW)$ is

$$-2\,\mathrm{Tr}(H^T(K_1^{-1})^TZ^T((P^TP)^{-1})^T\Sigma_{trn}^T\beta_U\beta_U^T\Sigma_{trn}Z^{-1}HH^TZP^\dagger)$$

Then the rest of the proof for this term is identical to the one in the proof of Lemma 8.

Finally, the last term in $\mathrm{Tr}(W^T\beta\beta^TW)$ is

$$\mathrm{Tr}((P^\dagger)^TZ^T(K_1^{-1})^THH^T(Z^{-1})^T\Sigma_{trn}^T\beta_U\beta_U^T\Sigma_{trn}Z^{-1}HH^TK_1^{-1}P^\dagger)$$

The rest of the proof is the same again, after using the cyclic invariance of the trace.

□

F.2 Proof of Corollary 1, The Distribution Shift Bound

We first prove Theorem 2, bounding the difference in generalization error in terms of the change in the test set. Recall the theorem below.

Theorem 2 (Test Set Shift Bound)**.** *Under the assumptions of Theorem 1, consider a linear regressor W_{opt} trained on training data $X_{trn} = U\Sigma_{trn}V_{trn}^T$ with Σ_{trn} such that $\sigma_r(X_{trn}) > M$, and tested on test data $X_{tst,1} = UL_1$ and $X_{tst,2} = UL_2$ with noise $A_{tst,1}, A_{tst,2}$ with the same variance η_{tst^2}/d. Then, the generalization errors $\mathcal{R}_1$ and $\mathcal{R}_2$ differ for $c < 1$ by*

$$|\mathcal{R}_2 - \mathcal{R}_1| \le \frac{\sigma_1(\beta)^2}{N_{tst}}\frac{\eta_{trn}^4 r}{(\sigma_r(X_{trn})^2 f(c) + \eta_{trn}^2)^2}\|L_2L_2^T - L_1L_1^T\|_F + o\left(\frac{1}{N}\right)$$

where $f(c) = c$ for $c < 1$ and $f(c) = 1$ for $c \ge 1$. We add $O(\|\Sigma_{trn}\|_F^2/N^2)$ to the bound when $c > 1$.

Proof. We will first show this for $c < 1$. Let $\mathcal{R}_i := \mathcal{R}(W_{opt}, X_{tst,i})$. Remember that the test error is given by

$$\begin{aligned}\mathcal{R}_i = &\frac{\eta_{trn}^4}{N_{tst}}\left\|\beta_U^T(\Sigma_{trn}^2c + \eta_{trn}^2 I)^{-1}L_i\right\|_F^2 \\ &+ \eta_{tst}^2\eta_{trn}^2\frac{1}{d}\frac{c^2}{1-c}\,\mathrm{Tr}\left(\beta_U\beta_U^T\Sigma_{trn}^2\left(\Sigma_{trn}^2 + \frac{1}{\eta_{trn}^2}I\right)\left(\Sigma_{trn}^2c + \eta_{trn}^2 I\right)^{-2}\right) + o\left(\frac{1}{N}\right)\end{aligned}$$

Note that the second term above has no dependence on $X_{tst,i}$, so the difference is given by

$$\begin{aligned}\mathcal{R}_2 - \mathcal{R}_1 &= \frac{\eta_{trn}^4}{N_{tst}}\left(\left\|\beta_U^T(\Sigma_{trn}^2c + \eta_{trn}^2 I)^{-1}L_2\right\|_F^2 - \left\|\beta_U^T(\Sigma_{trn}^2c + \eta_{trn}^2 I)^{-1}L_1\right\|_F^2\right) \\ &\quad + o\left(\frac{1}{N}\right) \\ &= \frac{\eta_{trn}^4}{N_{tst}}Tr\left((\Sigma_{trn}^2c + \eta_{trn}^2 I)^{-1}\beta_U\beta_U^T(\Sigma_{trn}^2c + \eta_{trn}^2 I)^{-1}(L_2L_2^T - L_1L_1^T)\right) + o\left(\frac{1}{N}\right) \\ &\overset{(i)}{\le} \frac{\eta_{trn}^4}{N_{tst}}\|(\Sigma_{trn}^2c + \eta_{trn}^2 I)^{-1}\beta_U\beta_U^T(\Sigma_{trn}^2c + \eta_{trn}^2 I)^{-1}\|_F\|(L_2L_2^T - L_1L_1^T)\|_F + o\left(\frac{1}{N}\right) \\ &= \frac{\eta_{trn}^4}{N_{tst}}\|\beta_U\beta_U^T(\Sigma_{trn}^2c + \eta_{trn}^2 I)^{-2}\|_F\|(L_2L_2^T - L_1L_1^T)\|_F + o\left(\frac{1}{N}\right) \\ &\overset{(ii)}{\le} \frac{\eta_{trn}^4}{N_{tst}}\|\beta_U\beta_U^T\|_2\|(\Sigma_{trn}^2c + \eta_{trn}^2 I)^{-2}\|_F\|(L_2L_2^T - L_1L_1^T)\|_F + o\left(\frac{1}{N}\right)\end{aligned}$$

where (i) above is by the Cauchy-Schwarz inequality for the Frobenius norm and (ii) above holds since $\|AB\|_F \leq \|A\|_2 \|B\|_F$. So, for Σ_{trn} with lower bounded diagonal entries $\sigma_i > M$, we have that

$$\begin{aligned}
|\mathcal{R}_2 - \mathcal{R}_1| &\leq \frac{\eta_{trn}^4 r}{N_{tst}(\sigma_r(X_{trn})^2 c + \eta_{trn}^2)^2} \|\beta_U \beta_U^T\|_2 \|(L_2 L_2^T - L_1 L_1^T)\|_F + o\left(\frac{1}{N}\right) \\
&= \frac{\eta_{trn}^4 r}{N_{tst}(\sigma_r(X_{trn})^2 c + \eta_{trn}^2)^2} \|U^T \beta \beta^T U\|_2 \|(L_2 L_2^T - L_1 L_1^T)\|_F + o\left(\frac{1}{N}\right) \\
&= \frac{\eta_{trn}^4 r}{N_{tst}(\sigma_r(X_{trn})^2 c + \eta_{trn}^2)^2} \|\beta \beta^T\|_2 \|(L_2 L_2^T - L_1 L_1^T)\|_F + o\left(\frac{1}{N}\right) \\
&= \frac{\sigma_1(\beta)^2}{N_{tst}} \frac{\eta_{trn}^4 r}{(\sigma_r(X_{trn})^2 c + \eta_{trn}^2)^2} \|L_2 L_2^T - L_1 L_1^T\|_F + o\left(\frac{1}{N}\right)
\end{aligned}$$

Similarly, for $c > 1$, we have that

$$|\mathcal{R}_2 - \mathcal{R}_1| \leq \frac{\sigma_1(\beta)^2}{N_{tst}} \frac{\eta_{trn}^4 r}{(\sigma_r(X_{trn})^2 + \eta_{trn}^2)^2} \|L_2 L_2^T - L_1 L_1^T\|_F + O\left(\frac{\|\Sigma_{trn}\|_F^2}{N^2}\right) + o\left(\frac{1}{N}\right)$$

□

We now prove our corollary below.

Corollary 1 (Distribution Shift Bound)**.** *Let W_{opt} be tested on test data $X_{tst,1} = UL_1$ and $X_{tst,2} = UL_2$ generated possibly dependently from distributions supported in the span of U with mean $U\mu_i$ and covariance $\Sigma_{U,i} = U\Sigma_i U^T$ respectively. Let $f(c) = c$ for $c < 1$ and $f(c) = 1$. Then, the difference in* generalization errors *$\mathcal{G}_i := \mathbb{E}_{X_{tst,i}}[\mathcal{R}(W_{opt}, X_{tst,i})]$ is bounded for $c < 1$ by*

$$|\mathcal{G}_2 - \mathcal{G}_1| \leq \frac{\sigma_1(\beta)^2 \eta_{trn}^4 r}{(\sigma_r(X_{trn})^2 f(c) + \eta_{trn}^2)^2} \|\Sigma_2 - \Sigma_1 + \mu_2 \mu_2^T - \mu_1 \mu_1^T\|_F + o\left(\frac{1}{N}\right).$$

We add $O(\|\Sigma_{trn}\|_F^2 / N^2)$ to the bound when $c \geq 1$.

Proof. Let $\bar{L}_i := L_i - [\mu_i \, \mu_i \, \dots \, \mu_i]$ be the centered version of the test data matrix. In that case, $\mathbb{E}_{X_{tst,i}}[\bar{L}_i] = \mathbb{E}_{X_{tst,i}}[U^T \bar{X}_{tst,i}] = 0$ and

$$\mathbb{E}_{X_{tst,i}}[\bar{L}_i \bar{L}_i^T] = \mathbb{E}_{X_{tst,i}}[U^T \bar{X}_{tst,i} \bar{X}_{tst,i}^T U] = N_{tst} \Sigma_i$$

Now note the following elementary computation.

$$\begin{aligned}
\mathbb{E}_{X_{tst,i}}[L_i L_i^T] &= \mathbb{E}_{X_{tst,i}}[(\bar{L}_i + [\mu_i \, \mu_i \, \dots \, \mu_i])(\bar{L}_i + [\mu_i \, \mu_i \, \dots \, \mu_i])^T] \\
&= \mathbb{E}_{X_{tst,i}}[\bar{L}_i \bar{L}_i^T] + 0 + 0 + N_{tst} \mu_i \mu_i^T \\
&= N_{tst} \Sigma_{trn} + N_{tst} \mu_i \mu_i^T
\end{aligned}$$

We can now follow the initial part of the proof of Theorem 2 to get the following for $c < 1$.

$$\begin{aligned}
\mathcal{G}_2 - \mathcal{G}_1 &= \frac{\eta_{trn}^4}{N_{tst}} Tr\left(\beta_U \beta_U^T (\Sigma_{trn}^2 c + \eta_{trn}^2 I)^{-2} (\mathbb{E}_{X_{tst},2}[L_2 L_2^T] - \mathbb{E}_{X_{tst},1}[L_1 L_1^T])\right) + o\left(\frac{1}{N}\right) \\
&= \eta_{trn}^4 Tr\left(\beta_U \beta_U^T (\Sigma_{trn}^2 c + \eta_{trn}^2 I)^{-2} (\Sigma_2 - \Sigma_1 + \mu_2 \mu_2^T - \mu_1 \mu_1^T)\right) + o\left(\frac{1}{N}\right)
\end{aligned}$$

Now, we can follow the rest of the proof of Theorem 2 to complete the proof. □

F.3 Proofs for Theorem 3, Out-of-Subspace Generalization

Theorem 3 (Out-of-Subspace Shift Bound). *If we have the same training data and solution W_{opt} assumptions as in Theorem 1. Then, for* **any** *X_{tst} for which there exists an L and an $\alpha > 0$ such that $\|X_{tst} - UL\|_F \leq \alpha$, and A_{tst} that satisfies 1,2 from Assumption 2, we have that the generalization error $\mathcal{R}(W_{opt}, X_{tst})$ satisfies*

$$|\mathcal{R}(W_{opt}, X_{tst}) - \mathcal{R}(W_{opt}, UL)| \leq \alpha^2 \sigma_1(W_{opt} + I)^2.$$

Proof. Here we see that

$$\begin{aligned} \|(I-W)X_{tst} - (I-W)UL\|_F^2 &= \|(I-W)(X_{tst} - UL)\|_F^2 \\ &\leq \sigma_1(W-I)^2 \|X_{tst} - UL\|_F^2 \\ &= \alpha^2 \sigma_1(W-I)^2 \end{aligned}$$

The inequality is due to Cauchy-Schwarz inequality. Then using the reverse triangle inequality, we have that

$$\left| \|(I-W)X_{tst}\|_F^2 - \|(I-W)UL\|_F^2 \right| \leq \alpha^2 \sigma_1(W+I)^2.$$

□

F.4 Proofs for Corollary 4, Generalization Error

Corollary 4 (Generalization Error). *In the setting of Theorem 1, if we further assume that the data X_{tst} is generated possibly dependently from distributions supported in the span of U with mean $U\mu$ and covariance $\Sigma_U = U\Sigma U^T$, then we can remove the $\frac{1}{N_{tst}}$ and replace L with $(\Sigma + \mu\mu^T)^{1/2}$ in the expression for test error to get the generalization error $\mathbb{E}_{X_{tst}}[\mathcal{R}(W_{opt}, X_{tst})]$.*

Proof. We begin by noting that the variance term is independent of X_{tst}. Hence we only need to focus on the bias term. Let $\bar{L} := L - [\mu\ \mu\ \ldots\ \mu]$ be the centered version of the test data matrix. In that case, $\mathbb{E}_{X_{tst,i}}[\bar{L}] = \mathbb{E}_{X_{tst,i}}[U^T \bar{X}_{tst,i}] = 0$ and

$$\mathbb{E}_{X_{tst,i}}[\bar{L}\bar{L}^T] = \mathbb{E}_{X_{tst,i}}[U^T \bar{X}_{tst,i} \bar{X}_{tst,i}^T U] = N_{tst}\Sigma$$

Now note the following elementary computation.

$$\begin{aligned} \mathbb{E}_{X_{tst,i}}[LL^T] &= \mathbb{E}_{X_{tst,i}}[(\bar{L} + [\mu\ \mu\ \ldots\ \mu])(\bar{L} + [\mu\ \mu\ \ldots\ \mu])^T] \\ &= \mathbb{E}_{X_{tst,i}}[\bar{L}\bar{L}^T] + 0 + 0 + N_{tst}\mu\mu^T \\ &= N_{tst}\Sigma_{trn} + N_{tst}\mu\mu^T \end{aligned}$$

Consider the following sequence on computations about the bias term.

$$\begin{aligned} &\mathbb{E}_{X_{tst}}\left[\frac{\eta_{trn}^4}{N_{tst}} \left\|\beta_U^T(\Sigma_{trn}^2 c + \eta_{trn}^2 I)^{-1} L\right\|_F^2\right] \\ &= \frac{\eta_{trn}^4}{N_{tst}} Tr\left(\beta_U^T(\Sigma_{trn}^2 c + \eta_{trn}^2 I)^{-1} \mathbb{E}_{X_{tst}}[LL^T](\Sigma_{trn}^2 c + \eta_{trn}^2 I)^{-1}\beta_U\right) \\ &= \frac{\eta_{trn}^4}{N_{tst}} Tr\left(\beta_U^T(\Sigma_{trn}^2 c + \eta_{trn}^2 I)^{-1}(\Sigma + \mu\mu^T)(\Sigma_{trn}^2 c + \eta_{trn}^2 I)^{-1}\beta_U\right) \\ &= \frac{\eta_{trn}^4}{N_{tst}} \left\|\beta_U^T(\Sigma_{trn}^2 c + \eta_{trn}^2 I)^{-1}(\Sigma + \mu\mu^T)^{1/2}\right\|_F^2 \end{aligned}$$

This establishes our claim. □

F.5 Proof for Theorem 4, Test Error for W^*

Theorem 4 (Test Error for W^*). *In the same setting as Theorem 1, we have that $W^* = \beta_U^T\left(I + \frac{\eta_{trn}^2}{c}\Sigma_{trn}^{-2}\right)^{-1} U^T$ and*

$$\mathcal{R}(W^*, UL) = \frac{\eta_{trn}^4 N^2}{d^2}\left\|\beta_U^T\left(\Sigma_{trn}^2 + \frac{\eta_{trn}^2 N}{d} I\right)^{-1} L\right\|_F^2 + \frac{\eta_{tst}^2}{d} Tr\left(\beta_U \beta_U^T\left(I + \frac{\eta_{trn}^2 N}{d}\Sigma_{trn}^{-2}\right)^{-2}\right).$$

Proof. To prove the first part of the theorem, we first note that

$$\mathbb{E}_{A_{trn}}\left[\|Y_{trn} - W(X_{trn} + A_{trn})\|_F^2\right] = \|Y_{trn} - WX_{trn}\|_F^2 + \frac{\eta_{trn}^2 N}{d}\|W\|_F^2.$$

Solving this is equivalent to solving

$$\|\begin{bmatrix}Y_{trn} & 0\end{bmatrix} - W\begin{bmatrix}X_{trn} & \mu I\end{bmatrix}\|_F^2.$$

where $\mu^2 = \frac{\eta_{trn}^2 N}{d}$. We know from classical linear algebra that the solution to the above is

$$W^* = \begin{bmatrix}\beta^T X_{trn} & 0\end{bmatrix}\begin{bmatrix}X_{trn} & \mu I\end{bmatrix}^\dagger.$$

Using Lemmas 5 and 6 from [44], we have that if $X_{trn} = U\Sigma_{trn}V_{trn}^T$ where U is d by d, Σ_{trn} is d by d and V_{trn} is $N \times d$, then

$$\begin{bmatrix}X_{trn} & \mu I\end{bmatrix} = U \underbrace{\begin{bmatrix} \sqrt{\sigma_1(X_{trn})^2 + \mu^2} & 0 & \cdots & & & 0 \\ 0 & \ddots & 0 & & & \\ \vdots & & \sqrt{\sigma_r(X_{trn})^2 + \mu^2} & & & \vdots \\ & & 0 & \mu & 0 & \\ & & & 0 & \ddots & 0 \\ 0 & & & & 0 & \mu \end{bmatrix}}_{\hat{\Sigma}} \begin{bmatrix} V_{trn}\Sigma_{trn}\hat{\Sigma}^{-1} \\ \mu U\hat{\Sigma}^{-1} \end{bmatrix}^T.$$

Thus, we have that

$$W^* = \begin{bmatrix}\beta^T U\Sigma_{trn}V_{trn}^T & 0\end{bmatrix}\begin{bmatrix} V_{trn}\Sigma_{trn}\hat{\Sigma}^{-1} \\ \mu U\hat{\Sigma}^{-1} \end{bmatrix}\begin{bmatrix} \frac{1}{\sqrt{\sigma_1(X_{trn})^2+\mu^2}} & 0 & \cdots & & & 0 \\ 0 & \ddots & 0 & & & \\ \vdots & & \frac{1}{\sqrt{\sigma_r(X_{trn})^2+\mu^2}} & & & \vdots \\ & & 0 & \frac{1}{\mu} & 0 & \\ & & & 0 & \ddots & 0 \\ 0 & & & & 0 & \frac{1}{\mu} \end{bmatrix} U^T.$$

Simplifying, we get

$$\begin{aligned}
W^* &= \beta_U^T \Sigma_{trn}^2 \hat{\Sigma}^{-2} U^T \\
&= \beta_U^T \begin{bmatrix} \frac{\sigma_1(X_{trn})^2}{\sigma_1(X_{trn})^2+\mu^2} & 0 & \cdots & & & 0 \\ 0 & \ddots & 0 & & & \\ \vdots & & \frac{\sigma_r(X_{trn})^2}{\sigma_r(X_{trn})^2+\mu^2} & & & \vdots \\ & & 0 & 0 & 0 & \\ & & & 0 & \ddots & 0 \\ 0 & & & & 0 & 0 \end{bmatrix} U^T \\
&= \beta_U^T \Sigma_{trn}^2 (\Sigma_{trn}^2 + \mu^2 I)^{-1} U^T \\
&= \beta_U^T \Sigma_{trn}^2 \left(\Sigma_{trn}^2 + \frac{\eta_{trn}^2 N}{d} I\right)^{-1} U^T \\
&= \beta_U^T \left(I + \frac{\eta_{trn}^2 N}{d}\Sigma_{trn}^{-2}\right)^{-1} U^T
\end{aligned}$$

Hence we have finished proving the first part.

For the second part, we note that similar to before, we need to calculate

$$\frac{1}{N_{tst}}\mathbb{E}_{A_{tst}}\left[\|Y_{tst} - W^*(X_{tst} + A_{tst})\|_F^2\right] = \frac{1}{N_{tst}}\|Y_{tst} - W^*X_{tst}\|_F^2 + \frac{\eta_{tst}^2}{d}\|W^*\|_F^2.$$

For the first term recall that $X_{tst} = UL$ and $Y_{tst} = \beta^T X_{tst}$. Hence we have that

$$\begin{aligned}\frac{1}{N_{tst}}\|Y_{tst} - W^* X_{tst}\|_F^2 &= \frac{1}{N_{tst}}\left\|\beta_U^T\left(I - \left(I + \frac{\eta_{trn}^2 N}{d}\Sigma_{trn}^{-2}\right)^{-1}\right)L\right\|_F^2 \\ &= \frac{1}{N_{tst}}\frac{\eta_{trn}^4 N^2}{d^2}\left\|\beta_U^T\left(\Sigma_{trn}^2 + \frac{\eta_{trn}^2 N}{d}\right)^{-1} L\right\|_F^2\end{aligned}$$

For the second term, we have that

$$\begin{aligned}\frac{\eta_{tst}^2}{d}\|W^*\|_F^2 &= \frac{\eta_{tst}^2}{d}\operatorname{Tr}\left(\beta_U^T\left(I + \frac{\eta_{trn}^2 N}{d}\Sigma_{trn}^{-2}\right)^{-2}\beta_U\right) \\ &= \frac{\eta_{tst}^2}{d}\operatorname{Tr}\left(\beta_U\beta_U^T\left(I + \frac{\eta_{trn}^2 N}{d}\Sigma_{trn}^{-2}\right)^{-2}\right)\end{aligned}$$

$\square$

F.6 Proof for Corollary 2, Relative Excess Error

Corollary 2 (Relative Excess Error)**.** *Let* $\|\Sigma_{trn}\|_F^2 = \Omega(N^{1/2+\epsilon})$. *As* $d, N \to \infty$ *with* $d/N \to c$*, the relative excess error tends to* $\frac{c}{1-c}$ *in the underparametrized regime. In the overparametrized regime, when* $\|\Sigma_{trn}\|_F^2 = o(N)$*, it tends to* $\frac{1}{c-1}$ *and to* $\frac{1}{c-1} + k$ *for some constant* k *when* $\|\Sigma_{trn}\|_F^2 = \Theta(N)$.

Proof. Recall from Theorem 4 that the test error for W^* is given by

$$\mathcal{R}(W^*, UL) = \frac{\eta_{trn}^4 N^2}{d^2}\left\|\beta_U^T\left(\Sigma_{trn}^2 + \frac{\eta_{trn}^2 N}{d}I\right)^{-1} L\right\|^2 + \frac{\eta_{tst}^2}{d}Tr\left(\beta_U\beta_U^T\left(I + \frac{\eta_{trn}^2 N}{d}\Sigma_{trn}^{-2}\right)^{-2}\right)$$

We prove this for $c > 1$, the proof for $c < 1$ is analogous and in fact simpler. Notice that when $|\Sigma_{trn}\|_F^2 = \Omega(N^{1/2+\epsilon})$, in both $\mathcal{R}(W_{opt}, X_{tst})$ and $\mathcal{R}(W^*, X_{tst})$, the bias terms are $O(1/d^{1+2\epsilon})$ while the variance terms are $\Theta(1/d)$. In particular, as $d, N \to \infty$, with $d/N \to c$, the limit of the excess risk is given by only considering the variance terms and the estimation errors.

$$
\begin{aligned}
&\lim_{d,N\to\infty, d/N\to c} \frac{\mathcal{R}(W_{opt}, X_{tst}) - \mathcal{R}(W^*, X_{tst})}{\mathcal{R}(W^*, X_{tst})} \\
&= \lim_{d,N\to\infty, d/N\to c} \frac{\frac{\eta_{tst}^2}{d} Tr\left(\beta_U \beta_U^T \left(I + \frac{\eta_{trn}^2 N}{d}\Sigma_{trn}^{-2}\right)^{-2}\right) - \frac{\eta_{tst}^2}{d}\frac{c}{c-1}\operatorname{Tr}(\beta_U\beta_U^T(I+\eta_{trn}^2\Sigma_{trn}^{-2})^{-1})}{\frac{\eta_{tst}^2}{d} Tr\left(\beta_U \beta_U^T \left(I + \frac{\eta_{trn}^2 N}{d}\Sigma_{trn}^{-2}\right)^{-2}\right)} \\
&\quad + \lim_{d,N\to\infty, d/N\to c} \frac{O\left(\frac{\|\Sigma_{trn}\|_F^2}{N^2}\right) + o\left(\frac{1}{N}\right)}{\frac{\eta_{tst}^2}{d} Tr\left(\beta_U \beta_U^T \left(I + \frac{\eta_{trn}^2 N}{d}\Sigma_{trn}^{-2}\right)^{-2}\right)} \\
&= \lim_{d,N\to\infty, d/N\to c} \frac{Tr\left(\beta_U \beta_U^T \left(I + \frac{\eta_{trn}^2}{c}\Sigma_{trn}^{-2}\right)^{-2}\right) - \frac{c}{c-1}\operatorname{Tr}(\beta_U\beta_U^T(I+\eta_{trn}^2\Sigma_{trn}^{-2})^{-1})}{Tr\left(\beta_U \beta_U^T \left(I + \frac{\eta_{trn}^2}{c}\Sigma_{trn}^{-2}\right)^{-2}\right)} \\
&\quad + \lim_{d,N\to\infty, d/N\to c} \frac{O\left(\frac{c\|\Sigma_{trn}\|_F^2}{N}\right) + o\left(c\right)}{\eta_{tst}^2 Tr\left(\beta_U \beta_U^T \left(I + \frac{\eta_{trn}^2}{c}\Sigma_{trn}^{-2}\right)^{-2}\right)} \\
&= \lim_{d,N\to\infty, d/N\to c} \frac{Tr\left(\beta_U\beta_U^T\right) - \frac{c}{c-1}\operatorname{Tr}(\beta_U\beta_U^T)}{Tr\left(\beta_U\beta_U^T\right)} + \lim_{d,N\to\infty, d/N\to c} \frac{O\left(\frac{c\|\Sigma_{trn}\|_F^2}{N}\right) + o(1)}{\eta_{tst}^2 Tr\left(\beta_U\beta_U^T\right)} \\
&= 1 - \frac{c}{c-1} + \lim_{d,N\to\infty, d/N\to c} O\left(\frac{\|\Sigma_{trn}\|_F^2}{N}\right) \\
&= \begin{cases} \frac{1}{c-1} & ; \|\Sigma_{trn}\|_F^2 = o(N) \\ \frac{1}{c-1} + k & ; \|\Sigma_{trn}\|_F^2 = \Theta(N) \end{cases}
\end{aligned}
$$

for some unknown problem-dependent constant k. This establishes the claim for $c > 1$, and the proof for when $c < 1$ is analogous and in fact simpler.

□

F.7 Proofs for Theorem 5, IID Training Data With Isotropic Covariance

Theorem 5 (I.I.D. Training Data With Isotropic Covariance)**.** *Let $c = d/N$ and $c_r = r/N$. Then if $c < 1$*

$$
\begin{aligned}
\mathbb{E}_{X_{trn}}[\mathcal{R}] = \frac{\eta_{trn}^4}{N_{tst}}\|(\Sigma_{trn}^2 c + \eta_{trn}^2 I)^{-1}L\|_F^2 \\
+ \eta_{tst}^2 \frac{r}{d}\frac{1}{1-c}\left(T_1(c_r, \eta_{trn}^2/c) + \frac{1}{\eta_{trn}^2}T_2(c_r, \eta_{trn}^2/c)\right) + o\left(\frac{1}{N}\right)
\end{aligned}
$$

and if $c > 1$

$$
\mathbb{E}_{X_{trn}}[\mathcal{R}] = \frac{\eta_{trn}^4}{N_{tst}}\|(\Sigma_{trn}^2 + \eta_{trn}^2 I)^{-1}L\|_F^2 + \eta_{tst}^2\frac{r}{d}\frac{c}{c-1}T_3(c_r, \eta_{trn}^2) + O\left(\frac{1}{N}\right)
$$

where $T_1(c_r, z) = T_3(cr, z) - zT_2(cr, z)$, and

$$
T_2(c_r, z) = \frac{1 + c_r + zc_r}{2\sqrt{(1 - c_r + c_r z)^2 + 4c_r^2 z}} - \frac{1}{2},\ T_3(c_r, z) = \frac{1}{2} + \frac{1 + zc_r - \sqrt{(1 - c_r + zc_r)^2 + 4c_r^2 z}}{2c_r}.
$$

Proof. Then if X_{trn} is the data matrix, the singular values squared for X_{trn} are the eigenvalues of

$$
X_{trn}^T X_{trn} = Z^T U^T U Z = Z^T Z
$$

Then $Z^T Z$ is a $N \times N$ matrix, and due to the normalization of the variance of the entries, this is a Wishart Matrix. Further, we know that the eigenvalue distribution can be approximated by the Marchenko Pastur distribution with shape parameter r/N [45–50].

Then we have that for the $c < 1$ case, we have the variance is

$$\frac{1}{d}\frac{c}{1-c}\sum_{i=1}^{r}\frac{1}{c^2}\left(\frac{\sigma_i^4}{(\sigma_i^2+\sigma_{trn}^2/c)^2}+\frac{1}{\sigma_{trn}^2}\frac{\sigma_i^2}{(\sigma_i^2+\sigma_{trn}^2)^2}\right)$$

Then we simplify this as the following.

$$\frac{r}{d}\frac{1}{c(1-c)}\left(\mathbb{E}\left[\frac{\sigma_i^4}{(\sigma_i^2+\sigma_{trn}^2/c)^2}\right]+\frac{1}{\sigma_{trn}^2}\mathbb{E}\left[\frac{\sigma_i^2}{(\sigma_i^2+\sigma_{trn}^2)^2}\right]\right)$$

If λ is an eigenvalue of the training data gram matrix, then the variance term of the generalization error has terms of the following form.

$$\frac{\lambda^2}{(\lambda+1/c)^2},\quad\frac{\lambda}{(\lambda+1/c)^2},\quad\frac{\lambda}{\lambda+1}$$

The value of these for the Marchenko Pastur distribution can be found in [44].

$$\mathbb{E}\left[\frac{\lambda}{\lambda+\eta_{trn}^2}\right]=\frac{1}{2}+\frac{1+\eta_{trn}^2c_r-\sqrt{(1-c_r+\eta_{trn}^2c_r)^2+4c_r^2\eta_{trn}^2}}{2c_r}$$

$$\mathbb{E}\left[\frac{\lambda}{(\lambda+\eta_{trn}^2)^2}\right]=\frac{1+c_r+\eta_{trn}^2c_r}{2\sqrt{(1-c_r+c_r\eta_{trn}^2)^2+4c_r^2\eta_{trn}^2}}-\frac{1}{2}+o(1)$$

$$\mathbb{E}\left[\frac{\lambda^2}{(\lambda+\eta_{trn}^2)^2}\right]=\mathbb{E}\left[\frac{\lambda}{\lambda+\eta_{trn}^2}\right]-\eta_{trn}^2\left(\mathbb{E}\left[\frac{\lambda}{(\lambda+\eta_{trn}^2)^2}\right]\right)$$

$c_r = r/N$

The proofs for the rest of the terms are similar. □

F.8 Proofs for Corollary 7, IID Training and Test Data With Isotropic Covariance

Corollary 7 (I.I.D. Train and Tests Data With Isotropic Covariance)**.** *Let $c = d/N$ and $c_r = r/N$. Then if $c < 1$*

$$\mathbb{E}_{X_{trn}}[\mathcal{R}]=\eta_{trn}^4\cdot r\cdot\kappa\cdot T_4(c_r,\eta_{trn}^2/c)+\frac{r}{d}\frac{1}{1-c}\left(T_1(c_r,\eta_{trn}^2/c)+\frac{1}{\eta_{trn}^2}T_2(c_r,\eta_{trn}^2/c)\right)+o\left(\frac{1}{N}\right)$$

and if $c > 1$

$$\mathbb{E}_{X_{trn}}[\mathcal{R}]=\eta_{trn}^4\cdot r\cdot\kappa\cdot T_4(c_r,\eta_{trn}^2)+\frac{r}{d}\frac{c}{c-1}T_3(c_r,\eta_{trn}^2)+O\left(\frac{1}{N}\right)$$

where $T_1(c_r, z) = T_3(c_r, z) - zT_2(c_r, z)$, and

$$T_2(c_r,z)=\frac{1+c_r+zc_r}{2\sqrt{(1-c_r+c_rz)^2+4c_r^2z}}-\frac{1}{2},\ T_3(c_r,z)=\frac{1}{2}+\frac{1+zc_r-\sqrt{(1-c_r+zc_r)^2+4c_r^2z}}{2c_r},$$

$$T_4(c_r,z)=\frac{zc_r^2+c_r^2+zc_r-2c_r+1}{2z^2c_r\sqrt{(1-c_r+c_rz)^2+4c_r^2z}}-\frac{1}{2z^2}\left(1-\frac{1}{c_r}\right).$$

Proof. For the bias, we get

$$\frac{\eta_{trn}^4}{N_{tst}}\frac{1}{c^2}\mathbb{E}\left[\frac{1}{(\sigma_i^2+\eta_{trn}^2/c)^2}\right]\|L\|_F^2$$

The value of these for the Marchenko Pastur distribution can be found in [44].

$$\mathbb{E}\left[\frac{1}{(\lambda+\eta_{trn}^2)^2}\right]=\frac{\eta_{trn}^2c_r^2+c_r^2+\eta_{trn}^2c_r-2c_r+1}{2\eta_{trn}^4c_r\sqrt{4\eta_{trn}^2c_r^2+(1-c_r+\eta_{trn}^2c_r)^2}}+\frac{1}{2\eta_{trn}^4}\left(1-\frac{1}{c_r}\right)$$

□

G Numerical Details

In this section, we include the computational details required to reproduce the data and figures in the paper. The code for the experiments can be found in the following anonymized repository [Link].

G.1 Data

For our transfer learning results, we use real datasets namely CIFAR [39], STL10 [40] and SVHN [41]. We will mostly be working with the training and test split of CIFAR, training split of STL10 and training split of SVHN. We will also use the test split of STL10 for our data augmentation results, refer figure 3 and section G.5, to avoid overlaps between training and test data.

To verify the application of our results to I.I.D. data, we generate datasets from certain distributions, the details of which are presented in the upcoming sections.

The test data is normalized so that each coordinate has mean zero and a standard deviation of 5. This is done before we do any other pre-processing.

G.2 Compute Time

For figures 7, 8, 9 and 6, we use the same training data from CIFAR train split. Thus, we combine our code implementation for these figures. This saves up compute time for mean empirical error since inversion of the matrix $X_{trn} + A_{trn}$, for obtaining W_{opt}, occurs once for each empirical run for all 4 figures. The code was implemented using Google Colab with A100 Nvidia GPU which took approximately 1 hour for the 200 trials for each value of r. Since the results are computed for 4 values of r, the entire experiment was completed within approximately 4 hours.

Figures 2 and 3 took approximately 4 hours each using A100 Nvidia GPU on Google Colab. Figures 4 and 5 were computed together in approximately 40 minutes. Figure 10 took approximately 1 hour to compute. Figure 11 only took around 10 minutes due to less number of N values and only 50 trials. All the above was implemented using A100 GPU on Colab. Figure 7c took approximately 4.5 hours using T4 Nvidia GPU on Google Colab.

G.3 Principal Component Regression

We use four datasets for the set of results obtained through principal component regression namely, CIFAR train split, CIFAR test split, STL10 dataset and SVHN dataset.

G.3.1 In-Subspace

For figure 7a, the test data lies in the same low-dimensional subspace as the training dataset. The experimental setting is as follows.

- Training data, of order $d \times N$, is sampled from flattened CIFAR train split such that $d = 3072$ and N ranges between 1050 and 10500 with an increment of 550 for the results.
- We project our training data over the first r principal components where r refers to the rank and varies as 25, 50, 100 and 150.
- Test datasets, of order $d \times N_{tst}$, are sampled from CIFAR test split, STL10 train split and SVHN train split where $d = 3072$ and $N_{tst} = 2500$.
- We also project these test datasets onto the low-dimensional subspace using the projection matrices.
- For denoising, we generate Gaussian noise matrix A_{trn} with norm $\sqrt{N}$ for the training data and A_{tst} with norm $\sqrt{N_{tst}}$ for the test datasets.

The theoretical error is calculated using the formula in Theorem 1 and the empirical error is the mean squared error.

G.3.2 Out-of-Subspace

Next, we test our formulas for test datasets which lie outside the training distribution space.

Small α We detail the numerical setup required to generate figure 7b.

- Training data, of order $d \times N$, is sampled from flattened CIFAR train split such that $d = 3072$ and N ranges between 1050 and 10500 with an increment of 550 for the results.
- We project our training data over the first r principal components where r refers to the rank and varies as 25, 50, 100 and 150.
- Test datasets, of order $d \times N_{tst}$, are sampled from CIFAR test split, STL10 train split and SVHN train split where $d = 3072$ and $N_{tst} = 2500$.
- We project these test datasets onto the low-dimensional subspace using the projection matrices.
- We add a small amount of full-dimensional Gaussian noise to the projected datasets to generate out-of-subspace datasets with small α. Here, we consider the case where $\alpha = 0.1$.
- For denoising, we generate Gaussian noise matrix A_{trn} with norm $\sqrt{N}$ for the training data and A_{tst} with norm $\sqrt{N_{tst}}$ for the test datasets.

The empirical error shown in figure 7b is the square root of the mean squared error. The theoretical bounds on the error are calculated using Theorem 3.

Large α. For figure 6, the experimental setup is as follows.

- Training data, of order $d \times N$, is sampled from flattened CIFAR train split such that $d = 3072$ and N ranges between 1050 and 10500 with an increment of 550 for the results.
- We project our training data over the first r principal components where r refers to the rank and varies as 25, 50, 100 and 150.
- Test datasets, of order $d \times N_{tst}$, are sampled from CIFAR test split, STL10 train split and SVHN train split where $d = 3072$ and $N_{tst} = 2500$.
- We do not project these test datasets onto the low-dimensional subspace. We retain their high dimensions. The values of α for different values of r are provided in figure 6.
- For denoising, we generate Gaussian noise matrix A_{trn} with norm $\sqrt{N}$ for the training data and A_{tst} with norm $\sqrt{N_{tst}}$ for the test datasets.

G.4 Linear Regression

To consider the linear regression case for figure 8,

- Training data, of order $d \times N$, is sampled from flattened CIFAR train split such that $d = 3072$ and N ranges between 1050 and 10500 with an increment of 550 for the results.
- We project our training data over the first r principal components where r refers to the rank and varies as 25, 50, 100 and 150.
- Gaussian noise matrix with norm $\sqrt{N}$ is added to the training data.
- We generate normally-distributed β_{opt} of order $d \times 1$ with norm 1. The learned estimator is computed as $\beta^T = \beta_{opt}^T W$ where W is the minimum norm solution to the least squares denoising problem. For theoretical error, we compute $\hat{\beta}^T = \beta_{opt} U$.
- Test datasets, of order $d \times N_{tst}$, are sampled from CIFAR test split, STL10 train split and SVHN train split where $d = 3072$ and $N_{tst} = 2500$.
- We also project these test datasets onto the low-dimensional subspace using the projection matrices.
- Gaussian noise matrix with norm $\sqrt{N_{tst}}$ is added to the test datasets.
- Finally, the test datasets, X_{tst}, are replaced with $\beta^T X_{tst}$ to compute the error for the linear regression problem.

G.5 Data Augmentation

To emphasize the application of our results to non-I.I.D. data, we consider two cases of data augmentation to our training data.

G.5.1 Without Independence

The experimental setting to obtain the empirical generalization error is as follows.

- We sample 1000 images from the CIFAR train split as the first batch of our training data. For experimental results
- We augment the above batch with the same batch to vary N between 1000 and 6000 with an increment of 1000. We project the dataset onto its first r principal components where $r = 25, 50, 100$ and 150.
- We add gaussian noise with norm $\sqrt{N}$ to the training data as before. Note that the noise on augmented batches would be independent of the noise in the original batch. This is the only assumption required for our result.
- Test datasets, of order $d \times N_{tst}$, sampled from CIFAR test split, STL10 train split and SVHN train split where $d = 3072$ and $N_{tst} = 2500$ are also projected onto the low-dimensional subspace.

We calculate the theoretical generalization error for more values of c to obtain smoother curves. Note that the left singular vectors i.e., the columns of matrix U, do not change when we augment our training batches. We utilize this to speed-up our computation for theoretical curves.

- We sample 1000 images from the CIFAR train split as the first batch of our training data.
- We obtain the projection matrix $P = UU^T$ and the matrix $L = U^T X_{tst}$ from the SVD of the first batch itself.
- The generalization error is computed from the formula in Theorem 1 for values of N between 1000 and 6000 with an increment of 50.
- We scale the singular values by a factor of $N/1000$ to account for the augmenting.

G.5.2 Without Identicality

To generate figure 3,

- We use training data, of order $d \times N$, such that $d = 3072$ and N ranges between 1050 and 10500 with an increment of 550 for the results.
- We use $N/2$ images from the CIFAR training split and $N/2$ images from the STL10 training split concatenated together for our training data.
- We project our training data over the first r principal components where r refers to the rank and varies as 25, 50, 100 and 150.
- Test datasets, of order $d \times N_{tst}$, are sampled from CIFAR test split, STL10 test split and SVHN train split where $d = 3072$ and $N_{tst} = 2500$. This is done to avoid any overlaps between training and test data.
- We also project these test datasets onto the low-dimensional subspace using the projection matrices.
- For denoising, we generate Gaussian noise matrix A_{trn} with norm $\sqrt{N}$ for the training data and A_{tst} with norm $\sqrt{N_{tst}}$ for the test datasets.

G.6 I.I.D. Data

We also perform experiments to verify our results in cases where training and test datasets are I.I.D. The numerical details for those experiments are presented in this section.

G.6.1 I.I.D. Test Data

To generate figure 9,

- Training data, of order $d \times N$, is sampled from flattened CIFAR train split such that $d = 3072$ and N ranges between 1050 and 10500 with an increment of 550 for the results.
- We project our training data over the first r principal components where r refers to the rank and varies as 25, 50, 100 and 150.

- We generate L from Gaussian distribution of norm $\sqrt{N_{tst}}$ where $N_{tst} = 2500$.
- We obtain our I.I.D. test data of order $d \times N_{tst}$ as $X_{tst} = UL$ where U contains the left singular vectors of the projected training data.
- For denoising, we generate Gaussian noise matrix A_{trn} with norm $\sqrt{N}$ for the training data and A_{tst} with norm $\sqrt{N_{tst}}$ for the test datasets.

G.6.2 I.I.D. Train Data

To generate figure 10,

- We generate the left singular matrix U from the SVD of a Gaussian matrix of order $d \times r$ where $M = 3072$ and $r = 50$.
- We generate the training matrix $X_{trn} = UZ$ where Z is of order $r \times N$ such that each column is normally distributed with mean 0 and variance $1/r$.
- Here, N varies from 1050 to 10500 with an increment of 550.
- Test datasets, of order $d \times N_{tst}$, are sampled from CIFAR test split, STL10 train split and SVHN train split where $d = 3072$ and $N_{tst} = 2500$.
- We also project these test datasets onto the r-dimensional subspace using projection matrices.
- For denoising, we generate Gaussian noise matrix A_{trn} with norm $\sqrt{N}$ for the training data and A_{tst} with norm $\sqrt{N_{tst}}$ for the test datasets.

G.6.3 I.I.D Train and Test Data

To generate figure 11,

- We generate the left singular matrix U from the SVD of a Gaussian matrix of order $d \times r$ where $M = 3072$ and $r = 50$.
- We generate the training matrix $X_{trn} = UZ$ where Z is of order $r \times N$ such that each column is normally distributed with mean 0 and variance $1/r$.
- Here, N varies from 500 to 6010 with an increment of 550 for the empirical markers and with an increment of 55 for theoretical values on the solid curve.
- We generate L from Gaussian distribution of norm $\sqrt{N_{tst}}$ where $N_{tst} = 5000$.
- We obtain our I.I.D. test data of order $d \times N_{tst}$ as $X_{tst} = UL$ where U contains the left singular vectors of the projected training data.
- For denoising, we generate Gaussian noise matrix A_{trn} with norm $\sqrt{N}$ for the training data and A_{tst} with norm $\sqrt{N_{tst}}$ for the test datasets.

G.7 Full Dimensional Denoising

To generate figure 7c,

- Training data, of order $d \times N$, is sampled from flattened CIFAR train split such that $d = 3072$ and N ranges between 1050 and 10500 with an increment of 550 for the results.
- We project our training data over the first r principal components where r is the minimum of d and N. This implies that the data is full dimensional.
- Test datasets, of order $d \times N_{tst}$, are sampled from CIFAR test split, STL10 train split and SVHN train split where $d = 3072$ and $N_{tst} = 2500$.
- We also project these test datasets onto the low-dimensional subspace using the projection matrices.
- For denoising, we generate Gaussian noise matrix A_{trn} with norm $\sqrt{N}$ for the training data and A_{tst} with norm $\sqrt{N_{tst}}$ for the test datasets.

G.8 Optimal η_{trn}

To generate figures 4 and 5,

- Training data, of order $d \times N$, is sampled from flattened CIFAR train split such that $d = 3072$ and N ranges between 500 and 5500 as {500, 750, 1000, 1250, 1500, 1750, 2000, 2250, 2500, 2600, 2700, 2800, 2900, 3000, 3020, 3130, 3200, 3300, 3400, 3500, 3750, 4000, 4250, 4500, 4750, 5000, 5250, 5500}.
- We project our training data over the first r principal components where $r = 50$.
- Test datasets, of order $d \times N_{tst}$, are the training dataset with new noise and sampled from CIFAR test split, STL10 train split and SVHN train split where $d = 3072$ and $N_{tst} = N$.
- We compute generalization error for 2000 η_{trn} values ranging from 1/3.5 to 100 for each N from our formula in Theorem 1.
- We report the optimal η_{trn} found to minimise the generalization error in figure 4 and the optimal generalization error in figure 5.