Risk Phase Transitions in Spiked Regression: Alignment Driven Benign and Catastrophic Overfitting

author names withheld

Under Review for the Workshop on High-dimensional Learning Dynamics, 2024

Abstract

This paper analyzes the generalization error of minimum-norm interpolating solutions in linear regression using spiked covariance data models. The paper characterizes how varying spike strengths and target-spike alignments affect risk, especially in overparameterized settings. The study presents an exact expression for the generalization error, leading to a comprehensive classification of benign, tempered, and catastrophic overfitting regimes based on spike strength, the aspect ratio $c = d/n$ (particularly as $c \to \infty$), and target alignment. Notably, in well-specified aligned problems, increasing spike strength can surprisingly induce catastrophic overfitting before achieving benign overfitting. The paper also reveals that target-spike alignment is not always advantageous, identifying specific, sometimes counterintuitive, conditions for its benefit or detriment. Alignment with the spike being detrimental is empirically demonstrated to persist in nonlinear models.

Keywords: Generalization, Random Matrix Theory, Spiked Covariance, Benign/Tempered/Catastrophic Overfitting

1. Introduction

Understanding the generalization error of overparameterized models is a central challenge in modern machine learning. Phenomena such as double descent [7, 16] and benign overfitting [6, 21, 32] have spurred research underscoring the critical role of the data's spectral structure [6, 15, 16, 18, 24, 31–33]. The spiked covariance model is one commonly considered spectral structure [12]. In this model, the data matrix $X = Z + A \in \mathbb{R}^{d \times n}$, comprising n data points in $\mathbb{R}^d$, is decomposed into a rank-one signal component ("spike") Z and an isotropic noise component ("bulk") A. Spiked covariance models emerge naturally in practice, for instance, in the features learned by neural networks during training [1, 2, 13, 14, 23, 26, 34]. While recent studies have examined benign overfitting in spiked models [2, 18], they lack a systematic taxonomy spanning spike strength, target–spike alignment, model misspecification, and train–test covariate shift. This paper closes the gap for linear regression.

This work explores how general spike sizes and target alignments affect generalization error in least squares linear regression. We address two fundamental questions:

- **Q1:** For a fixed aspect ratio $c = d/n$, in asympototic proportional regime under what conditions does alignment of the target signal with the data spike improve or impair generalization?
- **Q2:** In the high-dimensional limit where $c \to \infty$, when do we observe benign, tempered, or catastrophic overfitting regimes?

Contributions We present precise characterization of the generalization performance of minimum-norm interpolating solutions in linear regression. Our exact risk decomposition pinpoints conditions for transitions between benign and catastrophic overfitting. This reveals alignment-dependent phenomena obscured by isotropic theories, clarifying how signal structure, data scaling, and overparameterization shape generalization.

Notation The subscript on $o, O, \omega, \Omega, \Theta$ will denote which quantity is being sent to infinity.

2. Problem Setting

We study the generalization of minimum-norm interpolators in high-dimensional linear regression. Using a spiked covariance data model, we quantify how spike strength and alignment influence generalization and the emergence of benign, tempered, or catastrophic overfitting.

Data Model. We consider a data matrix $X = Z + A \in \mathbb{R}^{d\times n}$ with *signal component* Z and *isotropic noise component* A that satisfy the following assumptions. Specifically, we shall that the population feature covariance is $\Sigma = \theta^2 uu^T + \tau^2 I_d$, modeling a rank-one perturbation of isotropic noise.

Assumption 1 (Signal) *Let $u \in \mathbb{R}^d$ be a fixed unit vector representing the spike direction. Then*

$$Z = \theta \cdot uv^T, \tag{1}$$

where $\theta > 0$ controls the spike strength, and the vector $v \in \mathbb{R}^n$ has i.i.d. standard normal entries.

Assumption 2 (Noise) *The entries of A have zero mean and variance τ^2. The matrix A satisfies:*

- *Its entries are uncorrelated and possess finite fourth moments.*
- *Its distribution is invariant under left and right orthogonal transformations.*
- *The empirical spectral distribution of $\frac{1}{\tau^2 d}AA^T$ converges to the Marchenko–Pastur law as $n, d \to \infty$ with $d/n \to c \in (0, \infty)$.*

Spike Strength Normalizations. We consider two key scaling regimes for the spike strength relative to the bulk noise. These lead to distinct generalization behaviors.

1) **Operator Norm Scaling ($\theta^2 = \gamma\tau^2$):** Here γ tunes the spike strength θ^2 relative to the noise variance τ^2. When $\gamma = (1+\sqrt{c})^2$, the spectral norm of the signal component Z is comparable to that of the noise component A. If $\gamma > (1+\sqrt{c})^2$, the spike emerges as an isolated eigenvalue beyond the bulk spectrum established by A, a phenomenon known as the Baik–Ben Arous–Péché (BBP) transition [5]. This scaling reflects spikes in learned neural network features [1, 26].
2) **Frobenius Norm Scaling ($\theta^2 = d\tau^2$):** Here $\theta^2 = d\tau^2$ matches expected signal and noise Frobenius norms ($\mathbb{E}[\|Z\|_F^2] = \mathbb{E}[\|A\|_F^2]$) and the spike has macroscopic proportion of the energy. Such strong signals can lead to improved sample complexity, potentially overcoming limitations observed in purely isotropic models [2, 24].

Target Model. Given $x_i = z_i + a_i$, the targets y are obtained as follows:

$$y_i = \alpha_Z z_i^T \beta_* + \alpha_A a_i^T \beta_* + \varepsilon_i, \tag{2}$$

where $\beta_* \in \mathbb{R}^d$ is the true underlying parameter vector with $\|\beta_*\| = \Theta(1)$. The terms z_i and a_i are the i-th columns of Z and A respectively. The observation noise ε_i are i.i.d. with $\mathbb{E}[\varepsilon_i] = 0$, $\mathbb{E}[\varepsilon_i^2] = \tau_\varepsilon^2$. The coefficients $\alpha_Z, \alpha_A \in \mathbb{R}$ control the target's dependence on the signal and noise components. If $\alpha_Z \neq \alpha_A$, the true data generating process for y differentially weights components of x_i, causing model misspecification for estimators unaware of this structure.

Generalization Risk. We study the minimum-norm interpolating ordinary least squares estimator: $\beta_{int} = X^\dagger y,$ with $\hat{y} = (\tilde{z} + \tilde{a})\beta_{int}$, where $X^\dagger$ denotes the Moore–Penrose pseudoinverse. Given a new test data points $(\tilde{x}, \tilde{y})$, where $\tilde{x} = \tilde{z} + \tilde{a}$ and targets $\tilde{y} = \tilde{\alpha}_Z \tilde{z}^T \beta_* + \tilde{\alpha}_A \tilde{a}^T \beta_* + \tilde{\varepsilon}$ with potentially with different coefficients $\tilde{\alpha}_Z, \tilde{\alpha}_A$, the generalization risk is defined as the expected squared prediction error:

$$\mathcal{R}(\beta_{int}) = \mathbb{E}_{X,\varepsilon,\{\tilde{x},\tilde{\varepsilon}\}} \left[(\tilde{y} - \hat{y})^2 \right]. \tag{3}$$

The expectation is over the training data (X, ε) and the test data realization $(\{\tilde{x}, \tilde{\varepsilon}\})$. We shall denote the asymptotic excess risk in the proportional regime as follows:

$$\mathcal{R}_c = \lim_{n,d\to\infty,\, d/n\to c} \mathcal{R}(\beta_{int}) - \tau_\varepsilon^2.$$

Remark 1 (Generalizing Prior Work) *This problem formulation encompasses several existing models as special cases. For instance, isotropic regression settings studied in [16] are recovered by setting $\theta = 0$ (no spike) and $\alpha_Z = 0$. Spike recovery models, such as in [31], correspond to specific choices like $\tau^2 = 1/d$, $\tau_\varepsilon^2 = 0$, and $\alpha_A = 0$. Our generalized setup allows for a nuanced investigation of the interplay between signal structure, target alignment, and overparameterization.*

Quantifying the Benefit of Alignment. A key aspect of our investigation is to determine when the alignment of the true parameter vector β_* with the data's principal spike direction u is beneficial for generalization. We define alignment as *beneficial* if the generalization risk $\mathcal{R}(\beta_{int})$ (or $\mathcal{R}_c$), is monotonically decreasing as a function of $(\beta_*^T u)^2 \in [0, 1]$. Conversely, alignment is *detrimental* if the risk is a monotonically increasing function of $(\beta_*^T u)^2$.

Characterizing Overfitting Regimes. Following [6, 21], we classify the asymptotic behavior of the excess risk, $\mathcal{R}_c$ as $c \to \infty$ as benign, tempered or catastrophic. We say the overfitting is **benign** if $\lim_{c\to\infty} \mathcal{R}_c$ is zero, **tempered** if this limit is positive and finite, **catastrophic** if this limit is infinite.

3. Theoretical Results

Our core theoretical contribution is a precise analytical formula for excess risk in the spiked covariance model. This result relies on Assumption 3, which encompasses both the operator norm scaling $(\theta^2 = \gamma\tau^2)$ and Frobenius norm scaling $(\theta^2 = d\tau^2)$ regimes. We develop our general risk theorem by analyzing progressively complex scenarios. Specifically, our forthcoming theorems provide specific conditions for benign, tempered, or catastrophic overfitting (as $c \to \infty$), and determine when, for finite c, alignment of β_* with spike u is beneficial or detrimental. In the main text, we only present the well specified case, the rest, including the mis-specified case, the general theorem, and extension to nonlinear cases, can be found in the appendix.

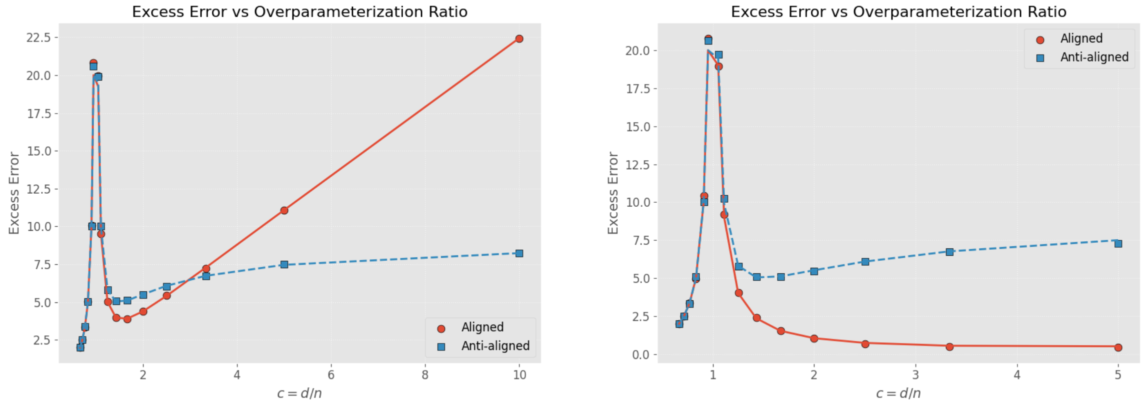
Assumption 3 (Scaling) *As $n, d \to \infty$ with $d/n \to c \in (0, \infty)$, we assume that θ^2 and τ^2 satisfy $\Omega(\tau^2) \le \theta^2 \le O(d\tau^2)$ and $\Omega\left(\frac{1}{d}\right) \le \tau^2 \le O(1)$.*

3.1. Well Specified Problem

We begin by analyzing the well-specified case, where the target y is a direct linear function of the observed covariates $X = Z + A$. This scenario is realized by setting:

$$\alpha_Z = \alpha_A = \tilde{\alpha}_Z = \tilde{\alpha}_A = \alpha > 0.$$

Consequently, $y_i = \alpha x_i^T \beta_* + \varepsilon_i$, and the model is properly specified.



(a) Operator norm scaling ($\theta^2 = c\tau^2$). Alignment can initially improve generalization for small c, but may lead to catastrophic overfitting as $c \to \infty$. Anti-alignment typically yields tempered risk.

(b) Equal Frobenius norm scaling ($\theta^2 = d\tau^2$). Alignment leads to benign overfitting, while anti-alignment results in tempered risk.

Figure 1: Excess error vs. overparameterization ratio $c = d/n$ in the well-specified case. Each plot shows the risk for aligned and anti-aligned targets under different spike scaling regimes. **The scatter plots are empirically obtained and the lines are theory.**

Theorem 2 (Well-Specified Risk) *Given data (X, y) and $(\tilde{X}, \tilde{y})$ generated according to Assumptions 1 (Signal), 2 (Noise), Equation 2 (Target Model), and Assumption 3 (Scaling). If the well-specification condition $\alpha_Z = \alpha_A = \tilde{\alpha}_Z = \tilde{\alpha}_A = \alpha > 0$ holds, the asymptotic excess risk $\mathcal{R}_c$ is given by:*

$$\mathcal{R}_c = \begin{cases} \tau_\varepsilon^2 \frac{c}{1-c} & \text{if } c < 1 \\ \tau_\varepsilon^2 \frac{1}{c-1} + \alpha^2\tau^2\left(1 - \frac{1}{c}\right)\left[\|\beta_*\|^2 + (\beta_*^T u)^2 \frac{\theta^2\tau^2c^2 - 2\theta^2\tau^2 c - \theta^4}{(\theta^2+\tau^2 c)^2}\right] & \text{if } c > 1 \end{cases}$$

where u is the unit vector defining the spike direction.

Remark 3 *If $\theta^2 = \gamma\tau^2$ with $\gamma = o(1)$ (a regime not allowed by Assumption 3 but useful for sanity checks), the coefficient of $(\beta_*^T u)^2$ vanishes, the risk expression aligns with that of isotropic models, such as in [16, Theorem 1].*

Operator Norm Scaling ($\theta^2 = \gamma\tau^2$). In this regime, where the spike strength θ^2 is $\gamma\tau^2$, the excess risk for $c > 1$ becomes:

$$\mathcal{R}_c = \alpha^2\tau^2\left(1 - \frac{1}{c}\right)\left(\|\beta_*\|^2 + \frac{\gamma c^2 - 2\gamma c - \gamma^2}{(\gamma + c)^2}(\beta_*^T u)^2\right) + \tau_\varepsilon^2 \frac{1}{c-1}.$$

The formula shows that alignment with the spike direction u is beneficial if and only if the coefficient of $(\beta_*^T u)^2$ is negative, which occurs when $\gamma > c(c-2)$. We consider different scalings for γ.

Case 1: $\gamma = \Theta_c(1)$ (constant with respect to c). The condition for beneficial alignment, $\gamma > c(c-2)$, interacts intricately with the BBP phase transition condition, $\gamma > (1 + \sqrt{c})^2$. Let $c_* \approx 4.212$ be the unique solution to $c(c-2) = (1 + \sqrt{c})^2$ for $c > 1$.

- For $1 < c < c_*$: Here, $c(c-2) < (1+\sqrt{c})^2$. If γ is in the range $c(c-2) < \gamma < (1+\sqrt{c})^2$, alignment is beneficial even though the BBP transition has *not* occurred (the spike is not resolved from the bulk).
- For $c > c_*$: Here, $c(c-2) > (1+\sqrt{c})^2$. For alignment to be beneficial ($\gamma > c(c-2)$), the BBP transition must have occurred (as $\gamma > c(c-2) \implies \gamma > (1+\sqrt{c})^2$). However, the BBP transition occurring is not sufficient for beneficial alignment. If $(1+\sqrt{c})^2 < \gamma < c(c-2)$, the BBP transition occurs, yet alignment is detrimental.

Regarding the type of overfitting as $c \to \infty$ (while γ remains constant):

$$\lim_{c\to\infty} \mathcal{R}_c = \alpha^2 \tau^2 \left(\|\beta_*\|^2 + \gamma(\beta_*^T u)^2 \right).$$

Since this limit is a positive constant, we consistently observe *tempered overfitting* when $\gamma = \Theta_c(1)$.

Case 2: $\gamma = \omega_c(1)$ (γ grows with c). The behavior depends on the growth rate of γ relative to c. The limit of the excess risk for $\beta_*^T u \neq 0$ as $c \to \infty$ is:

$$\lim_{c\to\infty} \mathcal{R}_c = \alpha^2\tau^2 \cdot \begin{cases} \infty & \text{if } \omega_c(1) \leq \gamma \leq o_c(c) \\ \|\beta_*\|^2 + (\frac{1}{\phi} - 1)(\beta_*^T u)^2 & \text{if } \gamma = \phi c^2 \text{ for const. } \phi > 0 \\ \|\beta_*\|^2 - (\beta_*^T u)^2 & \text{if } \gamma = \omega_c(c^2) \end{cases}$$

Surprisingly, while $\gamma = \Theta_c(1)$ gives tempered overfitting, increasing spike strength to $\omega_c(1) \leq \gamma \leq o_c(c^2)$ results in *catastrophic overfitting*, even though morally, this version of the problem has less noise. Additionally, we see that this catastrophic overfitting is not present in the anti-aligned ($\beta_*^T u$) case. More, aligned with intuition, we see that further increasing the size of the spike improves the generalization performance. Specifically, we get *tempered overfitting* if $\gamma = \phi c^2$ and *benign overfitting* if $\gamma = \omega_c(c^2)$, $\beta_* \parallel u$ and $\|\beta_*\| = 1$.

For $\gamma = c$, the $(\beta_*^T u)^2$ coefficient is $(c-3)/4$. Thus, for $1 < c < 3$, alignment is beneficial and for $c > 3$, alignment becomes detrimental. As $c \to \infty$, if $\beta_* \parallel u$, the excess risk grows approximately as $\alpha^2\tau^2\frac{c}{4}(\beta_*^T u)^2$, indicating *catastrophic overfitting*. In contrast, if $\beta_* \perp u$, the excess risk grows like $\alpha^2\tau^2(1-1/c)\|\beta_*\|^2$, leading to *tempered overfitting*. This transition is illustrated in Figure 1a.

Frobenius Norm Scaling ($\theta^2 = d\tau^2$). The excess risk for $c > 1$ simplifies to:

$$\mathcal{R}_{c>1} = \alpha^2\tau^2 \left(1 - \frac{1}{c}\right) \left(\|\beta_*\|^2 - (\beta_*^T u)^2\right) + \tau_\varepsilon^2 \frac{1}{c-1}.$$

We have a few observations. First, if $\beta_* \parallel u$ and $\|\beta_*\| = 1$, the excess risk $\mathcal{R}_c$ tends to 0 as $c \to \infty$ (*benign overfitting*). Second, if β_* is not perfectly aligned with u, $\mathcal{R}_c \to \alpha^2\tau^2(\|\beta_*\|^2 - (\beta_*^T u)^2) > 0$ as $c \to \infty$ (*tempered overfitting*). Finally, the coefficient of $(\beta_*^T u)^2$ in the risk formula is negative. Hence, in contrast with the operator norm regime, *alignment is always beneficial* in this regime for $c > 1$, and we visualize these behaviors in Figure 1b.

Takeaways for the Well-Specified Case. Spike scaling profoundly impacts overfitting, especially with target alignment. For aligned targets, increasing spike strength can drive transitions from tempered → catastrophic → tempered → benign overfitting, while anti-alignment ($\beta_* \perp u$) can mitigate catastrophic overfitting. Additionally, alignment with the spike is not always beneficial.

References

[1] Jimmy Ba, Murat A Erdogdu, Taiji Suzuki, Zhichao Wang, Denny Wu, and Greg Yang. High-dimensional asymptotics of feature learning: How one gradient step improves the representation. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=akddwRG6EGi.

[2] Jimmy Ba, Murat A Erdogdu, Taiji Suzuki, Zhichao Wang, and Denny Wu. Learning in the presence of low-dimensional structure: A spiked random matrix perspective. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=HlIAoCHDWW.

[3] Zhidong Bai and Wang Zhou. Large sample covariance matrices without independence structures in columns. 2008.

[4] Jinho Baik and Jack W Silverstein. Eigenvalues of large sample covariance matrices of spiked population models. *Journal of multivariate analysis*, 97(6):1382–1408, 2006.

[5] Jinho Baik, Gérard Ben Arous, and Sandrine Péché. Phase transition of the largest eigenvalue for nonnull complex sample covariance matrices. 2005.

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

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

[8] Florent Benaych-Georges and Raj Rao Nadakuditi. The eigenvalues and eigenvectors of finite, low rank perturbations of large random matrices. *Advances in Mathematics*, 227(1):494–521, 2011.

[9] Florent Benaych-Georges and Raj Rao Nadakuditi. The singular values and vectors of low rank perturbations of large rectangular random matrices. *Journal of Multivariate Analysis*, 111: 120–135, 2012.

[10] Yuan Cao, Quanquan Gu, and Mikhail Belkin. Risk bounds for over-parameterized maximum margin classification on sub-gaussian mixtures. *Advances in Neural Information Processing Systems*, 34:8407–8418, 2021.

[11] Niladri S. Chatterji and Philip M. Long. Finite-sample analysis of interpolating linear classifiers in the overparameterized regime. *Journal of Machine Learning Research*, 22(129):1–30, 2021. URL http://jmlr.org/papers/v22/20-974.html.

[12] Romain Couillet and Zhenyu Liao. *Random Matrix Methods for Machine Learning*. Cambridge University Press, 2022. doi: 10.1017/9781009128490. https://zhenyu-liao.github.io/book/.

[13] Alex Damian, Jason D. Lee, and Mahdi Soltanolkotabi. Neural networks can learn representations with gradient descent, 2022. URL https://arxiv.org/abs/2206.15144.

[14] Yatin Dandi, Florent Krzakala, Bruno Loureiro, Luca Pesce, and Ludovic Stephan. How two-layer neural networks learn, one (giant) step at a time. *Journal of Machine Learning Research*, 25(349):1–65, 2024. URL http://jmlr.org/papers/v25/23-1543.html.

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

[16] Trevor Hastie, Andrea Montanari, Saharon Rosset, and Ryan J Tibshirani. Surprises in high-dimensional ridgeless least squares interpolation. *Annals of statistics*, 50(2):949, 2022.

[17] Kedar Karhadkar, Erin George, Michael Murray, Guido F Montufar, and Deanna Needell. Benign overfitting in leaky relu networks with moderate input dimension. *Advances in Neural Information Processing Systems*, 37:36634–36682, 2024.

[18] Chinmaya Kausik, Kashvi Srivastava, and Rishi Sonthalia. Double descent and overfitting under noisy inputs and distribution shift for linear denoisers. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=HxfqTdLIRF.

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

[20] Tengyuan Liang and Alexander Rakhlin. Just interpolate. *The Annals of Statistics*, 48(3):1329–1347, 2020.

[21] Neil Mallinar, James B Simon, Amirhesam Abedsoltan, Parthe Pandit, Mikhail Belkin, and Preetum Nakkiran. Benign, tempered, or catastrophic: A taxonomy of overfitting. *arXiv preprint arXiv:2207.06569*, 2022.

[22] V A Marchenko and Leonid A. Pastur. Distribution of eigenvalues for some sets of random matrices. *Mathematics of The Ussr-sbornik*, 1:457–483, 1967.

[23] Charles H. Martin and Michael W. Mahoney. Implicit self-regularization in deep neural networks: Evidence from random matrix theory and implications for learning. *Journal of Machine Learning Research*, 22(165):1–73, 2021. URL http://jmlr.org/papers/v22/20-410.html.

[24] Song Mei, Theodor Misiakiewicz, and Andrea Montanari. Generalization error of random feature and kernel methods: hypercontractivity and kernel matrix concentration. *Applied and Computational Harmonic Analysis*, 59:3–84, 2022.

[25] Carl D. Meyer, Jr. Generalized Inversion of Modified Matrices. *SIAM Journal on Applied Mathematics*, 1973.

[26] Behrad Moniri, Donghwan Lee, Hamed Hassani, and Edgar Dobriban. A theory of non-linear feature learning with one gradient step in two-layer neural networks. *arXiv preprint arXiv:2310.07891*, 2023.

[27] Alireza Mousavi-Hosseini, Denny Wu, Taiji Suzuki, and Murat A Erdogdu. Gradient-based feature learning under structured data. *Advances in Neural Information Processing Systems*, 36: 71449–71485, 2023.

[28] Vidya Muthukumar, Kailas Vodrahalli, Vignesh Subramanian, and Anant Sahai. Harmless interpolation of noisy data in regression. *IEEE Journal on Selected Areas in Information Theory*, 1(1):67–83, 2020.

[29] Raj Rao Nadakuditi. Optshrink: An algorithm for improved low-rank signal matrix denoising by optimal, data-driven singular value shrinkage. *IEEE Transactions on Information Theory*, 60(5):3002–3018, 2014.

[30] Ohad Shamir. The implicit bias of benign overfitting. In Po-Ling Loh and Maxim Raginsky, editors, *Proceedings of Thirty Fifth Conference on Learning Theory*, volume 178 of *Proceedings of Machine Learning Research*, pages 448–478. PMLR, 02–05 Jul 2022. URL https://proceedings.mlr.press/v178/shamir22a.html.

[31] 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. ISSN 2835-8856. URL https://openreview.net/forum?id=FdMWtpVT1I.

[32] Alexander Tsigler and Peter L Bartlett. Benign overfitting in ridge regression. *Journal of Machine Learning Research*, 24(123):1–76, 2023.

[33] Yutong Wang, Rishi Sonthalia, and Wei Hu. Near-interpolators: Rapid norm growth and the trade-off between interpolation and generalization. In *International Conference on Artificial Intelligence and Statistics*, pages 4483–4491. PMLR, 2024.

[34] Zhichao Wang, Denny Wu, and Zhou Fan. Nonlinear spiked covariance matrices and signal propagation in deep neural networks. *arXiv preprint arXiv:2402.10127*, 2024.

[35] Denny Wu and Ji Xu. On the optimal weighted ℓ_2 regularization in overparameterized linear regression. *Advances in Neural Information Processing Systems*, 33:10112–10123, 2020.

Contents

Appendix A. Detailed Contributions & Related Works

Our primary contributions are as follows:

- **Precise Risk Characterization:** We derive an exact generalization error decomposition (Theorem 7) into interpretable bias, variance, data noise, and alignment terms.
- **Comprehensive Categorization of Overfitting Regimes:** We precisely classify benign, tempered, or catastrophic overfitting regimes based on spike strength, overparameterization ($c = d/n$), and target alignment (Table 1). Surprisingly, for well-specified aligned problems, increasing spike strength can induce catastrophic overfitting before achieving benign overfitting. Misspecified problems show distinct transitions, often precluding benign overfitting.
- **Conditions for Beneficial Alignment:** Challenging conventional wisdom, we show spike alignment is not always beneficial and depends on spike strength meeting critical thresholds (Table 2). For misspecified problems, beneficial alignment requires α_Z/α_A in a specific, non-trivial range. Counterintuitively, very strong spike dependence (α_Z/α_A) can render alignment detrimental.
- **Empirical Validation:** [1] Empirical validation confirms our theoretical phenomena, including surprising negative alignment impacts, persist in nonlinear models, underscoring broader relevance.

Benign Overfitting in Linear Regression. Significant research has explored benign overfitting in linear regression [6, 10, 11, 17, 19–21, 28, 30, 32, 35]. Many studies assume a uniformly bounded largest covariance eigenvalue or lack precise characterizations of its interplay with target alignment and generalization. *Our work allows this eigenvalue to grow, offering precise performance characterizations based on this growth and alignment.* While Kausik et al. [18] considers spiked models, their focus is on noiseless, well-specified scenarios with specific spike scaling. *Our analysis is broader, encompassing observation noise, misspecification, and general spike scaling.*

Many prior works[17, 30, 32] on benign overfitting with low-rank signals plus isotropic noise require near-orthogonality between signal and noise, sometimes imposing strong conditions like $d = \Omega(n^2 \log n)$. *We instead consider the proportional regime* $d/n \to c = \Theta(1)$*, subsequently examining* $c \to \infty$. This setting is morally similar to allowing $d = \omega(n)$ and aligns with approaches like [17] which, for classification, shows misclassification probability can be upper bounded by $Ce^{-d/n}$, vanishing as $d/n \to \infty$.

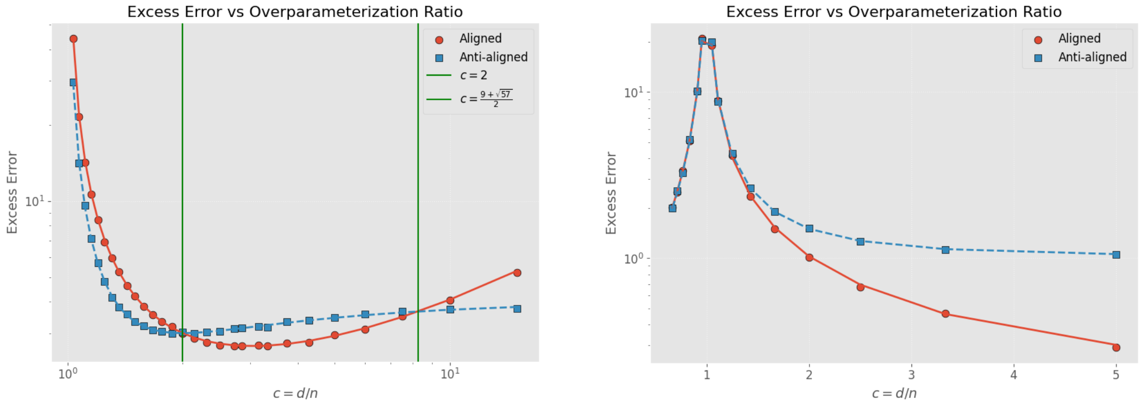
Generalization Error with Spiked Covariance. While recovering spike properties [8, 9, 18, 29, 31] and analyzing generalization error in spiked models [1, 2, 26, 27] are active research areas, existing analyses often characterize generalization implicitly (e.g., via fixed-point equations) or focus on specific spike strengths/alignments. *In contrast, we provide explicit, generic formulae for generalization error, enabling precise categorization of overfitting regimes and conditions for beneficial spike alignment.*

Appendix B. Misspecified Cases & Summary

B.1. Mis-Specified Case and no Covariate Shift

We next consider misspecified targets y with differing dependence on spike Z and noise A feature components. Specifically, we assume $\alpha_Z \neq \alpha_A$ but introduce no covariate shift between training and test distributions, i.e., $\tilde{\alpha}_Z = \alpha_Z$ and $\tilde{\alpha}_A = \alpha_A$. This scenario models situations where intrinsic

[1] Our code is available at the anonymous GitHub repository: link



(a) Under operator norm scaling ($\theta^2 = c\tau^2$) with $\alpha_Z = 1$, $\alpha_A = 2$, alignment initially improves generalization for small c, but becomes harmful beyond a critical point, leading to catastrophic overfitting.

(b) Under Frobenius norm scaling ($\theta = \sqrt{d}\tau$) with $\alpha_A = 1$ and $\alpha_Z = 1.1$, alignment remains better than anti-alignment across all c, but benign overfitting is not achieved unless $\alpha_Z = \alpha_A$.

Figure 2: Transition from beneficial to harmful alignment under mild misspecification. **The scatter plots are empirically obtained and the lines are theory.**

feature properties lead to differential correlations with the target, a common occurrence in practice. For notational convenience, we define $\Delta_c := \alpha_Z - \frac{\alpha_A}{c}$ with $\Delta_1 := \alpha_Z - \alpha_A$.

Theorem 4 *[Misspecified] Given data $Z, \tilde{Z}$ that satisfy Assumption 1, $A, \tilde{A}$ that satisfy Assumption 2 and $y, \tilde{y}$ according to Equation (2). If Assumption 3 holds with $\alpha_Z = \tilde{\alpha}_Z$, and $\alpha_A = \tilde{\alpha}_A$, then we have that*

$$\mathcal{R}_c = \begin{cases} \tau_\varepsilon^2 \frac{c}{1-c} + \tau^2 \left(\beta_*^T u\right)^2 \frac{\Delta_1^2}{1-c} \frac{\theta^2}{\theta^2+\tau^2} & c < 1 \\ \tau_\varepsilon^2 \frac{1}{c-1} + \alpha_A^2 \tau^2 \|\beta_*\|^2 \left(1 - \frac{1}{c}\right) + \tau^2 \left(\beta_*^T u\right)^2 \Delta_c^2 \frac{\theta^2}{\theta^2+\tau^2 c} \left[\frac{c}{c-1} \frac{\theta^2+\tau^2 c^2}{\theta^2+\tau^2 c} - 2\frac{\alpha_A}{\Delta_c}\right] & c > 1 \end{cases}$$

A key observation is that misspecification ($\alpha_Z \neq \alpha_A$) can itself induce double descent, even if $\tau_\varepsilon^2 = 0$. This contrasts with the well-specified case where, if $\tau_\varepsilon^2 = 0$, double descent is absent. However, in the misspecified case, we do not observe double descent if there is no alignment $\beta_*^T u = 0$.

Equal Operator Norm Case. For $\theta^2 = \gamma\tau^2$, the excess risk is

$$\mathcal{R} = \begin{cases} \tau^2 (\beta_*^T u)^2 \frac{\Delta_1^2}{1-c} \frac{\gamma}{\gamma+1} + \tau_\varepsilon^2 \frac{c}{1-c} & c < 1 \\ \tau^2 \frac{\gamma}{\gamma+c} (\beta_*^T u)^2 \Delta_c^2 \left[\left(\frac{c^2+\gamma}{\gamma+c} \frac{c}{c-1}\right) - 2\frac{\alpha_A}{\Delta_c}\right] + \alpha_A^2 \tau^2 \|\beta_*\|^2 \left(1 - \frac{1}{c}\right) + \tau_\varepsilon^2 \frac{1}{c-1} & c > 1 \end{cases}$$

For $c < 1$, the spike is *detrimental*. For $c > 1$, the behavior depends on α_Z/α_A. In particular, if

$$\frac{1}{c} \leq \frac{\alpha_Z}{\alpha_A} \leq \frac{1}{c} \left(\frac{3c^2 - \gamma + 2c\gamma - 2c}{(c^2 + \gamma)}\right),$$

then we have that the coefficient in front of $(\beta_*^T u)^2$ is negative. Thus, when α_Z/α_A lies between these thresholds, the spike *helps*, but the spike *is harmful* outside this range. As $c \to \infty$, if $\gamma = o_c(c^2)$, the beneficial region shrinks and *alignment increasingly harms generalization*. On the other hand, if the spike is big enough ($\gamma = \omega_c(c^2)$), we have that the beneficial region limits to $0 \leq \frac{\alpha_Z}{\alpha_A} \leq 2$. Figures 3a and 3b plot the coefficient of $(\beta_*^T u)^2$ for $c = 2$ and $c = 20$ for $\gamma = c$.

The upper bound on beneficial α_Z/α_A is surprising, as stronger target dependence on the spike might be expected to always favor alignment. Additionally, the dependence on the level of overparameterization c also offers new insights. Consider the example of $\gamma = c$, and $\alpha_Z/\alpha_A = 2$. Then when $c < 2$ or $c > (9 + \sqrt{57})/2$, we have that the ratio is outside the beneficial region. Figure 2a shows that in the beneficial region, the aligned risk is lower than the anti-aligned risk. However, outside the beneficial region, the aligned risk becomes strictly larger than the anti-aligned counterpart.

Next, in terms of benign vs. tempered vs. catastrophic overfitting, we have that

$$\lim_{c\to\infty} \mathcal{R}_c = \begin{cases} \tau^2\left[\gamma\alpha_Z^2(\beta_*^T u)^2 + \alpha_A^2\|\beta_*\|^2\right] & \beta_* \not\perp u, \gamma = \Theta_c(1) \\ \infty & \beta_* \not\perp u, \omega_c(1) \leq \gamma \leq o_c(c^2) \\ \tau^2\left[\alpha_A^2\|\beta_*\|^2 + \left(\alpha_Z^2\left(1+\frac{1}{\phi}\right) - 2\alpha_Z\alpha_A\right)(\beta_*^T u)^2\right] & \beta_* \not\perp u, \gamma = \phi c^2 \\ \tau^2(\alpha_A^2\|\beta_*\|^2 + (\alpha_Z^2 - 2\alpha_Z\alpha_A)(\beta_*^T u)^2) & \beta_* \not\perp u, \gamma = \omega_c(c^2) \\ \alpha_A^2\tau^2\|\beta_*\|^2 & \beta_* \perp u \end{cases}$$

For $\beta_* \not\perp u$, if $\omega_c(1) \leq \gamma \leq o_c(c^2)$ we have *catastrophic overfitting*. If $\gamma = \Theta_c(c^2)$, overfitting is tempered, with benign overfitting precluded (Appendix Proposition 1). If $\gamma = \omega_c(c^2)$, overfitting is again tempered with benign requiring returning to the well-specified case ($\alpha_A = \alpha_Z$).

Equal Frobenius Norm Case. For $\theta^2 = d\tau^2$, the excess risk becomes:

$$\mathcal{R}_{c>1} = \alpha_A^2\|\beta_*\|^2\left(1 - \frac{1}{c}\right) + (\beta_*^T u)^2\left[\frac{c}{c-1}\left(\alpha_Z - \frac{\alpha_A}{c}\right)^2 - 2\alpha_A\left(\alpha_Z - \frac{\alpha_A}{c}\right)\right] + \frac{\tau_\varepsilon^2}{c-1}.$$

For $c > 1$, the beneficial region for the ratio α_Z/α_A is defined by:

$$\frac{1}{c} \leq \frac{\alpha_Z}{\alpha_A} \leq 2 - \frac{1}{c}.$$

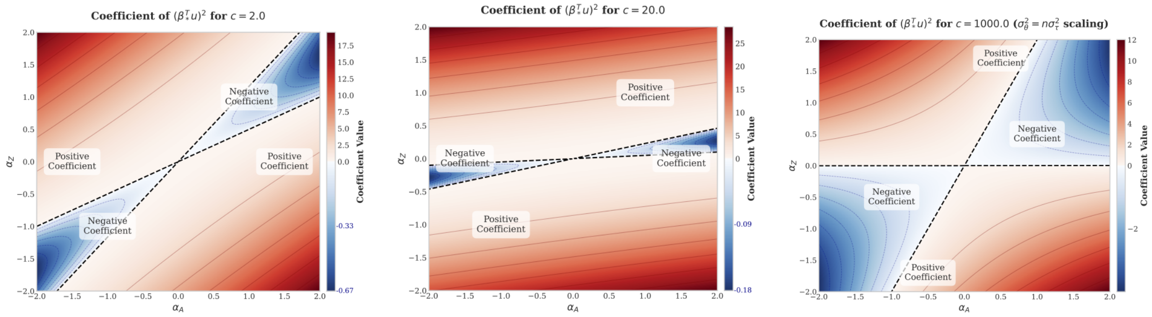
The beneficial region expands with c, making alignment increasingly beneficial in extreme overparameterization (Figure 3c). Beneficial alignment can also be seen in Figure 2b. Here $\alpha_Z/\alpha_A = 1.1$, which is in the beneficial region for $c > 10/9$. Finally, the overfitting is tempered unless $\alpha_A = \alpha_Z$.

B.2. Misspecified Target and Covariate Shift

Lastly, we consider the most challenging setting, where in addition to misspecifation, we also have covariate shift between train and test. Specifically, $\alpha_Z \neq \tilde{\alpha}_Z$ or $\alpha_A \neq \tilde{\alpha}_A$, hence we have the spike/noise importance differ between train and test.

Equal Operator Norm. In this case, we show the following.

Theorem 5 *Given data $Z, \tilde{Z}$ that satisfy Assumption 1, $A, \tilde{A}$ that satisfy Assumption 2 and $y, \tilde{y}$ according to Equation (2). If Assumption 3 holds, catastrophic overfitting occurs if $\tilde{\alpha}_Z = \alpha_Z$, $\beta_* \not\perp u$, and $\omega_c(1) \leq \gamma \leq o_c(c^2)$. Additionally, if $\tilde{\alpha}_Z \neq \alpha_Z$ with $\gamma = \omega_c(1)$ and $\beta_* \not\perp u$ we get catastrophic overfitting. Other scenarios yield tempered overfitting.*



(a) Operator norm scaling, $c = 2$. Here there is a large beneficial region.

(b) Operator norm scaling, $c = 20$. Here the beneficial region has shrunk

(c) Frobenius norm scaling, $c = 1000$. The beneficial region persists at extreme overparameterization.

Figure 3: Phase boundaries for spike alignment impact. Coefficient of $(\beta_*^T u)^2$ as a function of α_Z/α_A, indicating whether alignment improves or harms generalization.

Again, spike size and overfitting type show nuanced dependence. Additionally, different covariate shifts pose varying challenges. Specifically, if $\alpha_Z \neq \tilde{\alpha}_Z$, (target's spike dependence shifts), then catastrophic overfitting becomes unavoidable for sufficiently large spikes. That is unlike before, increasing the spike size does not mitigate catastrophic overfitting as it did before.

Equal Frobenius Norm. In this case, we have the following.

Theorem 6 *Given data $Z, \tilde{Z}$ that satisfy Assumption 1, $A, \tilde{A}$ that satisfy Assumption 2 and $y, \tilde{y}$ according to Equation (2). If Assumption 3 holds and $\alpha_Z \neq \tilde{\alpha}_Z$ then $\mathcal{R}_c = \infty$ for all $c \neq 1$. For $\alpha_Z = \tilde{\alpha}_Z$, we have that*

$$\lim_{c\to\infty} \mathcal{R}_c = \tau^2 \left[(\beta_*^T u)^2 (\alpha_Z^2 - 2\tilde{\alpha}_A \alpha_Z) + \|\beta_*\|^2 \tilde{\alpha}_A^2 \right].$$

Thus, if $\alpha_Z \neq \tilde{\alpha}_Z$, catastrophic overfitting is pervasive. When β_* and u are parallel, we have that

$$\tau^2 \|\beta_*\|^2 (\alpha_Z - \tilde{\alpha}_A)^2$$

This is benign if and only if $\alpha_Z = \tilde{\alpha}_A$. Notably, if training data is misspecified ($\alpha_A \neq \alpha_Z$) but test data is well-specified and matches the training spike dependence ($\alpha_Z = \tilde{\alpha}_Z = \tilde{\alpha}_A$), benign overfitting becomes achievable.

Tables 1 and 2 provide a summary of our observations.

Appendix C. General Theorem & Extension to Nonlinear Models

C.1. General Theorem

Prior results are special cases of our main theorem (Theorem 7). Its full form is complex (Appendix G). We present a high-level decomposition here.

Theorem 7 (Generalization Risk) *Suppose Assumption 1, Assumption 2, and Assumption 3 hold.*

$$\mathcal{R} = \mathbb{E}\left[\underbrace{\left\| \tilde{\alpha}_z \beta_*^T \tilde{Z} - \beta_{int}^T \tilde{Z} \right\|_F^2}_{Bias} + \underbrace{\tau^2 \left\| \beta_{int}^T \tilde{A} \right\|_F^2}_{Variance} + \underbrace{\tilde{\alpha}_A^2 \left\| \beta_*^T \tilde{A} \right\|_F^2}_{Data\ Noise} + \underbrace{\left(-2\tilde{\alpha}_A \beta_*^T \tilde{A}\tilde{A}^T \beta_{int} \right)}_{Target\ Alignment} \right]$$

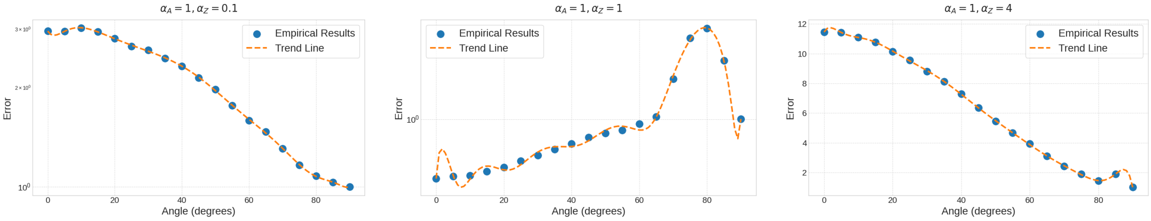
Table 1: Asymptotic Generalization Regimes. This table summarizes conditions for when overfitting is benign, tempered, or catastrophic in the limit where $d/n \to c$ and subsequently $c \to \infty$. The behavior depends on the spike scaling relative to the bulk, target alignment (β_* relative to spike direction u), and target specifications α_A, α_Z (train) and $\tilde{\alpha}_A, \tilde{\alpha}_Z$ (test). Here, θ^2 quantifies the scaled spike strength and τ^2 the scaled bulk variance; the two primary scaling regimes are operator norm based ($\theta^2 = \gamma\tau^2$) and Frobenius norm based ($\theta^2 = d\tau^2$). The ω, o, O, Θ are all as we send $c \to \infty$.

Scaling	Benign	Tempered	Catastrophic
Well-Specified, No Covariate Shift: $\alpha_A = \tilde{\alpha}_A = \alpha_Z = \tilde{\alpha}_Z = \alpha > 0$			
$\theta^2 = \gamma\tau^2$	$\gamma = \omega_c(c^2)$, $\beta_* \parallel u$	All other cases	$o_c(c^2) \geq \gamma \geq \omega_c(1)$, $\beta_* \not\perp u$
$\theta^2 = d\tau^2$	$\beta_* \parallel u$	$\beta_* \nparallel u$	Never
Misspecified, No Covariate Shift: $\alpha_A = \tilde{\alpha}_A, \alpha_Z = \tilde{\alpha}_Z, \alpha_A \neq \alpha_Z$			
$\theta^2 = \gamma\tau^2$	Never	All other cases	$o_c(c^2) \geq \gamma \geq \omega_c(1)$, $\beta_* \not\perp u$
$\theta^2 = d\tau^2$	Never	Always	Never
Misspecified with Covariate Shift: $\alpha_A \neq \tilde{\alpha}_A$ or $\alpha_Z \neq \tilde{\alpha}_Z$			
$\theta^2 = \gamma\tau^2$	Never	All other cases	$\alpha_Z \neq \tilde{\alpha}_Z, \beta_* \not\perp u, \gamma = \omega_c(1)$ or $\alpha_Z = \tilde{\alpha}_Z, \beta_* \not\perp u,$ $\omega_c(1) \leq \gamma \leq o_c(c^2)$
$\theta^2 = d\tau^2$	$\alpha_Z = \tilde{\alpha}_Z = \tilde{\alpha}_A,$ $\beta_* \parallel u$	All other cases	$\alpha_Z \neq \tilde{\alpha}_Z$ and $\beta_* \not\perp u$
Spike Recovery: $\alpha_A = \tilde{\alpha}_A = 0, \alpha_Z = \tilde{\alpha}_Z$			(Appendix F)
$\theta^2 = \gamma\tau^2$	$\gamma\tau^2 = o_c(1)$	$\gamma\tau^2 = \Theta_c(1)$	$\gamma\tau^2 = \omega_c(1)$
$\theta^2 = d\tau^2$	$\tau^2 = o_c(1)$	$\tau^2 = \Theta_c(1)$	Never

- **Bias.** This is the squared error between the learned predictor β_{int} and the true parameter β_* *projected onto the spike direction* u. In particular, the risk penalizes discrepancies only along the top eigen-direction of the population covariance Σ, reflecting the anistropic influence of the spike.
- **Variance.** The variance is equivalent to $\tau^2\|\beta_{int}\|_2$. This mirrors classical isotropic regression results [6, 16], but the norm $\|\beta_{int}\|^2$ itself is dependent upon the interaction between signal and noise components, the alignment between β_* and u, and the scaling parameters.
- **Data Noise.** The data noise term quantifies the contribution of the noise matrix A to the target outputs y_i through α_A. Even in the absence of observation noise ($\tau_\varepsilon^2 = 0$), target corruption via data noise can create an irreducible error floor.

Table 2: Conditions for Beneficial Spike Alignment at Finite Aspect Ratios ($c = d/n$). This table outlines the specific regions where alignment of the target signal with the data's principal spike direction improves generalization. Conditions depend on the problem setting (well-specified vs. mis-specified), the spike scaling regime (operator or frobenius norm based), the overparameterization level $c = d/n$, and the relative dependence of the targets y on the spike versus the bulk α_Z/α_A.

Setting	Alignment Beneficial Region
Well-Specified, Operator Norm	$\gamma > c(c-2)$
Well-Specified, Frobenius Norm	$c > 1$
Misspecified, No Covariate Shift, Operator Norm	$\frac{1}{c} \leq \frac{\alpha_Z}{\alpha_A} \leq \frac{1}{c}\left(\frac{3c^2-\gamma+2c\gamma-2c}{(c^2+\gamma)}\right)$
Misspecified, No Covariate Shift, Frobenius Norm	$\frac{1}{c} < \frac{\alpha_Z}{\alpha_A} < 2 - \frac{1}{c}$



(a) $\alpha_Z = 0.1$, alignment helps. **(b)** $\alpha_Z = 1$, mixed behavior. **(c)** $\alpha_Z = 4$, alignment hurts.

Figure 4: Alignment-phase transitions persist in deep networks. Generalization error vs. angle between spike direction u and ground-truth parameter β_* when fitting data with a 3-layer ReLU networks. The effect of alignment switches as α_Z increases, consistent with the phase transitions predicted by our theory. Experimental details are in AppendixE.

- **Target Alignment.** The alignment term measures the inner product between β_{int} and β_* with respect to the sample noise covariance. This cross-term captures how mismatch between β_{int} and β_*, especially when mediated by A, can amplify or dampen generalization error.

C.2. Extension: Nonlinear Models Also Exhibit Alignment Phase Transitions

While our theoretical focus is on linear regression, key phenomena like α_Z dependent non-monotonic alignment effects appear in nonlinear models as well. We test this by training 3-layer ReLU networks to predict y (Equation (2)) given X, where we vary the alignment angle between spike u and β_* and record the generalization error. Figure 4, shows our results for three α_Z values. For $\alpha_Z = 0.1$, increasing alignment with the spike is detrimental. For $\alpha_Z = 1$, alignment is beneficial, while for $\alpha_Z = 10$, alignment is detrimental again. This mirrors our theoretical findings that there is a region for beneficial alignment and a nuanced phase transition for different α_Z values.

Appendix D. Notation

Table 3 summarizes our notations used throughout the main text and Appendix.

Symbol	Description / Role	Typical scaling / range	First used
d, n	Data dimension and sample size	$d, n \to \infty$ with $c = d/n$ fixed	Sec. 2
c	Aspect ratio d/n	$(0, \infty)$	Sec. 2
τ^2	Noise variance in ambient bulk A	$\Theta(1)$ (or ρ^2/d under alt. notation)	Sec. 2
θ^2	Spike (signal) variance	$\theta^2 = \gamma\tau^2$ (operator-norm) or $\theta^2 = d\tau^2$ (Frobenius)	Sec. 2
γ	Spike-to-noise ratio $\gamma = \theta^2/\tau^2$ (effective outlier eigenvalue)	$[0, \infty)$; critical line $\gamma = (1 + \sqrt{c})^2$	Sec. 2
α_Z, α_A	Coeffs. weighting spike vs. bulk in *targets* y	$\Theta(1)$	Eq. (2)
$\tilde{\alpha}_Z$, $\tilde{\alpha}_A$	Same coefficients for *test* data (covariate shift)	$\Theta(1)$	Sec. 3
β_*	True parameter vector	$\\|\beta_*\\|_2 = 1$	Sec. 2
u	Spike direction in data covariance	$\\|u\\|_2 = 1$	Sec. 2
A, Z	Bulk noise matrix, rank-one signal matrix	$A_{ij} \sim \mathcal{N}(0, \tau^2/n),\ Z = \theta\, uv^\top$	Sec. 2
$\varepsilon, \tau_\varepsilon^2$	Label noise and its variance	IID, $\mathcal{N}(0, \tau_\varepsilon^2)$	Sec. 2

Table 3: Glossary of recurrent parameters and symbols. All $\Theta(1)$ constants are independent of n, d.

Appendix E. Non-Linear Experiment

We used 500 data points in 750 dimensional space, with a hidden width of 1000. We used full batch gradient descent for 100 epochs with a learning rate of 1e-4. Each data point is averaged over 50 trials. Equal Frobenius norm scaling was used for the size of the spike.

Appendix F. Spike Recovery Case

Finally, we consider the special case where the goal is to recover the spike direction u. In this setting, the target y depends only on the spike component Z, with no contribution from the noise A:

$$\alpha_A = \tilde{\alpha}_A = 0, \qquad \alpha_Z = \tilde{\alpha}_Z = \alpha > 0.$$

Thus, the target y is proportional to the signal Z plus possible observation noise ε.

Equal Operator Norm In this regime, we have that the risk is

$$\mathcal{R}_{c<1} = \frac{\gamma\alpha_Z^2\tau^2}{(1-c)(\gamma+1)}(\beta^T u)^2 + \frac{c}{1-c}\tau_\varepsilon^2, \qquad \mathcal{R}_{c>1} = \frac{\gamma c(c^2+\gamma)\alpha_Z^2\tau^2}{(c-1)(\gamma+c)^2}(\beta^T u)^2 + \frac{1}{c-1}\tau_\varepsilon^2$$

Here again, we see that when $\gamma = \Theta_c(1)$, we have tempered overfitting and $\omega_c(1) \leq \gamma \leq o_c(c^2)$, we have catastrophic overfitting and for $\gamma = \Omega_c(c^2)$ we get tempered overfitting again.

Equal Frobenius Norm. In this regime, we have that

$$R_{c<1} = \frac{\alpha_Z^2\tau^2}{1-c}(\beta^T u)^2 + \frac{c}{1-c}\tau_\varepsilon^2 \quad R_{c>1} = \frac{c\alpha_Z^2\tau^2}{c-1}(\beta^T u)^2 + \frac{1}{c-1}\tau_\varepsilon^2$$

This generalizes the spike recovery setting studied in [31], which assumed noiseless targets ($\tau_\varepsilon = 0$) and the equal Frobenius norm scaling. Our formula allows for observation noise and thus captures the more realistic case where the target y itself contains randomness not aligned with the spike. Here we see that we have tempered overfitting unless $\tau^2 = o(1)$, which is the case considered in [31].

Appendix G. Theorem 7

We have that as $n, d \to \infty$ with $d/n \to c \in (0, \infty)$, we have the following expressions for each term.
Bias: For $c < 1$, we have that the bias term is

$$\tilde{\theta}^2 \left[(\beta_*^T u)^2 \left(\tilde{\alpha}_Z - \alpha_Z + (\alpha_Z - \alpha_A) + \frac{\tau^2}{\theta^2 + \tau^2} \right)^2 + \tau_\varepsilon^2 \frac{c}{1-c} \frac{1}{d(\theta^2 + \tau^2)} \right].$$

If $c > 1$, we that the bias term is

$$\tilde{\theta}^2 (\beta_*^T u)^2 \left(\tilde{\alpha}_Z - \alpha_Z + \left(\alpha_Z - \frac{\alpha_A}{c} \right) \frac{\tau^2 c}{\theta^2 + \tau^2 c} \right)^2 + \tilde{\theta}^2 \left[\alpha_A^2 \frac{\|\beta_*\|^2}{d} \frac{c-1}{c} \frac{\theta^2 \tau^2 c}{(\theta^2 + \tau^2 c)^2} + \tau_\varepsilon^2 \frac{c}{c-1} \frac{\theta^2 + \tau^2}{n(\theta^2 + \tau^2 c)^2} \right].$$

Variance: For $c < 1$, we have that the variance term is

$$\alpha_A^2 \tilde{\tau}^2 \|\beta_*\|^2 + \tilde{\tau}^2 (\beta_*^T u)^2 \left[\frac{1}{1-c} \frac{\theta^4 + \theta^2 \tau^2 c}{(\theta^2 + \tau^2)^2} (\alpha_Z - \alpha_A)^2 + 2\alpha_A (\alpha_Z - \alpha_A) \frac{\theta^2}{\theta^2 + \tau^2} \right]$$
$$+ \tau_\varepsilon^2 \frac{\tilde{\tau}^2}{\tau^2} \left[\frac{c}{1-c} - \frac{\theta^2}{d(\theta^2 + \tau^2)} \frac{c}{1-c} \right].$$

For $c > 1$, we have that the variance term is

$$\tilde{\tau}^2 \|\beta_*\|^2 \left(\frac{\alpha_A^2}{c} - \frac{\alpha_A^2}{d} \frac{\theta^2}{\theta^2 + \tau^2 c} \right) + \tilde{\tau}^2 (\beta_*^T u)^2 \frac{c}{(c-1)} \frac{\theta^2}{\theta^2 + \tau^2 c} \left(\alpha_Z - \frac{\alpha_A}{c} \right)^2$$
$$+ \tau_\varepsilon^2 \frac{\tilde{\tau}^2}{\tau^2} \left(\frac{1}{c-1} - \frac{\theta^2}{d(\theta^2 + \tau^2 c)} \frac{c}{c-1} \right).$$

Data Noise: For all c, we have that

$$\tilde{\alpha}_A^2 \tilde{\tau}^2 \|\beta_*\|^2.$$

Target Alignment: For $c < 1$, we have that the alignment term is

$$-2\tilde{\alpha}_A \tilde{\tau}^2 \left((\alpha_Z - \alpha_A) \frac{\theta^2}{\theta^2 + \tau^2} (\beta_*^T u)^2 + \alpha_A \|\beta_*\|^2 \right).$$

For $c > 1$, we have that the alignment term is

$$-2\tilde{\alpha}_A \tilde{\tau}^2 \left(\left(\alpha_Z - \frac{\alpha_A}{c} \right) \frac{\theta^2}{\theta^2 + \tau^2 c} (\beta_*^T u)^2 + \alpha_A \|\beta_*\|^2 \left(\frac{1}{c} - \frac{1}{d} \frac{\theta^2}{\theta^2 + \tau^2 c} \right) \right).$$

Error terms: The largest error terms for all c are:

$$o(1) + O\left(\frac{1}{n}\right).$$

Remark: We note that the above theorem is very general and captures all of the theorems in the main text as special cases. It is worth noting that the theorem also incorporates different signal and bulk strengths for test data, namely for $\tilde{\theta}$ and $\tilde{\tau}$.

Appendix H. Proof of Theorem 7

Theorem 7 (Generalization Risk) *Suppose Assumption 1, Assumption 2, and Assumption 3 hold.*

$$\mathcal{R} = \mathbb{E}\left[\underbrace{\left\|\tilde{\alpha}_z\beta_*^T\tilde{Z} - \beta_{int}^T\tilde{Z}\right\|_F^2}_{Bias} + \underbrace{\tau^2\left\|\beta_{int}^T\tilde{A}\right\|_F^2}_{Variance} + \underbrace{\tilde{\alpha}_A^2\left\|\beta_*^T\tilde{A}\right\|_F^2}_{Data\ Noise} + \underbrace{\left(-2\tilde{\alpha}_A\beta_*^T\tilde{A}\tilde{A}^T\beta_{int}\right)}_{Target\ Alignment}\right]$$

The proof will be broken up into roughly three steps

1. **Decompose the error into four terms.** We shall refer to these terms as the 1) bias, 2) variance, 3) data noise, and 4) target alignment.

2. **Simplify the expressions.** We shall then use the result from [25] to simplify the expression for each of the four terms. In particular, we shall express each term as the product of dependent functions of the eigenvalues of X.

3. **Random matrix theory estimate.** We then use standard results from random matrix theory such as [3, 4, 22] to obtain a closed-form formula of the risk.

In order to better align with existing results and use them accordingly, we change our scalings for now and switch back after our derivation. In particular, consider the covariance matrix and data matrix, we now make:

$$Z = \eta u v^T, \quad \text{where } \theta = \frac{\eta}{\sqrt{n}} \text{ and } \|v\| = 1,$$

$$\tau = \frac{\rho}{\sqrt{d}} \quad \text{for the variance of the noise matrix } A,$$

where ρ, κ have the corresponding scalings that we need to match. Note that with the change of variable, the covariance matrix can be rewritten as:

$$\Sigma = \theta^2 uu^T + \tau^2 I_d = \frac{\kappa^2}{n}uu^T + \frac{\rho^2}{d}I_d.$$

In the general theorem, we can accommodate more room for distribution shift, that is, we consider $\tilde{n}$ test data with $\tilde{Z} = \tilde{\theta}u\tilde{v}^T$ and test noise matrix with variance $\tilde{\tau}$. We define the corresponding $\tilde{\eta}$, $\tilde{\rho}$, and suppose $\tilde{n} = \Theta(n)$, $\tilde{\tau} = \Theta(\tau)$, and $\tilde{\theta} = \Theta(\theta)$.

H.1. Decompose Error

Using the fact that $\tilde{A}$ has been zero entries and is independent of $\tilde{Z}$, we see that we can decompose the error as follows:

$$\begin{aligned}
&\frac{1}{\tilde{n}}\left\|\beta_*^T(\tilde{\alpha}_z\tilde{Z} + \tilde{\alpha}_A\tilde{A}) - \beta_{int}^T(\tilde{Z} + \tilde{A})\right\|_F^2 \\
\overset{\mathbb{E}}{=}&\frac{1}{\tilde{n}}\left\|\tilde{\alpha}_z\beta_*^T\tilde{Z} - \beta_{int}^T\tilde{Z}\right\|_F^2 + \frac{1}{\tilde{n}}\left\|\tilde{\alpha}_A\beta_*^T\tilde{A} - \beta_{int}^T\tilde{A}\right\|_F^2 \\
\overset{\mathbb{E}}{=}&\underbrace{\frac{1}{\tilde{n}}\left\|\tilde{\alpha}_z\beta_*^T\tilde{Z} - \beta_{int}^T\tilde{Z}\right\|_F^2}_{Bias} + \underbrace{\frac{1}{\tilde{n}}\left\|\beta_{int}^T\tilde{A}\right\|_F^2}_{Variance} + \underbrace{\frac{1}{\tilde{n}}\tilde{\alpha}_A^2\left\|\beta_*^T\tilde{A}\right\|_F^2}_{Data\ Noise} + \underbrace{\left(-\frac{2}{\tilde{n}}\tilde{\alpha}_A\beta_*^T\tilde{A}\tilde{A}^T\beta_{int}\right)}_{Target\ Alignment}
\end{aligned}$$

We compute these four terms one by one in the following sections.

H.2. Simplifying Terms and RMT estimates

This section simplifies the four terms and provides the random matrix theory estimates.

H.2.1. BIAS

Using Theorem 14, we have that if $c < 1$

$$\tilde{\alpha}_z \beta_*^T \tilde{Z} - \beta_{int}^T \tilde{Z} = \left[\tilde{\alpha}_Z - \alpha_Z + \frac{\xi}{\gamma_1}(\alpha_Z - \alpha_A)\right] \beta_*^T \tilde{Z} + \frac{\tilde{\eta}}{\eta}\frac{\xi}{\gamma_1}\varepsilon^T p_1 \tilde{v}^T,$$

and if $c > 1$

$$\tilde{\alpha}_z \beta_*^T \tilde{Z} - \beta_{int}^T \tilde{Z} = \beta_*^T \left[(\tilde{\alpha}_Z - \alpha_Z)I + \frac{\xi}{\gamma_2}(\alpha_Z I - \alpha_A A A^\dagger)\right] \tilde{Z} - \alpha_A \frac{\eta\|s\|^2}{\gamma_2}\beta_*^T h^T u^T \tilde{Z} + \frac{\tilde{\eta}}{\eta}\frac{\xi}{\gamma_2}\varepsilon^T p_2 \tilde{v}^T.$$

Then we are interested in the norm. Hence, we see that in $c < 1$ case, we have that

$$\begin{aligned}\|\tilde{\alpha}_z \beta_*^T \tilde{Z} - \beta_{int}^T \tilde{Z}\|^2 &= \left[\tilde{\alpha}_Z - \alpha_Z + \frac{\xi}{\gamma_1}(\alpha_Z - \alpha_A)\right]^2 \beta_*^T \tilde{Z}\tilde{Z}^T \beta_*^T + \frac{\tilde{\eta}^2}{\eta^2}\frac{\xi^2}{\gamma_1^2}\varepsilon^T p_1 \tilde{v}^T \tilde{v} p_1 \varepsilon \\ &\quad + 2\left[\tilde{\alpha}_Z - \alpha_Z + \frac{\xi}{\gamma_1}(\alpha_Z - \alpha_A)\right] \frac{\tilde{\eta}}{\eta}\frac{\xi}{\gamma_1}\varepsilon^T p_1 \tilde{v}^T \tilde{Z}^T \beta_*^T.\end{aligned}$$

Then taking the expectation with respect of ε, we get that the cross term disappears and a trace in the second term (Theorem 25).

$$\mathbb{E}_\varepsilon\left[\|\tilde{\alpha}_z \beta_*^T \tilde{Z} - \beta_{int}^T \tilde{Z}\|^2\right] = \left[\tilde{\alpha}_Z - \alpha_Z + \frac{\xi}{\gamma_1}(\alpha_Z - \alpha_A)\right]^2 \beta_*^T \tilde{Z}\tilde{Z}^T \beta_*^T + \frac{\tilde{\eta}^2}{\eta^2}\frac{\xi^2}{\gamma_1^2}\tau_\varepsilon^2 \operatorname{Tr}\left(p_1 \tilde{v}^T \tilde{v} p_1^T\right).$$

Using the fact that $\tilde{Z} = \tilde{\eta} u \tilde{v}^T$ and that $\tilde{v}$ has unit norm, we get

$$\mathbb{E}_\varepsilon\left[\|\tilde{\alpha}_z \beta_*^T \tilde{Z} - \beta_{int}^T \tilde{Z}\|^2\right] = \left[\tilde{\alpha}_Z - \alpha_Z + \frac{\xi}{\gamma_1}(\alpha_Z - \alpha_A)\right]^2 \tilde{\eta}^2 (\beta_*^T u)^2 + \frac{\tilde{\eta}^2}{\eta^2}\frac{\xi^2}{\gamma_1^2}\tau_\varepsilon^2 \|p_1\|^2.$$

Then using Theorem 28, we see the second term is equal to

$$\tilde{\eta}^2 \tau_\varepsilon^2 \frac{c}{1-c}\frac{1}{\eta^2 c + \rho^2} + o\left(\frac{\tilde{\eta}^2}{\eta^2\rho^2}\right) + O\left(\frac{\tilde{\eta}^2}{\eta^2\rho^2 n}\right) = \tau_\varepsilon^2 \frac{\tilde{\eta}^2 c}{1-c}\frac{1}{\eta^2 c + \rho^2} + o\left(\frac{1}{\rho^2}\right) + O\left(\frac{1}{\rho^2 n}\right).$$

For the first term, expanding the coefficient, we get

$$(\tilde{\alpha}_Z - \alpha_Z)^2 + \frac{1}{\eta^2}\frac{\eta^2\xi^2}{\gamma_1^2}(\alpha_Z - \alpha_A)^2 + \frac{1}{\eta}2\frac{\eta\xi}{\gamma_1}(\alpha_Z - \alpha_A)(\tilde{\alpha}_Z - \alpha_Z).$$

Taking the expectation and using Theorem 21, the last term is

$$2\frac{\rho^2}{\eta^2 c + \rho^2}(\alpha_Z - \alpha_A)(\tilde{\alpha}_Z - \alpha_Z) + o\left(\frac{1}{\eta^2\rho^2}\right) + O\left(\frac{1}{\eta n \rho}\right).$$

For the middle term, we write

$$\frac{1}{\eta^2}\frac{\eta^2\xi^2}{\gamma_1^2} = \frac{1}{\eta^2}\cdot\frac{\eta^2}{\gamma_1}\cdot\frac{\xi^2}{\gamma_1}.$$

Then using Theorem 20 and Theorem 21, we have that the expectation of this is

$$\frac{1}{\eta^2}\left[\left(\frac{\rho^2\eta^2}{\eta^2 c+\rho^2}+o\left(\frac{1}{\rho^2}\right)\right)\left(\frac{\rho^2}{\eta^2 c+\rho^2}+o\left(\frac{1}{\eta^2\rho^2}\right)+O\left(\frac{1}{n}\right)\right)+O\left(\frac{1}{\eta^2 n}+\frac{1}{n^{1.5}}\right)\right].$$

We can simplify this as:

$$\frac{\rho^4}{(\eta^2 c+\rho^2)^2}+o\left(\frac{1}{\eta^4}\right)+O\left(\frac{\rho^2}{\eta^2 n}\right)$$

Thus, we have that this term is

$$\left[(\tilde{\alpha}_Z-\alpha_Z)+\frac{\rho^2}{\eta^2 c+\rho^2}(\alpha_Z-\alpha_A)\right]^2+o\left(\frac{1}{\eta^4}\right)+O\left(\frac{\rho^2}{\eta^2 n}\right)+o\left(\frac{1}{\eta^2\rho}\right)+O\left(\frac{1}{\eta n\rho}\right).$$

Thus, the non-error terms are

$$\tilde{\eta}^2\left(\left[(\tilde{\alpha}_Z-\alpha_Z)+\frac{\rho^2}{\eta^2 c+\rho^2}(\alpha_Z-\alpha_A)\right]^2(\beta_*^T u)^2+\tau_\varepsilon^2\frac{c}{1-c}\frac{1}{\eta^2 c+\rho^2}\right),$$

with an error term of

$$\tilde{\eta}^2\left[o\left(\frac{1}{\eta^2\rho^2}\right)+O\left(\frac{\rho^2}{\eta^2 n}\right)+O\left(\frac{1}{\eta n\rho}\right)\right]=o\left(\frac{1}{\rho^2}\right)+O\left(\frac{\rho^2}{n}\right)+O\left(\frac{\eta}{n\rho}\right).$$

Dividing by $\tilde{n}$, we then have the desired results:

$$\frac{\tilde{\eta}^2}{\tilde{n}}\left(\left[(\tilde{\alpha}_Z-\alpha_Z)+\frac{\rho^2}{\eta^2 c+\rho^2}(\alpha_Z-\alpha_A)\right]^2(\beta_*^T u)^2+\tau_\varepsilon^2\frac{c}{1-c}\frac{1}{\eta^2 c+\rho^2}\right)+o\left(\frac{1}{\eta^2\rho n}\right)+O\left(\frac{\rho^2}{n^2}\right)+O\left(\frac{\eta}{n^2\rho}\right).$$

Case: $c>1$ To help simplify notation we shall let $c_1=\tilde{\alpha}_Z-\alpha_Z$. We shall look at this term by term. We begin, with the last term squared. Here we see that using Theorem 28

$$\begin{aligned}\left\|\frac{\tilde{\eta}}{\eta}\frac{\xi}{\gamma_2}\varepsilon^T p_2\tilde{v}^T\right\|^2&=\left(\frac{\tilde{\eta}}{\eta}\right)^2\frac{\xi^2}{\gamma_2^2}\|p_2\|^2\\&\stackrel{E}{=}\left(\frac{\tilde{\eta}}{\eta}\right)^2\tau_\varepsilon^2\left[\frac{\eta^2}{c-1}\frac{\eta^2 c+\rho^2}{(\eta^2+\rho^2)^2}+o\left(\frac{1}{\rho^2}+\frac{1}{n}\right)+O\left(\frac{1}{n}\right)\right]\\&=\tilde{\eta}^2\frac{\tau_\varepsilon^2}{c-1}\frac{\eta^2 c+\rho^2}{(\eta^2+\rho^2)^2}+o\left(\frac{1}{\rho^2}+\frac{1}{n}\right)+O\left(\frac{1}{n}\right).\end{aligned}$$

For the middle term, since $\tilde{Z}=\tilde{\eta}u\tilde{v}$, we have that

$$\left\|\alpha_A\frac{\eta\|s\|^2}{\gamma_2}\beta_*^T h^T u^T\tilde{Z}\right\|^2=\alpha_A^2\tilde{\eta}^2\frac{\eta^2\|s\|^4}{\gamma_2^2}\beta_*^T h^T h\beta_*=\alpha_A^2\frac{\tilde{\eta}^2}{\eta^2}\frac{\eta^4\|s\|^4}{\gamma_2^2}\beta_*^T h^T h\beta_*.$$

By Equation 7, we first have that

$$\mathbb{E}\left[\frac{\eta^4\|s\|^4}{\gamma_2^2}\right]=\left(1-\frac{1}{c}\right)^2\frac{\rho^4\eta^4}{(\eta^2+\rho^2)^2}+o(\rho^2)+O\left(\frac{\rho^2}{n}\right),\ \mathrm{Var}\left(\frac{\eta^4\|s\|^4}{\gamma_2^2}\right)=O\left(\frac{\rho^4}{n}\right).$$

Then using this and Lemma 36, we get

$$\mathbb{E}\left[\frac{\eta^4\|s\|^4}{\gamma_2^2}\beta_*^T h^T h\beta_*\right] = \frac{\|\beta_*\|^2}{d}\left(\frac{c-1}{c}\right)\frac{\eta^4\rho^2}{(\eta^2+\rho^2)^2} + o\left(\frac{\rho^2}{n}\right) + O\left(\frac{1}{n^{1.5}}\right).$$

Thus the final expression for the middle term is

$$\begin{aligned}\mathbb{E}\left[\left\|\alpha_A\frac{\eta\|s\|^2}{\gamma_2}\beta_*^T h^T u^T\tilde{Z}\right\|^2\right] &= \alpha_A^2\frac{\tilde{\eta}^2}{\eta^2}\left[\frac{\|\beta_*\|^2}{d}\left(\frac{c-1}{c}\right)\frac{\eta^4\rho^2}{(\eta^2+\rho^2)^2} + o\left(\frac{\rho^2}{n}\right) + O\left(\frac{1}{n^{1.5}}\right)\right] \\ &= \alpha_A^2\tilde{\eta}^2\frac{\|\beta_*\|^2}{d}\left(\frac{c-1}{c}\right)\frac{\eta^2\rho^2}{(\eta^2+\rho^2)^2} + o\left(\frac{\rho^2}{n}\right) + O\left(\frac{1}{n^{1.5}}\right).\end{aligned}$$

The first term can be broken into three terms

$$c_1^2\tilde{\eta}^2(\beta_*^T u)^2 + \tilde{\eta}^2\frac{\xi^2}{\gamma_2^2}\beta_*^T(\alpha_Z I - \alpha_A AA^\dagger)uu^T((\alpha_Z I - \alpha_A AA^\dagger)\beta_* + 2c_1\tilde{\eta}^2\frac{\xi}{\gamma_2}\beta_*^T(\alpha_Z I - \alpha_A AA^\dagger)uu^T\beta_*.$$

Not that for the second and third term, we have that ξ, γ_2 only depend on the singular values of A and the rest only depend on the singular vectors. Hence, these terms are independent. For the middle term, we have

$$\beta_*^T(\alpha_Z I - \alpha_A AA^\dagger)uu^T((\alpha_Z I - \alpha_A AA^\dagger)\beta_* \overset{E}{=} \left(\alpha_Z - \frac{\alpha_A}{c}\right)^2(\beta_*^T u)^2.$$

While for the last term, we have

$$\beta_*^T(\alpha_Z I - \alpha_A AA^\dagger)uu^T\beta_* \overset{E}{=} \left(\alpha_Z - \frac{\alpha_A}{c}\right)(\beta_*^T u)^2.$$

Thus putting it together, we get

$$\begin{aligned}\tilde{\eta}^2(\beta_*^T u)^2\left[c_1^2 + \frac{\xi^2}{\gamma_2^2}\left(\alpha_Z - \frac{\alpha_A}{c}\right)^2 + 2c_1\frac{\xi}{\gamma_2}\left(\alpha_Z - \frac{\alpha_A}{c}\right)\right] &\overset{E}{=} \tilde{\eta}^2(\beta_*^T u)^2\left[c_1 + \frac{\xi}{\gamma_2}\left(\alpha_Z - \frac{\alpha_A}{c}\right)\right]^2 \\ &= \tilde{\eta}^2(\beta_*^T u)^2\left[\tilde{\alpha}_Z - \alpha_Z + \frac{\xi}{\gamma_2}\left(\alpha_Z - \frac{\alpha_A}{c}\right)\right]^2.\end{aligned}$$

Then with the same argument as for the $c < 1$, taking the expectation with respect to the singular values, we get

$$\tilde{\eta}^2(\beta_*^T u)^2\left[\tilde{\alpha}_Z - \alpha_Z + \frac{\rho^2}{\eta^2+\rho^2}\left(\alpha_Z - \frac{\alpha_A}{c}\right)\right]^2 + o\left(\frac{1}{\rho^2}\right) + O\left(\frac{\eta}{\rho n}\right).$$

The final term we need is the cross term between the first and second terms, and the cross terms with ε have mean zero. This cross term can be broken up into sub-terms, for which we compute their errors using Theorem 18, Theorem 19, Theorem 21, and Theorem 24. These cross-terms respectively become:

$$\begin{aligned}-2\tilde{\eta}^2 c_1\frac{\eta\|s\|^2}{\gamma_2}\beta_*^T uh\beta_* &= -2\frac{\tilde{\eta}^2}{\eta}c_1\frac{\eta^2\|s\|^2}{\gamma_2}\beta_*^T uh\beta_* = O\left(\frac{\eta}{\rho n}\right), \\ -2\alpha_Z\tilde{\eta}^2\frac{\eta\xi\|s\|^2}{\gamma_2^2}\beta_*^T uh\beta_* &= -2\alpha_Z\frac{\tilde{\eta}^2}{\eta^2}\frac{\eta^2\|s\|^2}{\gamma_2}\frac{\eta\xi}{\gamma_2}\beta_*^T uh\beta_* = O\left(\frac{1}{\sqrt{\rho}n}\right),\end{aligned}$$

$$2\alpha_A\tilde{\eta}^2\frac{\eta\xi\|s\|^2}{\gamma_2^2}\beta_*^TAA^\dagger uh\beta_* = -2\alpha_A\frac{\tilde{\eta}^2}{\eta^2}\frac{\eta^2\|s\|^2}{\gamma_2}\frac{\eta\xi}{\gamma_2}\beta_*^TAA^\dagger uh\beta_* = O\left(\frac{1}{\sqrt{\rho}n}\right).$$

Thus the cross term concentrates to zero at a controlled rate. Thus collecting all the terms and dividing by $\tilde{n}$, we get that the non-error terms are

$$\frac{\tilde{\eta}^2}{\tilde{n}}\left[(\beta_*^Tu)^2\left(\tilde{\alpha}_Z-\alpha_Z+\left(\alpha_Z-\frac{\alpha_A}{c}\right)\frac{\rho^2}{\eta^2+\rho^2}\right)^2+\alpha_A^2\frac{\|\beta_*\|^2}{d}\left(\frac{c-1}{c}\right)\frac{\eta^2\rho^2}{(\eta^2+\rho^2)^2}+\frac{\tau_\varepsilon^2}{c-1}\frac{\eta^2c+\rho^2}{(\eta^2+\rho^2)^2}\right]$$

with an error term of $o\left(\frac{\rho^2}{n^2}\right)+O\left(\frac{1}{n}\right)$.

H.2.2. VARIANCE

Lemma 17 implies that the expectation will be the weighted sum of the expressions from Lemmas 29, 30, 31, 32. Informally,

$$\frac{\tilde{\rho}^2}{d}\left(\alpha_Z^2 * \text{Lemma } 29 + 2\alpha_Z\alpha_A * \text{Lemma } 31 + \alpha_A^2 * \text{Lemma } 30 + \text{Lemma } 32\right).$$

This yields that for $c<1$, after simplification, the variance is

$$\begin{aligned}\frac{\tilde{\rho}^2}{d}&\left[\alpha_A^2\|\beta_*\|^2+(\beta_*^Tu)^2\left[(\alpha_Z-\alpha_A)^2\frac{\eta^2(\eta^2+\rho^2)}{(\eta^2c+\rho^2)^2}\frac{c^2}{1-c}+2\alpha_A(\alpha_Z-\alpha_A)\frac{\eta^2c}{\eta^2c+\rho^2}\right]\right.\\&\left.+\tau_\varepsilon^2\left(\frac{c}{1-c}\frac{d}{\rho^2}-\frac{\eta^2}{\rho^2(\eta^2c+\rho^2)}\frac{c^2}{1-c}\right)\right].\end{aligned}$$

For $c>1$, we similarly simplify it to:

$$\begin{aligned}\frac{\tilde{\rho}^2}{d}&\left[\|\beta_*\|^2\left(\frac{\alpha_A^2}{c}-\frac{\alpha_A^2}{d}\frac{\eta^2}{\eta^2+\rho^2}\right)+(\beta_*^Tu)^2\frac{c}{c-1}\frac{\eta^2}{\eta^2+\rho^2}\left(\alpha_Z-\frac{\alpha_A}{c}\right)^2\right.\\&\left.+\tau_\varepsilon^2\left(\frac{d}{\rho^2}\frac{1}{c-1}-\frac{\eta^2}{\rho^2(\eta^2+\rho^2)}\frac{c}{c-1}\right)\right]\end{aligned}$$

H.2.3. DATA NOISE

The data noise term is the simplest to understand. Preliminary calculation gives us.

$$\frac{1}{\tilde{n}}\tilde{\alpha}_A^2\left\|\beta_*^T\tilde{A}\right\|_F^2\overset{E}{=}\frac{\tilde{\alpha}_A^2}{\tilde{n}}\frac{\tilde{\rho}^2\tilde{n}}{d}\|\beta_*\|^2=\frac{\tilde{\alpha}_A^2\tilde{\rho}^2}{d}\|\beta_*\|^2.$$

H.2.4. TARGET ALIGNMENT

To understand this term, we first note that $\tilde{A}$ is independent of everything else. Hence we replace $\tilde{A}\tilde{A}^T$ with its expectation $\frac{\tilde{\rho}^2\tilde{n}}{d}I$.

$$-\mathbb{E}_{\tilde{A}}\left[\frac{2}{\tilde{n}}\tilde{\alpha}_A\beta_*^T\tilde{A}\tilde{A}^T\beta_{int}\right]=-\frac{2}{\tilde{n}}\frac{\tilde{\rho}^2\tilde{n}}{d}\tilde{\alpha}_A\beta_*^T\beta_{int}=-\frac{2\tilde{\alpha}_A\tilde{\rho}^2}{d}\beta_*^T\beta_{int}.$$

Since ε has mean-zero entries that are independent of everything else. We see that

$$\begin{aligned} \beta_*^T \beta_{int} &= \beta_*^T \left((\alpha_z \beta_*^T Z + \varepsilon^T)(Z+A)^\dagger + \alpha_A \beta_*^T A (Z+A)^\dagger \right)^T && (4) \\ &\overset{E}{=} \beta_*^T \left(\alpha_z \beta_*^T Z (Z+A)^\dagger - \alpha_A \beta_*^T A (Z+A)^\dagger \right)^T && (5) \\ &= \alpha_z \beta_*^T (Z+A)^{\dagger T} Z^T \beta_* + \alpha_A \beta_*^T (Z+A)^{\dagger T} A^T \beta_*. && (6) \end{aligned}$$

From Theorem 26, we have that

$$\mathbb{E}\left[\beta_*^T (Z+A)^{\dagger T} Z^T \beta_*\right] = \begin{cases} \frac{\eta^2 c (u^T\beta_*)^2}{\rho^2+\eta^2 c} + o\left(\frac{1}{\rho^2}\right) + O\left(\frac{1}{\rho n}\right), & c<1 \\ \frac{\eta^2}{\eta^2+\rho^2}(\beta_*^T u)^2 + o\left(\frac{1}{\rho^2}\right) + O\left(\frac{1}{n}\right), & c>1 \end{cases}.$$

and from Theorem 27, we have that

$$\mathbb{E}\left[\beta_*^T (Z+A)^{\dagger T} A^T \beta_*\right] = \begin{cases} \|\beta_*\|^2 - \frac{\eta^2 c (u^T\beta_*)^2}{\rho^2+\eta^2 c} + o\left(\frac{1}{\rho^2}\right) + O\left(\frac{1}{\rho n}\right), & c<1 \\ \frac{\|\beta_*\|^2}{c} - \frac{\eta^2 \|\beta_*\|^2}{d(\eta^2+\rho^2)} - \frac{\eta^2 (u^T\beta_*)^2}{c(\eta^2+\rho^2)} + o\left(\frac{1}{\rho^2} + \frac{1}{n}\right) + +O\left(\frac{1}{\rho n}\right), & c>1 \end{cases}.$$

Thus for $c<1$, the entire interaction term now becomes

$$\begin{aligned} -\mathbb{E}\left[\frac{2}{\tilde{n}} \tilde{\alpha}_A \beta_*^T \tilde{A}\tilde{A}^T \beta_{int}\right] &= -\frac{2\tilde{\alpha}_A \tilde{\rho}^2}{d} \beta_*^T \beta_{int} \\ &= -\frac{2\tilde{\alpha}_A \tilde{\rho}^2}{d} \left(\alpha_z \frac{\eta^2 c (u^T\beta_*)^2}{\rho^2+\eta^2 c} + \alpha_A \left[\|\beta_*\|^2 - \frac{\eta^2 c (u^T\beta_*)^2}{\rho^2+\eta^2 c} \right] + o\left(\frac{1}{\rho^2}\right) + O\left(\frac{1}{\rho n}\right) \right) \\ &= -\frac{2\tilde{\alpha}_A \tilde{\rho}^2}{d} \left((\alpha_z - \alpha_A) \frac{\eta^2 c}{\rho^2+\eta^2 c} (\beta_*^T u)^2 + \alpha_A \|\beta_*\|^2 \right) + o\left(\frac{1}{n}\right) + O\left(\frac{\rho}{n^2}\right). \end{aligned}$$

For $c>1$, instead we have

$$\begin{aligned} -\mathbb{E}\left[\frac{2}{\tilde{n}} \tilde{\alpha}_A \beta_*^T \tilde{A}\tilde{A}^T \beta_{int}\right] &= -\frac{2\tilde{\alpha}_A \tilde{\rho}^2}{d} \beta_*^T \beta_{int} \\ &= -\frac{2\tilde{\alpha}_A \tilde{\rho}^2}{d} \left(\alpha_z \frac{\eta^2}{\eta^2+\rho^2} (\beta_*^T u)^2 + O\left(\frac{1}{n}\right) \right. \\ &\qquad \left. + \alpha_A \left[\frac{\|\beta_*\|^2}{c} - \frac{\eta^2 \|\beta_*\|^2}{d(\eta^2+\rho^2)} - \frac{\eta^2 (u^T\beta_*)^2}{c(\eta^2+\rho^2)} \right] + o\left(\frac{1}{\rho^2}\right) \right) \\ &= -\frac{2\tilde{\alpha}_A \tilde{\rho}^2}{d} \left(\left(\alpha_z - \frac{\alpha_A}{c} \right) \frac{\eta^2}{\rho^2+\eta^2} (\beta_*^T u)^2 \right. \\ &\qquad \left. + \alpha_A \|\beta_*\|^2 \left(\frac{1}{c} - \frac{\eta^2}{d(\eta^2+\rho^2)} \right) \right) + o\left(\frac{1}{n}\right) + O\left(\frac{\rho^2}{n^2}\right). \end{aligned}$$

We perform a change of variables $\rho = \tau\sqrt{d}$, $\tilde{\rho} = \tilde{\tau}\sqrt{d}$, $\eta = \theta\sqrt{n}$, $\tilde{\eta} = \tilde{\theta}\sqrt{\tilde{n}}$ and the result follows from $d/n \to c$. Note our biggest error terms are $o(1) + O\left(\frac{1}{n}\right)$.

Appendix I. Proof of Specific Cases and Overfitting

I.1. Proof of Theorem 2.

Proof We set $\alpha_Z = \alpha_A = \tilde{\alpha}_Z = \tilde{\alpha}_A = \alpha$, $\tilde{\theta} = \theta$, $\tilde{\tau} = \tau$ in the above Theorem 7 and note that it greatly simplifies each term. Algebra shows that for $c < 1$

$$\text{Bias} = \tau_\varepsilon^2 \frac{c}{1-c} \frac{\theta^2}{d(\theta^2+\tau^2)}, \quad \text{Variance} = \alpha^2\tau^2\|\beta_*\|^2 + \tau_\varepsilon^2 \frac{c}{1-c}\left[1 - \frac{\theta^2}{d(\theta^2+\tau^2)}\right],$$

$$\text{Data Noise} = \alpha^2\tau^2\|\beta_*\|^2, \quad \text{Target Alignment} = -2\alpha^2\tau^2\|\beta_*\|^2,$$

While for $c > 1$, we can first send $d, n \to \infty$ and many terms become asymptotically 0. In the end, we get that:

$$\text{Bias} = \alpha^2\theta^2(\beta_*^T u)^2 \left(1 - \frac{1}{c}\right)^2 \left(\frac{\tau^2 c}{\theta^2 + \tau^2 c}\right)^2, \quad \text{Data Noise} = \alpha^2\tau^2\|\beta_*\|^2,$$

$$\text{Variance} = \alpha^2\tau^2\|\beta_*\|^2 \frac{1}{c} + \alpha^2\tau^2(\beta_*^T u)^2 \frac{\theta^2}{\theta^2+\tau^2 c}\left(1 - \frac{1}{c}\right) + \tau_\varepsilon^2 \frac{1}{c-1}.$$

$$\text{Target Alignment} = -2\alpha^2\tau^2 \left(\left(1 - \frac{1}{c}\right)\frac{\theta^2}{\theta^2+\tau^2 c}(\beta_*^T u)^2 + \|\beta_*\|^2 \frac{1}{c}\right),$$

Adding these terms together, we see with simple algebra that many terms cancel or can be combined, establishing the stated formula. ■

I.2. Proof of Theorem 4.

Proof We set $\alpha_Z = \tilde{\alpha}_Z$, $\alpha_A = \tilde{\alpha}_A$, $\tilde{\theta} = \theta$, $\tilde{\tau} = \tau$, and send $d, n \to \infty$ in Theorem 7. Recall that $\Delta_c = \alpha_Z - \frac{\alpha_A}{c}$ and $\Delta_1 = \alpha_Z - \alpha_A$. Then some algebra shows that for $c < 1$,

$$\text{Bias} = \theta^2(\beta_*^T u)^2 \Delta_1^2 \left(\frac{\tau^2}{\theta^2+\tau^2}\right)^2, \quad \text{Data Noise} = \alpha_A^2\tau^2\|\beta_*\|^2,$$

$$\text{Target Alignment} = -2\alpha_A^2\tau^2\|\beta_*\|^2 \; - 2\alpha_A\tau^2(\beta_*^T u)^2 \Delta_1 \frac{\theta^2}{\theta^2+\tau^2},$$

$$\text{Variance} = \alpha_A^2\tau^2\|\beta_*\|^2 + \tau_\varepsilon^2 \frac{c}{1-c} + \tau^2(\beta_*^T u)^2 \left[\frac{1}{1-c}\frac{\theta^4 + \theta^2\tau^2 c}{(\theta^2+\tau^2)^2}\Delta_1^2 + 2\alpha_A\Delta_1 \frac{\theta^2}{\theta^2+\tau^2}\right].$$

For $c > 1$, we have that

$$\text{Bias} = \theta^2(\beta_*^T u)^2 \Delta_c^2 \left(\frac{\tau^2 c}{\theta^2+\tau^2 c}\right)^2, \quad \text{Data Noise} = \alpha_A^2\tau^2\|\beta_*\|^2,$$

$$\text{Target Alignment} = -2\alpha_A^2\tau^2 \frac{\|\beta_*\|^2}{c} - 2\alpha_A\tau^2(\beta_*^T u)^2 \Delta_c \frac{\theta^2}{\theta^2+\tau^2 c},$$

$$\text{Variance} = \alpha_A^2\tau^2 \frac{\|\beta_*\|^2}{c} + \tau_\varepsilon^2 \frac{1}{c-1} + \tau^2(\beta_*^T u)^2 \frac{c}{1-c}\frac{\theta^2}{\theta^2+\tau^2 c}\Delta_c^2.$$

We proceed by adding these terms together and the results follow from algebra. ■

I.3. Proof of Theorem 5.

Proof We set $\tilde{\theta} = \theta$ and $\tilde{\tau} = \tau$ in Theorem 7 and have the regime of equal operator norm $\theta^2 = \gamma\tau^2$. Since we are interested in the limit $c \to \infty$, we only consider the overparameterized case $c > 1$. We first take the limit $d, n \to \infty$ and have that:

$$\text{Bias} = \tau^2(\beta_*^T u)^2 \left(\sqrt{\gamma}(\tilde{\alpha}_Z - \alpha_Z) + \left(\alpha_Z - \frac{\alpha_A}{c}\right)\frac{c\sqrt{\gamma}}{\gamma + c}\right)^2, \quad \text{Data Noise} = \tilde{\alpha}_A^2\tau^2\|\beta_*\|^2,$$

$$\text{Target Alignment} = -2\tilde{\alpha}_A\tau^2\left(\left(\alpha_Z - \frac{\alpha_A}{c}\right)\frac{\gamma}{\gamma + c}(\beta_*^T u)^2 + \alpha_A\frac{\|\beta_*\|^2}{c}\right),$$

$$\text{Variance} = \tau^2\alpha_A^2\frac{\|\beta_*\|^2}{c} + \tau^2(\beta_*^T u)^2\frac{c}{(c-1)}\frac{\gamma}{\gamma + c}\left(\alpha_Z - \frac{\alpha_A}{c}\right)^2 + \tau_\varepsilon^2\left(\frac{1}{c-1}\right).$$

The rest follows from simple calculus: if $\tilde{\alpha}_Z \neq \alpha_Z$, $\gamma = \omega_c(1)$, and $\beta_*^T u \neq 0$, the bias will diverge and other terms are controlled, yielding catastrophic. If $\tilde{\alpha}_Z = \alpha_Z$, $\omega_c(1) \leq \gamma \leq o_c(c^2)$, and $\beta_*^T u \neq 0$, a similar thing happens. In other cases, all of these terms are controlled and become finite values in the limit $\lim_{c\to\infty}\mathcal{R}_c - \tau_\varepsilon^2$, giving us tempered overfitting.

$$\lim_{c\to\infty}\mathcal{R}_c = \begin{cases} \tilde{\alpha}_A^2\tau^2\|\beta_*\|^2 & \beta \perp u \\ \tau^2\left[\gamma\tilde{\alpha}_Z^2(\beta_*^T u)^2 + \tilde{\alpha}_A^2\|\beta_*\|^2\right] & \beta \not\perp u, \gamma = \Theta_c(1) \\ \infty & \alpha_Z \neq \tilde{\alpha}_Z, \beta_* \not\perp u, \gamma = \omega(1) \\ \infty & \alpha_Z = \tilde{\alpha}_Z, \beta_* \not\perp u, \omega(1) \leq \gamma \leq o(c^2) \\ \tau^2\left[\left(\frac{\phi}{(\phi+1)^2}\alpha_Z^2 - 2\tilde{\alpha}_A\alpha_Z\right)(\beta_*^T u)^2 + \alpha_A^2\|\beta_*\|^2\right] & \alpha_Z = \tilde{\alpha}_Z, \beta_* \not\perp u, \gamma = \phi c^2 \\ \tau^2\left[(\alpha_Z^2 - 2\tilde{\alpha}_A\alpha_Z)(\beta_*^T u)^2 + \alpha_A^2\|\beta_*\|^2\right] & \alpha_Z = \tilde{\alpha}_Z, \beta_* \not\perp u, \gamma = \omega(c^2) \end{cases}$$

■

I.4. Proof of Theorem 6.

Proof We start with the first part and assume that $\alpha_Z \neq \tilde{\alpha}_Z$. Similarly, we have that $\tilde{\theta} = \theta$ and $\tilde{\tau} = \tau$ in Theorem 7. To achieve equal Frobenius norm, we set $\theta^2 = d\tau^2$ and send $d, n \to \infty$ so several terms would vanish.

In particular, for $c < 1$, we have that

$$\text{Bias} = \theta^2(\beta_*^T u)^2\left(\tilde{\alpha}_Z - \alpha_Z + (\alpha_Z - \alpha_A)\frac{\tau^2}{\theta^2 + \tau^2}\right)^2 = \tau^2(\beta_*^T u)^2\left(\sqrt{d}(\tilde{\alpha}_Z - \alpha_Z) + (\alpha_Z - \alpha_A)\frac{\sqrt{d}}{d+1}\right)^2,$$

It is clear that this term becomes ∞ since the term inside the parentheses scales with d. Note that the variance and data noise are non-negative, and target alignment is controlled. We have that $\mathcal{R}_c = \infty$ for $c \in (0, 1)$.

For $c > 1$, the same logic follows, and we also note that:

$$\text{Bias} = \theta^2(\beta_*^T u)^2\left(\tilde{\alpha}_Z - \alpha_Z + \left(\alpha_Z - \frac{\alpha_A}{c}\right)\frac{\tau^2 c}{\theta^2 + \tau^2 c}\right)^2 = \tau^2(\beta_*^T u)^2\left(\sqrt{d}(\tilde{\alpha}_Z - \alpha_Z) + \left(\alpha_Z - \frac{\alpha_A}{c}\right)\frac{\sqrt{d}c}{d+c}\right)^2,$$

which scales with d with other terms controlled. Hence, $\mathcal{R}_c = \infty$ for all $c \neq 1$.

Now assume that $\alpha_Z = \tilde{\alpha}_Z$. Since we are interested in $c \to \infty$, we only consider $c > 1$. First, from algebra and taking the limit for d, n, we have that:

$$\text{Bias} = \tau^2(\beta_*^T u)^2 \left(\left(\alpha_Z - \frac{\alpha_A}{c} \right) \frac{c\sqrt{d}}{d+c} \right)^2 \to 0, \quad \text{Data Noise} = \tilde{\alpha}_A^2 \tau^2 \|\beta_*\|^2,$$

$$\text{Target Alignment} = -2\tilde{\alpha}_A \tau^2 \left(\left(\alpha_Z - \frac{\alpha_A}{c} \right) (\beta_*^T u)^2 + \alpha_A \frac{\|\beta_*\|^2}{c} \right),$$

$$\text{Variance} = \tau^2 \alpha_A^2 \frac{\|\beta_*\|^2}{c} + \tau^2(\beta_*^T u)^2 \frac{c}{(c-1)} \left(\alpha_Z - \frac{\alpha_A}{c} \right)^2 + \tau_\varepsilon^2 \left(\frac{1}{c-1} \right).$$

We now take $c \to \infty$ and many terms vanish in this limit, yielding:

$$\lim_{c\to\infty} \mathcal{R}_c = -2\tilde{\alpha}_A \alpha_Z \tau^2 (\beta_*^T u)^2 + \tau^2 (\beta_*^T u)^2 \alpha_Z^2 + \tilde{\alpha}_A^2 \tau^2 \|\beta_*\|^2 = \tau^2 \left[(\beta_*^T u)^2 (\alpha_Z^2 - 2\tilde{\alpha}_A \alpha_Z) + \|\beta_*\|^2 \tilde{\alpha}_A^2 \right].$$

■

Proposition 1 (Non–existence of a canceling scale parameter) *Let $\alpha_A, \alpha_Z > 0$ be fixed scalars, let $u, \beta_* \in \mathbb{R}^d$ be fixed vectors, and set*

$$a := \|\beta_*\|^2 > 0, \qquad b := \left(\beta_*^\top u\right)^2 \in [0, a].$$

For every positive real number ϕ define

$$f(\phi) = \alpha_A^2 a + \left(\alpha_Z^2 \left(1 + \frac{1}{\phi}\right) - 2\alpha_Z \alpha_A \right) b.$$

Then

$$f(\phi) > 0 \quad \textit{for all } \phi > 0.$$

Consequently the equation $f(\phi) = 0$ has no solution with $\phi \in (0, \infty)$.

Proof If $b = 0$ (i.e. β_* is orthogonal to u) we have $f(\phi) = \alpha_A^2 a > 0$, so no positive ϕ can cancel the expression. Hence assume $b > 0$.

Writing $r := b/a \in (0, 1]$ we obtain

$$f(\phi) = a\Big[\alpha_A^2 + \alpha_Z(\alpha_Z - 2\alpha_A)\, r + \frac{\alpha_Z^2 r}{\phi}\Big]. \tag{$*$}$$

Since $r \leq 1$,

$$\alpha_A^2 + \alpha_Z(\alpha_Z - 2\alpha_A)\, r \;\geq\; \alpha_A^2 + \alpha_Z(\alpha_Z - 2\alpha_A) \;=\; \left(\alpha_A - \alpha_Z\right)^2 \;\geq 0.$$

Thus the square bracket in $(*)$ is the sum of a non–negative term and a strictly positive term.

■

Appendix J. Helper Results

In this section, we detail helpful lemmas that we will need according to the η, ρ scalings.

J.1. Results from prior work

We begin by recalling results from prior work. We state them here for completeness.

Theorem 8 (Theorems 3, 5 of [25]) *Define the following helper functions* $h = v^T A^\dagger$, $k = A^\dagger u$, $t = v^T(I - A^\dagger A)$, $\xi = 1 + \eta v^T A^\dagger u$, $s = (I - AA^\dagger)u$, $\gamma_1 = \eta^2\|t\|^2\|k\|^2 + \xi^2$, $\gamma_2 = \eta^2\|s\|^2\|h\|^2 + \xi^2$ *and*

$$p_1 = -\frac{\eta^2\|k\|^2}{\xi}t^T - \eta k, \qquad q_1^T = -\frac{\eta\|t\|^2}{\xi}k^T A^\dagger - h.$$
$$p_2 = -\frac{\eta^2\|s\|^2}{\xi}A^\dagger h^T - \eta k, \qquad q_2^T = -\frac{\eta\|h\|^2}{\xi}s^T - h,$$

Then we have that

$$(Z + A)^\dagger = \begin{cases} A^\dagger + \frac{\eta}{\xi}t^T k^T A^\dagger - \frac{\xi}{\gamma_1}p_1 q_1^T, & c < 1 \\ A^\dagger + \frac{\eta}{\xi}A^\dagger h^T s^T - \frac{\xi}{\gamma_2}p_2 q_2^T, & c > 1 \end{cases}.$$

Proposition 2 (Proposition 2 from [31]) *In the setting from Section 2*

$$Z(Z + A)^\dagger = \begin{cases} \frac{\eta\xi}{\gamma_1}uh + \frac{\eta^2\|t\|^2}{\gamma_1}uk^T A^\dagger, & c < 1 \\ \frac{\eta\xi}{\gamma_2}uh + \frac{\eta^2\|h\|^2}{\gamma_2}us^T, & c > 1 \end{cases}.$$

In addition to the above linear algebra results, we also need some random matrix theory estimates from prior work.

Lemma 9 (Lemma 7 from [31]) *In the setting of Section 2, we have that*

1. $\mathbb{E}[\|h\|^2] = \begin{cases} \frac{1}{\rho^2}\frac{c^2}{1-c} & c < 1 \\ \frac{1}{\rho^2}\frac{c}{c-1} & c > 1 \end{cases} + o\left(\frac{1}{\rho^2}\right)$ *and* $\mathrm{Var}(\|h\|^2) = O\left(\frac{1}{\rho^4 n}\right)$
2. $\mathbb{E}[\|k\|^2] = \frac{1}{\rho^2}\frac{c}{1-c} + o\left(\frac{1}{\rho^2}\right)$ *and* $\mathrm{Var}(\|k\|^2) = O\left(\frac{1}{\rho^4 n}\right)$
3. $\mathbb{E}[\|s\|^2] = 1 - \frac{1}{c}$ *and* $\mathrm{Var}(\|s\|^2) = O\left(\frac{1}{n}\right)$
4. $\mathbb{E}[\|t\|^2] = 1 - c$ *and* $\mathrm{Var}(\|t\|^2) = O\left(\frac{1}{n}\right)$
5. $\mathbb{E}\left[\frac{\xi}{\eta}\right] = \frac{1}{\eta}$ *and* $\mathrm{Var}\left(\frac{\xi}{\eta}\right) = O\left(\frac{1}{\max(n,d)}\frac{1}{\rho^2}\frac{c}{|1-c|}\right)$
6. $\mathbb{E}\left[\frac{\xi^2}{\eta^2}\right] = \frac{1}{\eta^2} + \frac{1}{\max(n,d)}\frac{1}{\rho^2}\frac{c}{|1-c|} + o\left(\frac{1}{\max(n,d)\rho^2}\right) = \frac{1}{\eta^2} + O\left(\frac{1}{\max(n,d)\rho^2}\right)$ *and* $\mathrm{Var}\left(\frac{\xi^2}{\eta^2}\right) = O\left(\frac{1}{\max(d,n)^2\rho^4}\right)$.

Proof Let $\zeta = \xi/\eta = 1/\eta + v^T A^\dagger u$. With $A = U\Sigma V^T$ (SVD), $A \in \mathbb{R}^{d\times n}$ having i.i.d. $\mathcal{N}(0, \rho^2/d)$ entries, and u, v fixed unit vectors, we have $\zeta = \sum_{i=1}^r \frac{1}{\sigma_i} b_i a_i$. Here $r = \min(d, n)$, $a = V^T v$, $b = U^T u$ are uniformly random on S^{n-1} and S^{d-1} respectively.

The fourth moment is $\mathbb{E}[\zeta^4] = \sum_{i,j,k,l} \mathbb{E}\left[\frac{1}{\sigma_i\sigma_j\sigma_k\sigma_l}\right] \mathbb{E}[b_ib_jb_kb_l]\mathbb{E}[a_ia_ja_ka_l]$. Non-zero terms require paired indices. Using exact spherical moments $\mathbb{E}[\zeta_i^4] = \frac{3}{M(M+2)}$ and $\mathbb{E}[\zeta_i^2 x_k^2] = \frac{1}{M(M+2)}$ $(i \neq k)$ for $x \in S^{M-1}$:

$$\begin{aligned}\mathbb{E}[\zeta^4] &= \sum_{i=1}^r \mathbb{E}\left[\frac{1}{\sigma_i^4}\right] \frac{9}{d(d+2)n(n+2)} + 3\sum_{i\neq k} \mathbb{E}\left[\frac{1}{\sigma_i^2\sigma_k^2}\right] \frac{1}{d(d+2)n(n+2)} \\ &= \underbrace{\frac{9S_4}{d(d+2)n(n+2)}}_{\text{Term 1}} + \underbrace{\frac{3\sum_{i\neq k}\mathbb{E}[1/(\sigma_i^2\sigma_k^2)]}{d(d+2)n(n+2)}}_{\text{Term 2}}\end{aligned}$$

where $S_4 = \sum_{i=1}^r \mathbb{E}[1/\sigma_i^4]$.

Leading Order Scaling: Let $N = \max(d, n)$, assume $n, d \to \infty$ with $d/n \to c \neq 1$. Lemma 5 from [31] tells

$$\mathbb{E}[1/\sigma_i^4] = O(1/\rho^4)$$

and

$$\mathbb{E}[1/(\sigma_i^2\sigma_k^2)] = O(1/\rho^4).$$

Term 1 has $r = \min(d, n)$ summands, hence

$$\text{Term 1} \sim O\left(\frac{r}{N^4\rho^4}\right) \sim O\left(\frac{1}{N^3\rho^4}\right).$$

Term 2 has $r(r-1) \approx r^2$ summands, hence

$$\text{Term 2} \sim O\left(\frac{r^2}{N^4\rho^4}\right) \sim O\left(\frac{1}{N^2\rho^4}\right).$$

Term 2 dominates. Thus,

$$\mathbb{E}[\zeta^4] = O\left(\frac{1}{\max(d,n)^2\rho^4}\right).$$

Variance $\mathrm{Var}(\zeta^2)$**:** $\mathrm{Var}(\zeta^2) = \mathbb{E}[\zeta^4] - (\mathbb{E}[\zeta^2])^2$. Let $S_2 = \sum_{i=1}^r \mathbb{E}[1/\sigma_i^2]$, then $(\mathbb{E}[\zeta^2])^2 = S_2^2/(d^2n^2)$. Neglecting the lower order Term 1 in $\mathbb{E}[\zeta^4]$ and approximating denominators $d(d+2) \approx d^2, n(n+2) \approx n^2$:

$$\mathrm{Var}(\zeta^2) \approx \frac{3\sum_{i\neq k}\mathbb{E}[1/(\sigma_i^2\sigma_k^2)]}{d^2n^2} - \frac{S_2^2}{d^2n^2} = \frac{1}{d^2n^2}\left[3\sum_{i\neq k}\mathbb{E}\left[\frac{1}{\sigma_i^2\sigma_k^2}\right] - S_2^2\right]$$

This is the leading order expression for the variance, which depends on the joint moments $\mathbb{E}[1/(\sigma_i^2\sigma_k^2)]$. The overall scaling is determined by the dominant terms:

$$\mathrm{Var}\left(\left(\frac{\xi}{\eta}\right)^2\right) = O\left(\frac{1}{\max(d,n)^2\rho^4}\right)$$

■

J.2. New Linear Algebra Calculations

Lemma 10 *If $\xi \neq 0$ and A has full rank, we have:*

$$\varepsilon^T(Z+A)^\dagger \tilde{Z} = \begin{cases} -\frac{\tilde{\eta}\xi}{\eta\gamma_1}\varepsilon^T p_1 \tilde{v}^T & c<1 \\ -\frac{\tilde{\eta}\xi}{\eta\gamma_2}\varepsilon^T p_2 \tilde{v}^T & c>1 \end{cases}.$$

Proof After substitutions, for $c < 1$ using Proposition 2 $\varepsilon^T(Z+A)^\dagger\tilde{Z}$ becomes:

$$\begin{aligned} &\varepsilon^T\left(A^\dagger + \frac{\eta}{\xi}t^T k^T A^\dagger - \frac{\xi}{\gamma_1}p_1\left(-\frac{\eta\|t\|^2}{\xi}k^T A^\dagger - h\right)\right)\tilde{Z} \\ &= \tilde{\eta}\varepsilon^T\left(A^\dagger u\tilde{v}^T + \frac{\eta}{\xi}t^T k^T A^\dagger u\tilde{v}^T - \frac{\xi}{\gamma_1}p_1\left(-\frac{\eta\|t\|^2}{\xi}k^T A^\dagger u - hu\right)\tilde{v}^T\right). \end{aligned}$$

Since $k = A^\dagger u$ and $hu = v^T A^\dagger u = \frac{\xi-1}{\eta}$, we then have that

$$\begin{aligned} &\tilde{\eta}\varepsilon^T\left(A^\dagger u\tilde{v}^T + \frac{\eta}{\xi}t^T k^T A^\dagger u\tilde{v}^T - \frac{\xi}{\gamma_1}p_1\left(-\frac{\eta\|t\|^2}{\xi}k^T A^\dagger u - hu\right)\tilde{v}^T\right) \\ &= \tilde{\eta}\varepsilon^T\left(k\tilde{v}^T + \frac{\eta\|k\|^2}{\xi}t^T\tilde{v}^T + \frac{\xi}{\gamma_1}p_1\left(\frac{\eta^2\|t\|^2\|k\|^2+\xi^2-\xi}{\xi\eta}\right)\tilde{v}^T\right) \\ &= \tilde{\eta}\varepsilon^T\left(k\tilde{v}^T + \frac{\eta\|k\|^2}{\xi}t^T\tilde{v}^T + \frac{1}{\gamma_1}p_1\left(\frac{\gamma_1-\xi}{\eta}\right)v_{tst}^T\right) \\ &= \tilde{\eta}\varepsilon^T\left(\frac{1}{\eta}\left(\frac{\eta^2\|k\|^2}{\xi}t^T + \eta k\right)\tilde{v}^T + \frac{1}{\eta}p_1\tilde{v}^T - \frac{\xi}{\eta\gamma_1}p_1\tilde{v}^T\right) \\ &= \varepsilon^T\left(-\frac{\tilde{\eta}}{\eta}p_1\tilde{v}^T + \frac{\tilde{\eta}}{\eta}p_1\tilde{v}^T - \frac{\tilde{\eta}\xi}{\eta\gamma_1}p_1\tilde{v}^T\right) \\ &= -\frac{\tilde{\eta}\xi}{\eta\gamma_1}\varepsilon^T p_1\tilde{v}^T, \end{aligned}$$

For $c > 1$, we note that the calculation is exactly the same. An example of such a calculation can be seen in the proof of Theorem 13. ■

Lemma 11 *In the setting of Section 2, we have:*

$$A(Z+A)^\dagger = \begin{cases} I - \frac{\eta\xi}{\gamma_1}uh + \frac{\eta^2\|t\|^2}{\gamma_1}uk^T A^\dagger, & c<1 \\ AA^\dagger + \frac{\eta\xi}{\gamma_2}h^T s^T - \frac{\eta^2\|s\|^2}{\gamma_2}h^T h - \frac{\eta^2\|h\|^2}{\gamma_2}AA^\dagger u s^T - \frac{\eta\xi}{\gamma_2}AA^\dagger uh, & c>1 \end{cases}.$$

Proof For $c < 1$, we see that Z, A are $d \times n$ with $d < n$. Then since A is assumed to have full rank, $Z + A$ has full rank with probability 1, and hence

$$(Z+A)(Z+A)^\dagger = I.$$

Thus, from Proposition 2,

$$A(Z+A)^\dagger = (Z+A)(Z+A)^\dagger - Z(Z+A)^\dagger = I - \frac{\eta\xi}{\gamma_1}uh - \frac{\eta^2\|t\|^2}{\gamma_1}uk^T A^\dagger.$$

For $c > 1$, since $(Z + A)(Z + A)^\dagger$ is no longer the identity matrix, we instead note $AA^\dagger$ and directly expand using Theorem 8:

$$\begin{aligned}A(Z+A)^\dagger &= A\left(A^\dagger + \frac{\eta}{\xi}A^\dagger h^T s^T - \frac{\xi}{\gamma_2}\left(\frac{\eta^2\|s\|^2}{\xi}A^\dagger h^T + \eta k\right)\left(\frac{\eta\|h\|^2}{\xi}s^T + h\right)\right)\\ &= AA^\dagger + \frac{\eta}{\xi}AA^\dagger h^T s^T - \frac{\xi}{\gamma_2}\left(\frac{\eta^2\|s\|^2}{\xi}AA^\dagger h^T + \eta AA^\dagger u\right)\left(\frac{\eta\|h\|^2}{\xi}s^T + h\right)\end{aligned}$$

Noting that $AA^\dagger h^T = AA^\dagger A^{\dagger T} v = A^{\dagger T} v = h^T$, we see that

$$\begin{aligned}A(Z+A)^\dagger &= AA^\dagger + \frac{\eta}{\xi}h^T s^T - \frac{\xi}{\gamma_2}\left(\frac{\eta^2\|s\|^2}{\xi}h^T + \eta AA^\dagger u\right)\left(\frac{\eta\|h\|^2}{\xi}s^T + h\right)\\ &= AA^\dagger + \frac{\eta}{\xi}h^T s^T - \frac{\eta^3\|s\|^2\|h\|^2}{\xi\gamma_2}h^T s^T - \frac{\eta^2\|s\|^2}{\gamma_2}h^T h - \frac{\eta^2\|h\|^2}{\gamma_2}AA^\dagger u s^T - \frac{\eta\xi}{\gamma_2}AA^\dagger u h.\end{aligned}$$

We can combine the coefficients in front of $h^T s^T$ to get

$$\frac{\eta}{\xi} - \frac{\eta^3\|s\|^2\|h\|^2}{\xi\gamma_2} = \frac{\eta(\eta^2\|s\|^2\|h\|^2 + \xi^2) - \eta^3\|s\|^2\|h\|^2}{\xi\gamma_2} = \frac{\eta\xi}{\gamma_2}.$$

The Lemma statement follows from here. ■

Lemma 12 *If $\xi \neq 0$ and A has full rank, we have:*

$$\beta_*^T Z(Z+A)^\dagger \tilde{Z} = \begin{cases}\left(1 - \frac{\xi}{\gamma_1}\right)\beta_*^T \tilde{Z} & c < 1\\ \left(1 - \frac{\xi}{\gamma_2}\right)\beta_*^T \tilde{Z} & c > 1\end{cases},$$

Proof Using Proposition 2 for $c < 1$, we get that

$$\begin{aligned}\beta_*^T Z(Z+A)^\dagger \tilde{Z} &= \beta_*^T\left(\frac{\eta\xi}{\gamma_1}uh + \frac{\eta^2\|t\|^2}{\gamma_1}uk^T A^\dagger\right)\tilde{Z}\\ &= \tilde{\eta}\beta_*^T\left(\frac{\eta\xi}{\gamma_1}uhu\tilde{v}^T + \frac{\eta^2\|t\|^2}{\gamma_1}uk^T A^\dagger u\tilde{v}^T\right)\\ &= \tilde{\eta}\beta_*^T\left(\frac{\eta\xi}{\gamma_1}uv^T A^\dagger u\tilde{v}^T + \frac{\eta^2\|t\|^2}{\gamma_1}uk^T A^\dagger u\tilde{v}^T\right)\end{aligned}$$

Note $\xi - 1 = \eta v^T A^\dagger u$, $kA^\dagger u = k^T k = \|k\|^2$. The above equation becomes

$$\tilde{\eta}\beta_*^T\left(\frac{\xi(\xi-1)}{\gamma_1} + \frac{\eta^2\|t\|^2\|k\|^2}{\gamma_1}\right)u\tilde{v}^T = \beta_*^T\left(\frac{\xi(\xi-1)}{\gamma_1} + \frac{\eta^2\|t\|^2\|k\|^2}{\gamma_1}\right)\tilde{Z}^T.$$

Using $\gamma_1 = \eta^2\|t\|^2\|k\|^2 + \xi^2$ to combine the coefficients, we have that

$$\frac{\xi(\xi-1)}{\gamma_1} + \frac{\eta^2\|t\|^2\|k\|^2}{\gamma_1} = \frac{-\xi + \xi^2 + \eta^2\|t\|^2\|k\|^2}{\gamma_1} = \frac{-\xi + \gamma_1}{\gamma_1} = 1 - \frac{\xi}{\gamma_1}.$$

Similarly, for $c > 1$, we obtain

$$\begin{aligned}\beta_*^T Z(Z+A)^\dagger \tilde{Z} &= \beta_*^T \left(\frac{\eta\xi}{\gamma_2} uh + \frac{\eta^2 \|h\|^2}{\gamma_2} us^T \right) \tilde{Z} \\ &= \tilde{\eta}\beta_*^T \left(\frac{\eta\xi}{\gamma_2} uhu\tilde{v}^T + \frac{\eta^2 \|h\|^2}{\gamma_2} us^T u\tilde{v}^T \right) \\ &= \tilde{\eta}\beta_*^T \left(\frac{\eta\xi}{\gamma_2} uv^T A^\dagger u\tilde{v}^T + \frac{\eta^2 \|h\|^2}{\gamma_2} us^T u\tilde{v}^T \right)\end{aligned}$$

Note $\xi - 1 = \eta v^T A^\dagger u$, $s^T u = \|s\|^2$. The above equation becomes

$$\tilde{\eta}\beta_*^T \left(\frac{\xi(\xi-1)}{\gamma_2} + \frac{\eta^2 \|s\|^2 \|h\|^2}{\gamma_2} \right) u\tilde{v}^T = \beta_*^T \left(\frac{\xi(\xi-1)}{\gamma_2} + \frac{\eta^2 \|s\|^2 \|h\|^2}{\gamma_2} \right) \tilde{Z}^T.$$

Using $\gamma_2 = \eta^2 \|s\|^2 \|h\|^2 + \xi^2$ to combine the coefficients, we have that

$$\frac{\xi(\xi-1)}{\gamma_2} + \frac{\eta^2 \|s\|^2 \|h\|^2}{\gamma_2} = \frac{-\xi + \xi^2 + \eta^2 \|t\|^2 \|k\|^2}{\gamma_2} = \frac{-\xi + \gamma_2}{\gamma_2} = 1 - \frac{\xi}{\gamma_2}.$$

The target expression follows:

$$\left(1 - \frac{\xi}{\gamma_2}\right) \beta_*^T \tilde{Z}.$$

■

Lemma 13 *If $\xi \neq 0$ and A has full rank, we have:*

$$\beta_*^T A(Z+A)^\dagger \tilde{Z} = \begin{cases} \frac{\xi}{\gamma_1} \beta_*^T \tilde{Z} & c < 1 \\ \frac{\eta\|s\|^2}{\gamma_2} \beta_*^T h^T u^T \tilde{Z} + \frac{\xi}{\gamma_2} \beta_*^T A A^\dagger \tilde{Z}. & c > 1 \end{cases},$$

Proof We begin with $c < 1$. Then since A is assumed to have full rank. We see that $Z + A$ has full column rank with probability 1, and hence

$$(Z+A)(Z+A)^\dagger = I$$

It follows from Theorem 12 that

$$\begin{aligned}\beta_*^T A(Z+A)^\dagger \tilde{Z} &= \beta_*^T (Z+A)(Z+A)^\dagger \tilde{Z} - \beta_*^T Z(Z+A)^\dagger \tilde{Z} \\ &= \beta_*^T \tilde{Z} - \left(1 - \frac{\xi}{\gamma_1}\right) \beta_*^T \tilde{Z} = \frac{\xi}{\gamma_1} \beta_*^T \tilde{Z}.\end{aligned}$$

For $c > 1$, we no longer have that $Z + A$ is full column rank. It is now full row rank. Hence we do not have

$$(Z+A)(Z+A)^\dagger = I$$

hence, we directly expand it using Theorem 8 and its helper variables:

$$\begin{aligned}
\beta_*^T A(Z+A)^\dagger \tilde{Z} &= \beta_*^T A \left(A^\dagger + \frac{\eta}{\xi} A^\dagger h^T s^T - \frac{\xi}{\gamma_2} p_2 q_2^T \right) \tilde{Z} \\
&= \tilde{\eta} \beta_*^T A \left(k \tilde{v}^T + \frac{\eta \|s\|^2}{\xi} A^\dagger h^T \tilde{v}^T - \frac{\xi}{\gamma_2} p_2 q_2^T u \tilde{v}^T \right) \\
&= \tilde{\eta} \beta_*^T A \left(-\frac{1}{\eta} p_2 \tilde{v}^T - \frac{\xi}{\gamma_2} p_2 \left(-\frac{\eta \|h\|^2}{\xi} s^T - h \right) u \tilde{v}^T \right) \\
&= \tilde{\eta} \beta_*^T A \left(-\frac{1}{\eta} p_2 \tilde{v}^T + \frac{\xi}{\gamma_2} p_2 \left(\frac{\eta \|s\|^2 \|h\|^2}{\xi} + \frac{\xi - 1}{\eta} \right) \tilde{v}^T \right) \\
&= \tilde{\eta} \beta_*^T A \left(-\frac{1}{\eta} p_2 \tilde{v}^T + \frac{\xi}{\gamma_2} p_2 \left(\frac{\eta^2 \|s\|^2 \|h\|^2 + \xi^2 - \xi}{\xi \eta} \right) \tilde{v}^T \right) \\
&= \tilde{\eta} \beta_*^T A \left(-\frac{1}{\eta} p_2 \tilde{v}^T + \frac{\xi}{\gamma_2} p_2 \left(\frac{\gamma_2 - \xi}{\xi \eta} \right) \tilde{v}^T \right) \\
&= \tilde{\eta} \beta_*^T A \left(-\frac{1}{\eta} p_2 \tilde{v}^T + \frac{1}{\eta} p_2 \tilde{v}^T - \frac{\xi}{\eta \gamma_2} p_2 \tilde{v}^T \right) \\
&= -\frac{\tilde{\eta} \xi}{\eta \gamma_2} \beta_*^T A p_2 \tilde{v}^T \\
&= \frac{\tilde{\eta} \xi}{\eta \gamma_2} \beta_*^T \left(\frac{\eta^2 \|s\|^2}{\xi} h^T + \eta A k \right) \tilde{v}^T \\
&= \frac{\tilde{\eta} \eta \|s\|^2}{\gamma_2} \beta_*^T h^T \tilde{v}^T + \frac{\xi}{\gamma_2} \beta_*^T A A^\dagger \tilde{Z}.
\end{aligned}$$

Noting that $\beta_*^T h^T$ is a scalar, we then introduce $1 = u^T u$ and get that

$$\frac{\tilde{\eta} \eta \|s\|^2}{\gamma_2} \beta_*^T h^T u^T u \tilde{v}^T = \frac{\eta \|s\|^2}{\gamma_2} \beta_*^T h^T u^T \tilde{Z}.$$

Thus, the final expression is

$$\frac{\eta \|s\|^2}{\gamma_2} \beta_*^T h^T u^T \tilde{Z} + \frac{\xi}{\gamma_2} \beta_*^T A A^\dagger \tilde{Z}.$$

■

Lemma 14 (Bias Term) *In the setting of Section 2, we have that if $c < 1$*

$$\tilde{\alpha}_z \beta_*^T \tilde{Z} - \beta_{\mathit{int}}^T \tilde{Z} = \left[\tilde{\alpha}_Z - \alpha_Z + \frac{\xi}{\gamma_1} (\alpha_Z - \alpha_A) \right] \beta_*^T \tilde{Z} + \frac{\tilde{\eta}}{\eta} \frac{\xi}{\gamma_1} \varepsilon^T p_1 \tilde{v}^T$$

and if $c > 1$

$$\tilde{\alpha}_z \beta_*^T \tilde{Z} - \beta_{\mathit{int}}^T \tilde{Z} = \beta_*^T \left[(\tilde{\alpha}_Z - \alpha_Z) I + \frac{\xi}{\gamma_2} (\alpha_Z I - \alpha_A A A^\dagger) \right] \tilde{Z} - \alpha_A \frac{\eta \|s\|^2}{\gamma_2} \beta_*^T h^T u^T \tilde{Z} + \frac{\tilde{\eta}}{\eta} \frac{\xi}{\gamma_2} \varepsilon^T p_2 \tilde{v}^T$$

Proof To simplify the bias term, we first need to the following calculation.

$$\begin{aligned}\tilde{\alpha}_z\beta_*^T\tilde{Z} - \beta_{int}^T\tilde{Z} &= \tilde{\alpha}_z\beta_*^T\tilde{Z} - (\beta_*^T(\alpha_z Z + \alpha_A A) + \varepsilon^T)(Z+A)^\dagger\tilde{Z} \\ &= \tilde{\alpha}_z\beta_*^T\tilde{Z} - \alpha_z\beta_*^T Z(Z+A)^\dagger - \alpha_A\beta_*^T A(Z+A)^\dagger\tilde{Z} - \varepsilon^T(Z+A)^\dagger\tilde{Z}\end{aligned}$$

Using Theorem 12, we get

$$\beta_*^T Z(Z+A)^\dagger\tilde{Z} = \begin{cases}\left(1-\frac{\xi}{\gamma_1}\right)\beta_*^T\tilde{Z} & c<1 \\ \left(1-\frac{\xi}{\gamma_2}\right)\beta_*^T\tilde{Z} & c>1\end{cases},$$

Using Theorem 13, we get

$$\beta_*^T A(Z+A)^\dagger\tilde{Z} = \begin{cases}\frac{\xi}{\gamma_1}\beta_*^T\tilde{Z} & c<1 \\ \frac{\eta\|s\|^2}{\gamma_2}\beta_*^T h^T u^T\tilde{Z} + \frac{\xi}{\gamma_2}\beta_*^T AA^\dagger\tilde{Z}. & c>1\end{cases},$$

and using Theorem 10

$$\varepsilon^T(Z+A)^\dagger\tilde{Z} = \begin{cases}-\frac{\tilde{\eta}\xi}{\eta\gamma_1}\varepsilon^T p_1\tilde{v}^T & c<1 \\ -\frac{\tilde{\eta}\xi}{\eta\gamma_2}\varepsilon^T p_2\tilde{v}^T & c>1\end{cases}.$$

Thus, adding the terms together, we see that for $c<1$, we get

$$\tilde{\alpha}_Z\beta_*^T\tilde{Z} - \alpha_Z\left(1-\frac{\xi}{\gamma_1}\right)\beta_*^T\tilde{Z} - \alpha_A\frac{\xi}{\gamma_1}\beta_*^T\tilde{Z} + \frac{\tilde{\eta}}{\eta}\frac{\xi}{\gamma_1}\varepsilon^T p_1\tilde{v}^T$$

Collecting relevant terms together, we get

$$\left[\tilde{\alpha}_Z - \alpha_Z + \frac{\xi}{\gamma_1}(\alpha_Z - \alpha_A)\right]\beta_*^T\tilde{Z} + \frac{\tilde{\eta}}{\eta}\frac{\xi}{\gamma_1}\varepsilon^T p_1\tilde{v}^T$$

On the other hand for $c>1$, we have

$$\tilde{\alpha}_Z\beta_*^T\tilde{Z} - \alpha_Z\left(1-\frac{\xi}{\gamma_2}\right)\beta_*^T\tilde{Z} - \alpha_A\left[\frac{\eta\|s\|^2}{\gamma_2}\beta_*^T h^T u^T\tilde{Z} + \frac{\xi}{\gamma_2}\beta_*^T AA^\dagger\tilde{Z}\right] + \frac{\tilde{\eta}}{\eta}\frac{\xi}{\gamma_2}\varepsilon^T p_2\tilde{v}^T$$

Collecting relevant terms together, we get

$$\beta_*^T\left[(\tilde{\alpha}_Z - \alpha_Z)I + \frac{\xi}{\gamma_2}(\alpha_Z I - \alpha_A AA^\dagger)\right]\tilde{Z} - \alpha_A\frac{\eta\|s\|^2}{\gamma_2}\beta_*^T h^T u^T\tilde{Z} + \frac{\tilde{\eta}}{\eta}\frac{\xi}{\gamma_2}\varepsilon^T p_2\tilde{v}^T$$

■

Lemma 15 (Squared Norms of p_1 and p_2) *Let* $p_1 = -\frac{\eta^2\|k\|^2}{\xi}t - \eta k$ *and* $p_2 = -\frac{\eta^2\|s\|^2}{\xi}A^\dagger h - \eta k$.

1. $\|p_1\|^2 = \dfrac{\eta^2\|k\|^2}{\xi^2}\gamma_1$.

2. $\|p_2\|^2 = \frac{\eta^4\|s\|^4}{\xi^2}\|A^\dagger h^T\|^2 + \frac{2\eta^3\|s\|^2}{\xi}h(A^\dagger)^T k + \eta^2\|k\|^2$.

Proof For p_1:

$$\begin{aligned}\|p_1\|^2 &= \left(-\frac{\eta^2\|k\|^2}{\xi}t - \eta k^T\right)\left(-\frac{\eta^2\|k\|^2}{\xi}t^T - \eta k\right)\\ &= \left(\frac{\eta^2\|k\|^2}{\xi}\right)^2\|t\|^2 + 2\frac{\eta^3\|k\|^2}{\xi}(tk) + \eta^2\|k\|^2.\end{aligned}$$

Using $tk = 0$ yields the first result, we can further simplify as

$$\frac{\eta^2\|k\|^2}{\xi^2}\left(\eta^2\|k\|^2\|t\|^2 + \xi^2\right) = \frac{\eta^2\|k\|^2}{\xi^2}\gamma_1$$

For p_2:

$$\begin{aligned}\|p_2\|^2 &= \left(-\frac{\eta^2\|s\|^2}{\xi}(A^\dagger h^T)^T - \eta k^T\right)\left(-\frac{\eta^2\|s\|^2}{\xi}A^\dagger h^T - \eta k\right)\\ &= \left(\frac{\eta^2\|s\|^2}{\xi}\right)^2\|A^\dagger h^T\|^2 + 2\frac{\eta^3\|s\|^2}{\xi}(A^\dagger h^T)^T k + \eta^2\|k\|^2\\ &= \frac{\eta^4\|s\|^4}{\xi^2}\|A^\dagger h^T\|^2 + \frac{2\eta^3\|s\|^2}{\xi}h(A^\dagger)^T k + \eta^2\|k\|^2.\end{aligned}$$

■

Lemma 16 (Squared Norms of q_1 and q_2) *Let $q_1^T = -\frac{\eta\|t\|^2}{\xi}k^T A^\dagger - h$ and $q_2^T = -\frac{\eta\|h\|^2}{\xi}s^T - h$.*

1. $\|q_1\|^2 = \frac{\eta^2\|t\|^4}{\xi^2}kA^\dagger A^{\dagger T}k + \frac{2\eta\|t\|^2}{\xi}k^T A^\dagger h^T + \|h\|^2.$
2. $\|q_2\|^2 = \frac{\|h\|^2}{\xi^2}\gamma_2.$

Proof For q_1,

$$\|q_1\|^2 = \left(-\frac{\eta\|t\|^2}{\xi}k^T A^\dagger - h\right)\left(-\frac{\eta\|t\|^2}{\xi}A^{\dagger T}k - h^T\right) = \frac{\eta^2\|t\|^4}{\xi^2}kA^\dagger A^{\dagger T}k + \frac{2\eta\|t\|^2}{\xi}k^T A^\dagger h^T + \|h\|^2.$$

For q_2:

$$\begin{aligned}\|q_2\|^2 = \left(-\frac{\eta\|h\|^2}{\xi}s^T - h\right)\left(-\frac{\eta\|h\|^2}{\xi}s - h^T\right) &= \frac{\eta^2\|h\|^4\|s\|^2}{\xi^2} + \|h\|^2 \quad \text{since } hs = 0\\ &= \frac{\|h\|^2(\eta^2\|h\|^2\|s\|^2 + \xi^2)}{\xi^2}\\ &= \frac{\|h\|^2}{\xi^2}\gamma_2.\end{aligned}$$

■

Lemma 17 (Preliminary Expansion of Variance) *In the setting of Section 2, we have*

$$\mathbb{E}\left[\frac{1}{\tilde{n}}\left\|\beta_{int}^T\tilde{A}\right\|_F^2\right] = \mathbb{E}\left[\frac{\tilde{\rho}^2\alpha_z^2}{d}\beta_*^T Z(Z+A)^\dagger(Z+A)^{\dagger T}Z\beta_* + \frac{\tilde{\rho}^2\alpha_A^2}{d}\beta_*^T A(Z+A)^\dagger(Z+A)^{\dagger T}A^T\beta_* \right.$$
$$\left. + \frac{2\tilde{\rho}^2\alpha_A\alpha_z}{d}\beta_*^T Z(Z+A)^\dagger(Z+A)^{\dagger T}A^T\beta_* + \frac{\tilde{\rho}^2}{d}\varepsilon^T(Z+A)^\dagger(Z+A)^{\dagger T}\varepsilon\right]$$

Proof *Since $\tilde{A}$ is independent of the other terms, we replace $\tilde{A}\tilde{A}^T$ with its expectation $\frac{\tilde{\rho}^2\tilde{n}}{d}I$.*

$$\frac{1}{\tilde{n}}\left\|\beta_{int}^T\tilde{A}\right\|_F^2 = \frac{1}{\tilde{n}}\beta_{int}^T\tilde{A}\tilde{A}^T\beta_{int} \overset{E}{=} \frac{1}{\tilde{n}}\frac{\tilde{\rho}^2\tilde{n}}{d}\tilde{\alpha}_A\beta_{int}^T\beta_{int} = \frac{\tilde{\rho}^2}{d}\|\beta_{int}\|^2.$$

We now plug in the expression for β_{int}. Since ε is a zero-mean vector and independent from other random variables, terms with only one ε have zero expectation. A straightforward expansion gives:

$$\begin{aligned}\frac{\tilde{\rho}^2}{d}\|\beta_{int}\|_F^2 &= \frac{\tilde{\rho}^2}{d}(\beta_*^T(\alpha_z Z + \alpha_A A) + \varepsilon^T)(Z+A)^\dagger(Z+A)^{\dagger T}(\beta_*^T(\alpha_z Z + \alpha_A A) + \varepsilon^T)^T \\ &\overset{E}{=} \frac{\tilde{\rho}^2\alpha_z^2}{d}\beta_*^T Z(Z+A)^\dagger(Z+A)^{\dagger T}Z\beta_* + \frac{\tilde{\rho}^2}{d}\varepsilon^T(Z+A)^\dagger(Z+A)^{\dagger T}\varepsilon \\ &+ \frac{\tilde{\rho}^2\alpha_A^2}{d}\beta_*^T A(Z+A)^\dagger(Z+A)^{\dagger T}A^T\beta_* \\ &+ \frac{2\tilde{\rho}^2\alpha_A\alpha_z}{d}\beta_*^T Z(Z+A)^\dagger(Z+A)^{\dagger T}A^T\beta_*\end{aligned}$$

■

J.3. New Estimates

We now present new random matrix theory estimates.

Lemma 18 (General Terms) *In the setting of Section 2 we have the following expectations:*

1. *For $c < 1$, $\mathbb{E}[\beta_*^T uk^T A^\dagger\beta_*] = \frac{c}{\rho^2(1-c)}(\beta_*^T u)^2 + o\left(\frac{1}{\rho^2}\right)$ and the variance is $O(1/(\rho^2 d))$.*
2. *For $c < 1$, $\mathbb{E}[k^T A^\dagger A^{\dagger T}k] = \frac{c^2}{\rho^4(1-c)^3} + o\left(\frac{1}{\rho^4}\right)$ and the variance is $O(1/(\rho^4 d^2))$.*
3. *For $c > 1$, $\mathbb{E}[\beta_*^T su^T\beta_*] = \frac{c-1}{c}(\beta_*^T u)^2$ and the variance is $O(1/d)$.*
4. *For $c > 1$, $\mathbb{E}[\beta_*^T AA^\dagger us^T\beta_*] = \frac{c-1}{c^2}(\beta_*^T u)^2 + o(1/n)$ and the variance is $O(1/(\rho^2 d))$.*
5. *For $c > 1$, $\mathbb{E}[\beta_*^T h^T h\beta_*] = \frac{\|\beta_*\|^2}{d}\frac{c}{\rho^2(c-1)} + o\left(\frac{\|\beta_*\|^2}{d\rho^2}\right)$ and the variance is $O(1/(\rho^4 d^2))$.*
6. *For $c > 1$, $\mathbb{E}[\|A^\dagger h^T\|^2] = \frac{1}{\rho^4}\frac{c^3}{(c-1)^3} + o\left(\frac{1}{\rho^4}\right)$ and the variance is $O(1/(\rho^8 d))$.*
7. *For $c > 1$, $\mathbb{E}[\|k\|^2] = \frac{1}{\rho^2}\frac{1}{c-1} + o\left(\frac{1}{\rho^2}\right)$ and the variance is $O(1/(\rho^4 n))$*

Proof For all three terms we will need the SVD $A = U\Sigma V^T$, with $A^\dagger = V\Sigma^\dagger U^T$.

For the first term, we note that

$$\beta_*^T u k^T A^\dagger \beta_* = (\beta_*^T u) u^T A^{\dagger T} A^\dagger \beta_*$$

$$\begin{aligned} &= (\beta_*^T u) \sum_{i=1}^d (u^T U)_i (U^T \beta_*)_i \frac{1}{\sigma_i^2(A)} \\ &= (\beta_*^T u) \sum_{i=1}^d u^T u_i \cdot \beta_*^T u_i \frac{1}{\sigma_i^2(A)} \end{aligned}$$

where u_i is the $i$the column of U. Noting that

$$(u^T \beta_*) = (u^T U U^T \beta_*)$$

Since permuting columns of a orthogonal matrix does not break orthogonality and U is uniformly random, we have that the marginals u_i are identical. Thus, we see that

$$\mathbb{E}[u^T u_1 \cdot \beta_*^T u_1] = \ldots = \mathbb{E}[u^T u_d \cdot \beta_*^T u_d] = \frac{1}{d}(u^T \beta_*)$$

Thus using Lemma 5 part 5 from [31] along with convergence, we get that

$$\begin{aligned} \mathbb{E}\left[\beta_*^T u k^T A^\dagger \beta_*\right] &= (\beta_*^T u) \sum_{i=1}^d \mathbb{E}[(u^T u u^T \beta_*)_i] \mathbb{E}\left[\frac{1}{\rho^2 \sigma_i^2(A/\rho^2)}\right] \\ &= \frac{1}{\rho^2}(\beta_*^T u)^2 \sum_{i=1}^d \frac{1}{d}\left(\frac{c}{1-c} + o(1)\right) \\ &= \frac{1}{\rho^2}\frac{c}{1-c}(\beta_*^T u)^2 + o\left(\frac{1}{\rho^2}\right) \end{aligned}$$

Similar to calculation in Theorem 19, we see that the variance of this term is $O(1/(\rho^2 d))$.

For the second term, we have that

$$\begin{aligned} k^T A^\dagger A^{\dagger T} k &= u^T ((AA^T)^\dagger)^2 u \\ &= u^T U ((\Sigma\Sigma^T)^\dagger)^2 U^T u \\ &= \sum_{i=1}^d (u^T u_i)^2 \frac{1}{\sigma_i^4(A)} \end{aligned}$$

Then using Lemma 5 part 6 from [31], we have that

$$\mathbb{E}[k^T A^\dagger A^{\dagger T} k] = \sum_{i=1}^d \mathbb{E}[(u^T u_i)^2] \mathbb{E}\left[\frac{1}{\sigma_i^4(A)}\right] = \sum_{i=1}^d \frac{1}{\rho^4}\frac{1}{d}\left(\frac{c^2}{(1-c)^3} + o(1)\right) = \frac{1}{\rho^4}\frac{c^2}{(1-c)^3} + o\left(\frac{1}{\rho^4}\right)$$

Similarly, we see that the variance is $O(1/(\rho^4 d^2))$.

For third term, we have that

$$\begin{aligned}\beta_*^T s u^T \beta_* &= \beta_*^T (I - AA^\dagger) u \cdot (u^T \beta_*) \\ &= (\beta_*^T u)^2 - \sum_{i=1}^{n} (\beta_*^T u_i)(u^T u_i)\end{aligned}$$

Taking the expectation, we get

$$(\beta_*^T u)^2 \left[1 - \sum_{i=1}^{n} \frac{1}{d}\right] = \left(1 - \frac{1}{c}\right)(\beta_*^T u)^2$$

the variance for this term is $O(1/d)$.

For the fourth term, we have that

$$\beta_*^T AA^\dagger u s^T \beta_* = \beta_*^T AA^\dagger u u^T \beta_* - (\beta_*^T AA^\dagger u)^2$$

From previous calculations, we have that

$$\mathbb{E}[\beta_*^T AA^\dagger u] = (\beta_*^T u)\mathbb{E}[\beta_*^T AA^\dagger u] = \frac{1}{c}(\beta_*^T u)^2$$

Using Proposition 3, we see that

$$\mathbb{E}[(\beta_*^T AA^\dagger u)^2] = \frac{1}{c^2}(\beta_*^T u)^2 + o(1)$$

Then, we get

$$\mathbb{E}[\beta_*^T AA^\dagger u s^T \beta_*] = \frac{c-1}{c^2}(\beta_*^T u)^2 + o(1)$$

The variance for this term is $O(1/(\rho^2 d))$.

For the fifth term, and final term, we have

$$\beta_*^T h^T h \beta_* = (\beta_*^T A^\dagger v)^2 = \sum_{i,j}^{n} (\beta_*^T U)_i (\beta_*^T U)_j \frac{1}{\sigma_i(A)\sigma_j(A)} (V^T v)_i (V^T v)_j$$

Since $\beta_*^T U$ is uniformly random and independent of everything else, we see that we only have the diagonal terms when we take the expectation. Using Lemma 5 part 1 from [31]

$$\mathbb{E}[\beta_*^T h^T h \beta_*] = \sum_{i=1}^{n} \frac{\|\beta_*\|^2}{d} \frac{1}{n} \frac{1}{\rho^2} \left(\frac{c}{c-1} + o(1)\right) = \frac{\|\beta_*\|^2}{d} \frac{1}{\rho^2} \frac{c}{c-1} + o\left(\frac{\|\beta_*\|^2}{d\rho^2}\right)$$

The variance for this term is $O(1/(d^2\rho^2))$.

For the sixth term we have that

$$\mathbb{E}\left[\|A^\dagger h^T\|^2\right] = \sum_{i=1}^{n} \frac{1}{n}\mathbb{E}\left[\frac{1}{\sigma_i^4}\right] = \frac{1}{\rho^4}\frac{c^3}{(c-1)^3} + o\left(\frac{1}{\rho^4}\right)$$

With a variance of $O(1/(\rho^8 d))$. ■

Lemma 19 (Zero Expectation) *In the setting of Section 2, we have the following expectations for*

1. $\forall c$, $\mathbb{E}[\beta_*^T uh\beta_*] = 0$ *and* $\mathrm{Var}(\beta_*^T uh\beta_*) = O(1/(\rho^2 d))$
2. *If* $c > 1$, $\mathbb{E}[\beta_*^T AA^\dagger uh\beta_*] = 0$ *and* $\mathrm{Var}(\beta_*^T AA^\dagger uh\beta_*) = O(1/(\rho^2 d^2))$
3. *If* $c > 1$, $\mathbb{E}[\beta_*^T sh\beta_*] = 0$ *and* $\mathrm{Var}(\beta_*^T sh\beta_*) = O(1/(\rho^2 d))$
4. $\forall c$, $\mathbb{E}[k^T A^\dagger h^T] = 0$ *and* $\mathrm{Var}(k^T A^\dagger h^T) = O(1/(\rho^6 d))$

Proof For all three terms we will need the SVD $A = U\Sigma V^T$, with $A^\dagger = V\Sigma^\dagger U^T$.

For the first term, we note that

$$\begin{aligned}\beta_*^T uh\beta_*^T &= (\beta_*^T u)\cdot v^T A^\dagger \beta_* \\ &= (\beta_*^T u)v^T V\Sigma^\dagger U^T\beta_* \\ &= (\beta_*^T u)\sum_{i=1}^{\min(n,d)} (v^T V)_i (U^T\beta_*)_i \frac{1}{\sigma_i(A)}\end{aligned}$$

Since A is isotropic Gaussian, we have that U, V are uniformly random orthogonal matrices. Thus, we have that $v^T V$ and $U^T\beta_*$ are uniformly random vectors on a spheres of radius $\|v\|$ and $\|\beta_*\|$ respectively. In particular, they are independent and have mean zero. Thus, we see that

$$\mathbb{E}\left[\beta_*^T uh\beta_*^T\right] = 0$$

Since $v^T V$ and $U^T\beta_*$ are independent. We have their we have that variance of this term is $O(1/(\rho^2 d))$

For the second term, we note that

$$\beta_*^T AA^\dagger u = \sum_{i=1}^{\min(n,d)} (\beta_*^T U)_i (U^T u)_i \text{ and } h\beta_* = \sum_{i=1}^{\min(n,d)} (v^T V)_i (U^T\beta_*)_i \frac{1}{\sigma_i(A)}$$

Multiplying the two together, we get that

$$\beta_*^T AA^\dagger uh\beta_* = \sum_{i,j}^{\min(n,d)} (\beta_*^T U)_i (U^T u)_i (v^T V)_j (U^T\beta_*)_j \frac{1}{\sigma_i(A)}$$

Noting that v^TV is a uniformly random mean zero vector independent of everything else in the summation. Hence the expectation is equal to zero.

$$\mathbb{E}\left[\beta_*^T AA^\dagger uh\beta_*\right] = 0$$

Using Theorem 37, the variance of this term is $O(1/(\rho^2 d^2))$

For the third term, we have that

$$\begin{aligned}\beta_*^T sh\beta_* &= \beta_*^T(I - AA^\dagger)uh\beta_* \\ &= \beta_*^T uh\beta_* - \beta_*^T AA^\dagger uh\beta_*\end{aligned}$$

Then using the previous two parts, we get that each term has mean zero. Thus, we get the needed result. Using Theorem 35, the variance of this term is $O(1/(\rho^2 d))$

Finally, we have that for the last term:

$$k^T A^\dagger h^T = uU\Sigma^{\dagger T}\Sigma^\dagger\Sigma^{\dagger T}V^T v$$

Hence using similar arguments to before, using the independence of U, Σ, V, we get mean zero and variance $O(1/(\rho^6 n))$ ■

Lemma 20 (Moments of γ_1/η^2) *We have:*

(i) For γ_1/η^2,

$$\mathbb{E}\left[\frac{\gamma_1}{\eta^2}\right] = \frac{c}{\rho^2} + \frac{1}{\eta^2} + o\left(\frac{1}{\rho^2}\right), \quad \mathrm{Var}\left(\frac{\gamma_1}{\eta^2}\right) = O\left(\frac{1}{n}\right).$$

(ii) For γ_2/η^2,

$$\mathbb{E}\left[\frac{\gamma_2}{\eta^2}\right] = \frac{1}{\rho^2} + \frac{1}{\eta^2} + o\left(\frac{1}{\rho^2}\right), \quad \mathrm{Var}\left(\frac{\gamma_2}{\eta^2}\right) = O\left(\frac{1}{n}\right).$$

Proof We decompose

$$\frac{\gamma_i}{\eta^2} = \zeta_i + Y, \quad i = 1, 2,$$

where

$$\zeta_1 = \|t\|^2\,\|k\|^2, \quad \zeta_2 = \|s\|^2\,\|h\|^2, \quad \text{and} \quad Y = \frac{\xi^2}{\eta^2}.$$

Expectation Estimates:

Even though the norm terms (e.g. $\|t\|^2$ and $\|k\|^2$) are not independent, we use the standard bound

$$\left|\mathbb{E}\left[\|t\|^2\,\|k\|^2\right] - \mathbb{E}[\|t\|^2]\,\mathbb{E}[\|k\|^2]\right| \le \sqrt{\mathrm{Var}\left(\|t\|^2\right)\mathrm{Var}\left(\|k\|^2\right)}.$$

By Theorem 9 we have

$$\mathrm{Var}(\|t\|^2) = O\left(\frac{1}{n}\right), \quad \mathrm{Var}(\|k\|^2) = O\left(\frac{1}{\rho^4 n}\right),$$

so that

$$\sqrt{\operatorname{Var}\left(\|t\|^2\right)\operatorname{Var}\left(\|k\|^2\right)} = O\left(\frac{1}{\rho^2 n}\right).$$

Therefore,

$$\mathbb{E}[\zeta_1] = \mathbb{E}[\|t\|^2\, \|k\|^2] = \mathbb{E}[\|t\|^2]\, \mathbb{E}[\|k\|^2] + O\left(\frac{1}{\rho^2 n}\right).$$

Using Theorem 9 again,

$$\mathbb{E}[\|t\|^2] = 1 - c, \quad \mathbb{E}[\|k\|^2] = \frac{1}{\rho^2}\frac{c}{1-c} + o\left(\frac{1}{\rho^2}\right),$$

which implies

$$\mathbb{E}[\zeta_1] = (1-c)\left(\frac{1}{\rho^2}\frac{c}{1-c}\right) + o\left(\frac{1}{\rho^2}\right) + o\left(\frac{1}{\rho^2}\right) = \frac{c}{\rho^2} + o\left(\frac{1}{\rho^2}\right).$$

Finally from Theorem 9 item 9,

$$\mathbb{E}[Y] = \frac{1}{\eta^2} + O\left(\frac{1}{n\rho^2}\right),$$

since for $c < 1$ we have $\max(n, d) = n$. Hence,

$$\mathbb{E}\left[\frac{\gamma_1}{\eta^2}\right] = \mathbb{E}[\zeta_1] + \mathbb{E}[Y] = \frac{c}{\rho^2} + \frac{1}{\eta^2} + o\left(\frac{1}{\rho^2}\right).$$

A similar argument applies for γ_2/η^2. Using items 9 and 9,

$$\mathbb{E}\left[\frac{\gamma_2}{\eta^2}\right] = \frac{1}{\rho^2} + \frac{1}{\eta^2} + o\left(\frac{1}{\rho^2}\right).$$

Variance Estimates:

By Theorem 35, we have that

$$\operatorname{Var}(\zeta_1) \leq O\left(\operatorname{Var}(\|k\|^2) + \operatorname{Var}(\|t\|^2)\right)$$

Thus, we see that

$$\operatorname{Var}(\zeta_1) = O\left(\frac{1}{n} + \frac{1}{\rho^4 n}\right) = O\left(\frac{1}{n}\right)$$

For both cases, by Theorem 36 the variance obeys

$$\operatorname{Var}\left(\frac{\gamma_i}{\eta^2}\right) \leq \left(\sqrt{\operatorname{Var}(\zeta_i)} + \sqrt{\operatorname{Var}(Y)}\right)^2$$

Also, from item 9

$$\operatorname{Var}(Y) = O\left(\frac{1}{\max(n,d)^2\rho^4}\right) = o\left(\frac{1}{\rho^4 n}\right).$$

Thus, we get that

$$\operatorname{Var}\left(\frac{\gamma_i}{\eta^2}\right) = O\left(\frac{1}{n}\right), \quad i = 1, 2.$$

■

Lemma 21 (Moments of ξ/γ_i) *Defining*

$$\frac{\eta\xi}{\gamma_i} = \frac{\xi/\eta}{\gamma_i/\eta^2}, \quad i = 1, 2,$$

we have the following estimates for the moments:

(i) For $\eta\xi/\gamma_1$,

$$\mathbb{E}\left[\frac{\eta\xi}{\gamma_1}\right] = \frac{\rho^2\eta}{\eta^2 c + \rho^2} + o\left(\frac{1}{\eta\rho^2}\right) + O\left(\frac{1}{n\rho}\right), \quad \mathrm{Var}\left(\frac{\eta\xi}{\gamma_1}\right) = O\left(\frac{1}{n}\right).$$

(ii) For $\eta\xi/\gamma_2$,

$$\mathbb{E}\left[\frac{\eta\xi}{\gamma_2}\right] = \frac{\rho^2\eta}{\eta^2 + \rho^2} + o\left(\frac{1}{\eta\rho^2}\right) + O\left(\frac{1}{n\rho}\right), \quad \mathrm{Var}\left(\frac{\eta\xi}{\gamma_2}\right) = O\left(\frac{1}{n}\right).$$

Proof Write

$$\frac{\eta\xi}{\gamma_i} = \frac{\xi/\eta}{\gamma_i/\eta^2}, \quad i = 1, 2.$$

Since both ξ/η and γ_i/η^2 concentrate (with vanishing variances) and are bounded away from zero with high probability, standard concentration bounds and the delta method (in particular, Lemma 33) imply that

$$\begin{aligned}
\mathbb{E}\left[\frac{\eta\xi}{\gamma_1}\right] &= \frac{\mathbb{E}[\xi/\eta]}{\mathbb{E}[\gamma_1/\eta^2]} + \sqrt{\mathrm{Var}([\gamma_i/\eta^2)\mathrm{Var}(\xi/\eta)} \\
&= \frac{1}{\eta} \cdot \left(\frac{\rho^2\eta^2}{\eta^2 c + \rho^2} + o\left(\frac{1}{\rho^2}\right)\right) + O\left(\frac{1}{n\rho}\right) \\
&= \frac{\rho^2\eta}{\eta^2 c + \rho^2} + o\left(\frac{1}{\eta\rho^2}\right) + O\left(\frac{1}{n\rho}\right)
\end{aligned}$$

For the variance, we have from Theorem 36

$$\begin{aligned}
\mathrm{Var}\left(\frac{\eta\xi}{\gamma_1}\right) &\leq \left(\frac{1}{\eta^2} O\left(\frac{1}{n}\right) + \frac{\rho^4\eta^4}{(\eta^2 c + \rho^2)^2} O\left(\frac{1}{n\rho^2}\right) + O\left(\frac{1}{n} + \frac{1}{n\rho^2}\right)\right) \\
&\leq O\left(\frac{1}{n}\right)
\end{aligned}$$

Similarly for the other term. ■

Lemma 22 (Moments of ξ/γ_i) *Defining*

$$\frac{\xi^2}{\gamma_i} = \frac{\xi^2/\eta^2}{\gamma_i/\eta^2}, \quad i = 1, 2,$$

we have the following estimates for the moments:

(i) For ξ^2/γ_1,

$$\mathbb{E}\left[\frac{\xi^2}{\gamma_1}\right] = \frac{\rho^2\eta^2}{\eta^2 c+\rho^2}\left(\frac{1}{\eta^2}+\frac{1}{n\rho^2}\frac{c}{1-c}\right)+o\left(\frac{1}{\eta^2\rho^2}\right)+O\left(\frac{1}{n}\right), \quad \mathrm{Var}\left(\frac{\xi^2}{\gamma_1}\right) = O\left(\frac{1}{n^2}\right)+O\left(\frac{1}{\eta^4 n}\right).$$

(ii) For ξ^2/γ_2,

$$\mathbb{E}\left[\frac{\xi^2}{\gamma_2}\right] = \frac{\rho^2\eta^2}{\eta^2+\rho^2}\left(\frac{1}{\eta^2}+\frac{1}{n\rho^2}\frac{1}{c-1}\right)+o\left(\frac{1}{\eta^2\rho^2}\right)+O\left(\frac{1}{n}\right), \quad \mathrm{Var}\left(\frac{\xi^2}{\gamma_2}\right) = O\left(\frac{1}{n^2}\right)+O\left(\frac{1}{\eta^4 n}\right).$$

Proof Using Theorem 9 and Theorem 21, we see that

$$\begin{aligned}\mathbb{E}\left[\frac{\xi^2/\eta^2}{\gamma_1/\eta^2}\right] &= \left(\frac{\rho^2\eta^2}{\eta^2 c+\rho^2}+o\left(\frac{1}{\rho^2}\right)\right)\cdot\left(\frac{1}{\eta^2}+o\left(\frac{1}{n\rho^2}\right)+O\left(\frac{1}{n\rho^2}\right)\right)+O\left(\frac{1}{\rho^2 n^{1.5}}\right)\\ &= \frac{\rho^2}{\eta^2 c+\rho^2}+o\left(\frac{1}{\eta^2\rho^2}\right)+O\left(\frac{1}{n}\right).\end{aligned}$$

$$\begin{aligned}\mathbb{E}\left[\frac{\xi^2/\eta^2}{\gamma_2/\eta^2}\right] &= \left(\frac{\rho^2\eta^2}{\eta^2+\rho^2}+o\left(\frac{1}{\rho^2}\right)\right)\cdot\left(\frac{1}{\eta^2}+o\left(\frac{1}{d\rho^2}\right)+O\left(\frac{1}{d\rho^2}\right)\right)+O\left(\frac{1}{\rho^2 n^{1.5}}\right)\\ &= \frac{\rho^2}{\eta^2+\rho^2}+o\left(\frac{1}{\eta^2\rho^2}\right)+O\left(\frac{1}{n}\right)\end{aligned}$$

Similarly, using Theorem 36, we have that

$$\mathrm{Var}\left(\frac{\xi^2/\eta^2}{\gamma_1/\eta^2}\right), \mathrm{Var}\left(\frac{\xi^2/\eta^2}{\gamma_2/\eta^2}\right) = O\left(\frac{1}{n^2}\right)+O\left(\frac{1}{\eta^4 n}\right)$$

■

Lemma 23 (Moments of $\|t\|^2(\|h\|^2)/\gamma_i$) *We have the following estimates for the moments:*

(i) For $\eta^2\|t\|^2/\gamma_1$,

$$\mathbb{E}\left[\frac{\eta^2\|t\|^2}{\gamma_1}\right] = \frac{\rho^2\eta^2}{\eta^2 c+\rho^2}(1-c)+o(1)+O\left(\frac{1}{n}\right), \quad \mathrm{Var}\left(\frac{\eta^2\|t\|^2}{\gamma_1}\right) = O\left(\frac{1}{n}\right).$$

(ii) For $\eta^2\|h\|^2/\gamma_1$,

$$\mathbb{E}\left[\frac{\eta^2\|h\|^2}{\gamma_1}\right] = \frac{\eta^2}{\eta^2 c+\rho^2}\frac{c^2}{1-c}+o\left(\frac{1}{\rho^2}\right)+O\left(\frac{1}{\rho^2 n}\right), \quad \mathrm{Var}\left(\frac{\eta^2\|h\|^2}{\gamma_1}\right) = O\left(\frac{1}{n}\right).$$

(iii) For $\eta^2\|h\|^2/\gamma_2$,

$$\mathbb{E}\left[\frac{\eta^2\|h\|^2}{\gamma_2}\right] = \frac{\eta^2}{\eta^2+\rho^2}\cdot\frac{c}{c-1}+o\left(\frac{1}{\rho^2}\right)+O\left(\frac{1}{\rho^2 n}\right), \quad \mathrm{Var}\left(\frac{\eta^2\|h\|^2}{\gamma_2}\right) = O\left(\frac{1}{n}\right).$$

Proof Similar to Lemma 23, since both $\|k\|^2$ and γ_i/η^2 concentrate (with vanishing variances) and are bounded away from zero with high probability, standard concentration bounds and the delta method (in particular, Lemma 33) imply that

$$\begin{aligned}\mathbb{E}\left[\frac{\eta^2\|t\|^2}{\gamma_1}\right] &= \frac{\mathbb{E}[\|t\|^2]}{\mathbb{E}[\gamma_1/\eta^2]} + \sqrt{\mathrm{Var}(\gamma_i/\eta^2)\mathrm{Var}(\|t\|^2)} \\ &= (1-c)\cdot\left(\frac{\rho^2\eta^2}{\eta^2 c+\rho^2} + o\left(\frac{1}{\rho^2}\right)\right) + O\left(\frac{1}{n}\right) \\ &= (1-c)\cdot\frac{\rho^2\eta^2}{\eta^2 c+\rho^2} + o\left(\frac{1}{\rho^2}\right) + O\left(\frac{1}{n}\right)\end{aligned}$$

For the variance, we have from Theorem 36

$$\begin{aligned}\mathrm{Var}\left(\frac{\eta^2\|t\|^2}{\gamma_1}\right) &\leq \left((1-c)^2 O\left(\frac{1}{n}\right) + \frac{\rho^4\eta^4}{(\eta^2 c+\rho^2)^2} O\left(\frac{1}{n}\right) + O\left(\frac{1}{n}+\frac{1}{n}\right)\right) \\ &\leq O\left(\frac{1}{n}\right)\end{aligned}$$

Similarly, we have that:

$$\begin{aligned}\mathbb{E}\left[\frac{\eta^2\|h\|^2}{\gamma_1}\right] &= \frac{\mathbb{E}[\|h\|^2]}{\mathbb{E}[\gamma_1/\eta^2]} + \sqrt{\mathrm{Var}(\gamma_i/\eta^2)\mathrm{Var}(\|h\|^2)} \\ &= \left(\frac{1}{\rho^2}\frac{c^2}{1-c} + o\left(\frac{1}{\rho^2}\right)\right)\cdot\left(\frac{\rho^2\eta^2}{\eta^2 c+\rho^2} + o\left(\frac{1}{\rho^2}\right)\right) + O\left(\frac{1}{\rho^2 n}\right) \\ &= \frac{\eta^2}{\eta^2 c+\rho^2}\frac{c^2}{1-c} + o\left(\frac{1}{\rho^2}\right) + O\left(\frac{1}{\rho^2 n}\right)\end{aligned}$$

For the variance, we have from Theorem 36

$$\begin{aligned}\mathrm{Var}\left(\frac{\eta^2\|h\|^2}{\gamma_1}\right) &\leq \left(\left(\frac{1}{\rho^2}\frac{c^2}{1-c}\right)^2 O\left(\frac{1}{n}\right) + \frac{\rho^4\eta^4}{(\eta^2 c+\rho^2)^2} O\left(\frac{1}{\rho^4 n}\right) + O\left(\frac{1}{n}+\frac{1}{\rho^4 n}\right)\right) \\ &\leq O\left(\frac{1}{n}\right).\end{aligned}$$

This proof is similar for the last term. ■

Lemma 24 (Moments of $\|s\|^2(\|k\|^2)/\gamma_i$) *We have the following estimates for the moments:*

(i) For $\eta^2\|k\|^2/\gamma_1$,

$$\mathbb{E}\left[\frac{\eta^2\|k\|^2}{\gamma_1}\right] = \frac{c}{1-c}\frac{\eta^2}{\eta^2 c+\rho^2} + o\left(\frac{1}{\rho^2}\right) + O\left(\frac{1}{\rho^2 n}\right), \quad \mathrm{Var}\left(\frac{\eta^2\|k\|^2}{\gamma_1}\right) = O\left(\frac{1}{n}\right).$$

(ii) For $\eta^2\|s\|^2/\gamma_2$,

$$\mathbb{E}\left[\frac{\eta^2\|s\|^2}{\gamma_2}\right] = \left(1-\frac{1}{c}\right)\frac{\rho^2\eta^2}{\eta^2+\rho^2} + o\left(1\right) + O\left(\frac{1}{n}\right), \quad \mathrm{Var}\left(\frac{\eta^2\|s\|^2}{\gamma_2}\right) = O\left(\frac{1}{n}\right).$$

(iii) For $\eta^2\|k\|^2/\gamma_2$,

$$\mathbb{E}\left[\frac{\eta^2\|k\|^2}{\gamma_2}\right] = \frac{1}{c-1}\frac{\eta^2}{\eta^2+\rho^2} + o\left(\frac{1}{\rho^2}\right) + O\left(\frac{1}{\rho^2 n}\right), \quad \mathrm{Var}\left(\frac{\eta^2\|k\|^2}{\gamma_2}\right) = O\left(\frac{1}{n}\right).$$

Proof Similar to Lemma 23, since both $\|k\|^2$ and γ_i/η^2 concentrate, Lemma 33 imply that

$$\begin{aligned}\mathbb{E}\left[\frac{\eta^2\|k\|^2}{\gamma_1}\right] &= \frac{\mathbb{E}[\|k\|^2]}{\mathbb{E}[\gamma_1/\eta^2]} + \sqrt{\mathrm{Var}(\gamma_1/\eta^2)\mathrm{Var}(\|k\|^2)} \\ &= \left(\frac{1}{\rho^2}\frac{c}{1-c} + o\left(\frac{1}{\rho^2}\right)\right)\left(\frac{\rho^2\eta^2}{\eta^2 c+\rho^2} + o\left(\frac{1}{\rho^2}\right)\right) + O\left(\frac{1}{\rho^2 n}\right) \\ &= \frac{c}{1-c}\frac{\eta^2}{\eta^2 c+\rho^2} + o\left(1\right) + O\left(\frac{1}{\rho^2 n}\right).\end{aligned}$$

For the variance, we have from Theorem 36

$$\begin{aligned}\mathrm{Var}\left(\frac{\eta^2\|k\|^2}{\gamma_1}\right) &\le \left(\frac{1}{\rho^2}\frac{c}{1-c}\right)^2 O\left(\frac{1}{n}\right) + \left(\frac{c}{\rho^2}+\frac{1}{\eta^2}\right)^2 O\left(\frac{1}{n}\right) + O\left(\frac{1}{n}\right) \\ &\le O\left(\frac{1}{n}\right).\end{aligned}$$

Similarly for the other term. ■

Lemma 25 *Suppose $\varepsilon \in \mathbb{R}^n$ whose entries have mean 0, variance τ_ε, and follow our noise assumptions. Then for any random matrix $Q \in \mathbb{R}^{n\times n}$ independent, we have*

$$\mathbb{E}_{\varepsilon,Q}\left[\varepsilon^T Q\varepsilon\right] = \tau_\varepsilon^2\mathbb{E}\left[\mathrm{Tr}(Q)\right].$$

Proof We have that

$$\varepsilon^T Q\varepsilon = \sum_{i=1}^n\sum_{j=1}^n \varepsilon_i\varepsilon_j q_{ij}.$$

We take the expectation of this sum. By the independence assumption and assumption $\mathbb{E}[\varepsilon_i\varepsilon_j] = 0$ when $i \neq j$, we then have

$$\mathbb{E}_{\varepsilon,Q}\left[\varepsilon^T Q\varepsilon\right] = \sum_{i=1}^n \mathbb{E}\left[\varepsilon_i^2\right]\mathbb{E}\left[q_{ii}\right] = \tau_\varepsilon^2\mathbb{E}\left[\sum_{i=1}^n q_{ii}\right] = \tau_\varepsilon^2\mathbb{E}\left[\mathrm{Tr}(Q)\right].$$

■

Lemma 26 *In the same setting as Section 2, we have that,*

$$\mathbb{E}\left[\beta_*^T(Z+A)^{\dagger T}Z^T\beta_*\right] = \begin{cases}\frac{\eta^2 c}{\rho^2+\eta^2 c}(\beta_*^T u)^2 + o\left(\frac{1}{\rho^2}\right) + O\left(\frac{1}{\rho n}\right), & c<1 \\ \frac{\eta^2}{\eta^2+\rho^2}(\beta_*^T u)^2 + o\left(\frac{1}{\rho^2}\right) + O\left(\frac{1}{n}\right), & c>1\end{cases}.$$

Proof For $c < 1$, from Proposition 2, we get that

$$\beta_*^T (Z + A)^{\dagger T} Z^T \beta_* = \frac{\eta\xi}{\gamma_1} \beta_*^T h^T u^T \beta_* + \frac{\eta^2 \|t\|^2}{\gamma_1} \beta_*^T A^{\dagger T} k u^T \beta_*$$

Let us now compute the expected value of both terms. For the first one, we have that by approximating the expectation of the product with the product of the expectations it has mean 0 with an error of

$$\sqrt{\text{Var}\left(\frac{\eta\xi}{\gamma_1}\right) \text{Var}\left(\beta_*^T h^T u^T \beta_*\right)} = O\left(\frac{1}{n\rho}\right).$$

Thus, we have that

$$\mathbb{E}\left[\frac{\eta\xi}{\gamma_1} \beta_*^T h^T u^T \beta_*\right] = O\left(\frac{1}{n\rho}\right)$$

Here we used Theorem 21 and Theorem 19. Using Theorem 19 and Theorem 23, we see that the mean of the second term is

$$\begin{aligned}\mathbb{E}\left[\frac{\eta^2 \|t\|^2}{\gamma_1} \beta_*^T A^{\dagger T} k u^T \beta_*\right] &= \left(\frac{\rho^2 \eta^2}{\eta^2 c + \rho^2}(1 - c) + o\left(1\right) + O\left(\frac{1}{n}\right)\right) \cdot \left(\frac{c}{\rho^2 (1 - c)} (\beta_*^T u)^2 + o\left(\frac{1}{\rho^2}\right)\right) + O\left(\frac{1}{\rho n}\right) \\ &= \frac{\eta^2 c}{\rho^2 + \eta^2 c} (\beta_*^T u)^2 + o\left(\frac{1}{\rho^2}\right) + O\left(\frac{1}{\rho n}\right)\end{aligned}$$

Similarly, for $c > 1$, with expectations from Proposition 2, Theorem 18, Theorem 19, it follows that

$$\begin{aligned}\beta_*^T (Z + A)^{\dagger T} Z^T \beta_* &= \beta_*^T \left(\frac{\eta\xi}{\gamma_2} uh + \frac{\eta^2 \|h\|^2}{\gamma_2} us^T\right)^T \beta_* \\ &= \frac{\eta\xi}{\gamma_2} \beta_*^T h^T u^T \beta_* + \frac{\eta^2 \|h\|^2}{\gamma_2} \beta_*^T s u^T \beta_*\end{aligned}$$

Similar to before the mean zero term is of order $o(1/(\rho n))$. While the second term is

$$\begin{aligned}\mathbb{E}\left[\frac{\eta^2 \|h\|^2}{\gamma_2} \beta_*^T s u^T \beta_*\right] &= \left(\frac{\eta^2}{\eta^2 + \rho^2} \frac{c}{c - 1} + o\left(\frac{1}{\rho^2}\right) + O\left(\frac{1}{\rho^2 n}\right)\right) \cdot \left(\frac{c - 1}{c} (\beta_*^T u)^2\right) + O\left(\frac{1}{n}\right) \\ &= \frac{\eta^2}{\eta^2 + \rho^2} (\beta_*^T u)^2 + o\left(\frac{1}{\rho^2}\right) + O\left(\frac{1}{n}\right)\end{aligned}$$

■

Lemma 27 *In the same setting as Section 2, we have that, for $c < 1$*

$$\mathbb{E}\left[\beta_*^T (Z + A)^{\dagger T} A^T \beta_*\right] = \|\beta_*\|^2 - \frac{\eta^2 c}{\rho^2 + \eta^2 c} (\beta_*^T u)^2 + o\left(\frac{1}{\rho^2}\right) + O\left(\frac{1}{\rho n}\right),$$

and for $c > 1$

$$\mathbb{E}\left[\beta_*^T (Z + A)^{\dagger T} A^T \beta_*\right] = \frac{1}{c} \|\beta_*\|^2 - \frac{1}{d} \frac{\eta^2}{\eta^2 + \rho^2} \|\beta_*\|^2 - \frac{1}{c} \frac{\eta^2}{\eta^2 + \rho^2} (\beta_*^T u)^2 + o\left(\frac{1}{\rho^2} + \frac{1}{n}\right) + O\left(\frac{1}{\rho n}\right).$$

Proof For $c < 1$, using that Theorem 11, we get

$$\begin{aligned}\beta_*^T (Z+A)^{\dagger T} A^T \beta_* &= \beta_*^T \left(I - Z(Z+A)^\dagger\right)^T \beta_* \\ &\overset{E}{=} \|\beta_*\|^2 - \frac{\eta^2 c}{\rho^2 + \eta^2 c}(\beta_*^T u)^2 + o\left(\frac{1}{\rho^2}\right) + O\left(\frac{1}{\rho n}\right)\end{aligned}$$

For $c > 1$, using Theorem 11, we get

$$\beta_*^T (Z+A)^{\dagger T} A^T \beta_* = \beta_*^T \left(AA^\dagger + \frac{\eta\xi}{\gamma_2} h^T s^T - \frac{\eta^2 \|s\|^2}{\gamma_2} h^T h - \frac{\eta^2 \|h\|^2}{\gamma_2} AA^\dagger u s^T - \frac{\eta\xi}{\gamma_2} AA^\dagger u h\right)^T \beta_*$$

Then we have that

$$\mathbb{E}\left[\beta_*^T AA^\dagger \beta_*^T\right] = \frac{1}{c}\|\beta_*\|^2$$

Next, using Theorem 19 and Theorem 20, we have that

$$\mathbb{E}\left[\beta_*^T \left(\frac{\eta\xi}{\gamma_2} h^T s^T\right)\beta_*\right] = O\left(\frac{1}{\rho n}\right) \text{ and } \mathbb{E}\left[\beta_*^T \left(\frac{\eta\xi}{\gamma_2} AA^\dagger u h\right)\beta_*\right] = O\left(\frac{1}{\rho n^{1.5}}\right)$$

Then using Theorem 18 and Theorem 24, we have that

$$\begin{aligned}\mathbb{E}\left[\beta_*^T \left(\frac{\eta^2 \|s\|^2}{\gamma_2} h^T h\right)\beta_*\right] &= \left(\frac{1}{d}\|\beta_*\|^2 \frac{c}{\rho^2(c-1)}\right)\cdot\left(\left(1 - \frac{1}{c}\right)\frac{\rho^2\eta^2}{\eta^2+\rho^2} + o(1) + O\left(\frac{1}{n}\right)\right) + O\left(\frac{1}{\rho^2 n^{1.5}}\right) \\ &= \frac{1}{d}\frac{\eta^2}{\eta^2+\rho^2}\|\beta_*\|^2 + o\left(\frac{1}{\rho^2 n}\right) + O\left(\frac{1}{\rho^2 n^{1.5}}\right)\end{aligned}$$

The final term is

$$\begin{aligned}\mathbb{E}\left[\beta_*^T \left(\frac{\eta^2 \|h\|^2}{\gamma_2} AA^\dagger u s^T\right)\beta_*\right] &= \left(\frac{c-1}{c^2}(\beta_*^T u)^2 + o\left(\frac{1}{n}\right)\right)\cdot\left(\frac{\eta^2}{\eta^2+\rho^2}\cdot\frac{c}{c-1} + o\left(\frac{1}{\rho^2}\right) + O\left(\frac{1}{\rho^2 n}\right)\right) + O\left(\frac{1}{\rho n}\right.\\ &= \frac{1}{c}\frac{\eta^2}{\eta^2+\rho^2}(\beta_*^T u)^2 + o\left(\frac{1}{\rho^2} + \frac{1}{n}\right) + O\left(\frac{1}{\rho n}\right)\end{aligned}$$

Putting it all together, we get that

$$\mathbb{E}\left[\beta_*^T (Z+A)^{\dagger T} A^T \beta_*\right] = \frac{1}{c}\|\beta_*\|^2 - \frac{1}{d}\frac{\eta^2}{\eta^2+\rho^2}\|\beta_*\|^2 - \frac{1}{c}\frac{\eta^2}{\eta^2+\rho^2}(\beta_*^T u)^2 + o\left(\frac{1}{\rho^2} + \frac{1}{n}\right) + O\left(\frac{1}{\rho n}\right)$$

■

Lemma 28 (Expectations involving p_1 and p_2) *In the setting of Section 2, we have that*

1. *For $c = d/n < 1$:*

$$\mathbb{E}\left[\frac{\xi^2}{\gamma_1^2}\|p_1\|^2\right] = \frac{c}{1-c}\frac{\eta^2}{\eta^2 c + \rho^2} + o\left(\frac{1}{\rho^2}\right) + O\left(\frac{1}{\rho^2 n}\right).$$

2. *For $c = d/n > 1$:*

$$\mathbb{E}\left[\frac{\xi^2}{\gamma_2^2}\|p_2\|^2\right] = \frac{\eta^2}{c-1}\frac{\eta^2 c + \rho^2}{(\eta^2+\rho^2)^2} + o\left(\frac{1}{\rho^2}+\frac{1}{n}\right) + O\left(\frac{1}{n}\right)$$

Proof Theorem 15 showed

$$\frac{\xi^2}{\gamma_1^2}\|p_1\|^2 = \frac{\eta^2\|k\|^2}{\gamma_1}.$$

Then Theorem 24 gives us the result.

Using Lemma 15 for p_2:

$$\frac{\xi^2}{\gamma_2^2}\|p_2\|^2 = \frac{1}{\gamma_2^2}\left[\eta^4\|s\|^4\|A^\dagger h^T\|^2 + 2\eta^3\xi\|s\|^2 k^T A^\dagger h^T + \eta^2\xi^2\|k\|^2\right]$$

For the first term we have that, we begin by noting that from Theorem 24

$$\begin{aligned}\mathbb{E}\left[\frac{\eta^4\|s\|^4}{\gamma_2^2}\right] &= \mathbb{E}\left[\frac{\eta^2\|s\|^2}{\gamma_2}\right]^2 + \mathrm{Var}\left(\frac{\eta^2\|s\|^2}{\gamma_2}\right)\\ &= \left(1-\frac{1}{c}\right)^2\frac{\rho^4\eta^4}{(\eta^2+\rho^2)^2} + o(1) + O\left(\frac{1}{n^2}\right) + o(\rho^2) + o\left(\frac{1}{n}\right) + O\left(\frac{\rho^2}{n}\right) + O\left(\frac{1}{n}\right)\\ &= \left(1-\frac{1}{c}\right)^2\frac{\rho^4\eta^4}{(\eta^2+\rho^2)^2} + o(\rho^2) + O\left(\frac{\rho^2}{n}\right) \qquad (7)\end{aligned}$$

Then using Theorem 36, we have that

$$\mathrm{Var}\left(\frac{\eta^4\|s\|^4}{\gamma_2^2}\right) = O\left(\mathbb{E}\left[\frac{\eta^2\|s\|^2}{\gamma_2}\right]^2\mathrm{Var}\left(\frac{\eta^2\|s\|^2}{\gamma_2}\right)\right) = O\left(\frac{\rho^4}{n}\right)$$

Thus, we have that

$$\begin{aligned}\mathbb{E}\left[\frac{\eta^4\|s\|^4}{\gamma_2^2}\|A^\dagger h^T\|^2\right] &= \left(\frac{1}{\rho^4}\frac{c^3}{(c-1)^3} + o\left(\frac{1}{\rho^4}\right)\right)\left(\left(1-\frac{1}{c}\right)^2\frac{\rho^4\eta^4}{(\eta^2+\rho^2)^2} + o(\rho^2) + O\left(\frac{\rho^2}{n}\right)\right) + O\left(\frac{1}{\rho^2 n}\right)\\ &= \frac{c}{c-1}\frac{\eta^4}{(\eta^2+\rho^2)^2} + o\left(\frac{1}{\rho^2}\right) + O\left(\frac{1}{\rho^2 n}\right)\end{aligned}$$

We write the second term as

$$2\cdot\frac{\eta\xi}{\gamma_2}\cdot\frac{\eta^2\|s\|^2}{\gamma_2}\cdot k^T A^\dagger h$$

where we see that each term concentrates. Since the last term has mean zero, we have that

$$\mathbb{E}\left[\cdot\frac{\eta\xi}{\gamma_2}\cdot\frac{\eta^2\|s\|^2}{\gamma_2}\cdot k^T A^\dagger h\right] = 0 + \sqrt{\mathrm{Var}\left(\frac{\eta\xi}{\gamma_2}\cdot\frac{\eta^2\|s\|^2}{\gamma_2}\right)\mathrm{Var}\left(k^T A^\dagger h\right)}$$

Hence we need Var $\left(\frac{\eta\xi}{\gamma_2}\cdot\frac{\eta^2\|s\|^2}{\gamma_2}\right)$, which we get from Theorem 36, Theorem 24, and Theorem 21 as follows:

$$\text{Var}\left(\frac{\eta\xi}{\gamma_2}\cdot\frac{\eta^2\|s\|^2}{\gamma_2}\right)=O\left(\frac{\rho^4}{\eta^2 n}\right)+O\left(\frac{\rho^4}{n}\right)=O\left(\frac{\rho^4}{n}\right)$$

Thus, we get that

$$\mathbb{E}\left[\cdot\frac{\eta\xi}{\gamma_2}\cdot\frac{\eta^2\|s\|^2}{\gamma_2}\cdot k^TA^\dagger h\right]=O\left(\frac{1}{\rho n}\right)$$

The final term can be written as

$$\frac{\xi^2}{\gamma_2}\cdot\frac{\eta^2\|k\|^2}{\gamma_2}=\frac{\xi^2/\eta^2}{\gamma_2/\eta^2}\cdot\frac{\eta^2\|k\|^2}{\gamma_2}$$

The final term can be written as

$$\frac{\xi^2}{\gamma_2}\cdot\frac{\eta^2\|k\|^2}{\gamma_2}=\frac{\xi^2/\eta^2}{\gamma_2/\eta^2}\cdot\frac{\eta^2\|k\|^2}{\gamma_2},\quad \mathbb{E}\left[\frac{\xi^2}{\gamma_2}\right]=\frac{\rho^2\eta^2}{\eta^2+\rho^2}\left(\frac{1}{\eta^2}+\frac{1}{n\rho^2}\frac{1}{c-1}\right)$$

Then Theorem 24 gives us

$$\mathbb{E}\left[\frac{\eta^2\|k\|^2}{\gamma_2}\right]=\frac{1}{c-1}\frac{\eta^2}{\eta^2+\rho^2}+o\left(\frac{1}{\rho^2}\right)+O\left(\frac{1}{\rho^2 n}\right),\quad \text{Var}\left(\frac{\eta^2\|k\|^2}{\gamma_2}\right)=O\left(\frac{1}{n}\right).$$

Thus, we see that the mean is

$$\frac{1}{c-1}\frac{\eta^2\rho^2}{(\eta^2+\rho^2)^2}+o\left(\frac{1}{n}\right)+o\left(\frac{1}{\eta^2\rho^2}\right)+O\left(\frac{1}{n}\right)$$

Finally, putting all three terms together we get

$$\begin{aligned}\mathbb{E}\left[\frac{\xi^2}{\gamma_2^2}\|p_2\|^2\right]&=\frac{c}{c-1}\frac{\eta^4}{(\eta^2+\rho^2)^2}+o\left(\frac{1}{\rho^2}\right)+\frac{1}{c-1}\frac{\eta^2\rho^2}{(\eta^2+\rho^2)^2}+o\left(\frac{1}{n}\right)+O\left(\frac{1}{n}\right)\\&=\frac{\eta^2}{c-1}\frac{\eta^2c+\rho^2}{(\eta^2+\rho^2)^2}+o\left(\frac{1}{\rho^2}+\frac{1}{n}\right)+O\left(\frac{1}{n}\right)\end{aligned}$$

■

The following lemmas deal with terms in the variances.

Lemma 29 *In the same setting as Section 2, we have that,*

$$\mathbb{E}\left[\beta_*^TZ(Z+A)^\dagger(Z+A)^{\dagger T}Z\beta_*\right]=\begin{cases}\frac{\eta^2(\eta^2+\rho^2)}{(\eta^2c+\rho^2)^2}\frac{c^2}{1-c}(\beta_*^Tu)^2+o\left(1\right)+O\left(\frac{1}{n}\right), & c<1\\ \frac{\eta^2}{\eta^2+\rho^2}\frac{c}{c-1}(\beta_*^Tu)^2+o\left(\frac{1}{\rho^2}\right)+O\left(\frac{1}{\rho^2 n}\right), & c>1\end{cases}$$

Proof We start with $c<1$ and expand this term using Proposition 2:

$$\beta_*^TZ(Z+A)^\dagger(Z+A)^{\dagger T}Z\beta_*=\frac{\eta^2\|h\|^2\xi^2}{\gamma_1^2}(\beta_*^Tu)^2+\frac{\eta^4\|t\|^4}{\gamma_1^2}(k^TA^\dagger A^{\dagger T}k)(\beta_*^Tu)^2+\frac{2\eta^3\|t\|^2\xi}{\gamma_1^2}k^TA^\dagger h^T(\beta_*^Tu)^2.$$

We then start plugging in the expectations of these terms and keep track of the “cumulative" variance of the sum. By Lemmas 9, 23, 22,

$$\begin{aligned}\mathbb{E}\left[\frac{\eta^2\|h\|^2\xi^2}{\gamma_1^2}(\beta_*^T u)^2\right] &= (\beta_*^T u)^2\mathbb{E}\left[\frac{\eta^2\|h\|^2}{\gamma_1}\right]\mathbb{E}\left[\frac{\xi^2}{\gamma_1}\right] + \sqrt{\mathrm{Var}\left(\frac{\eta^2\|h\|^2}{\gamma_1}\right)\mathrm{Var}\left(\frac{\xi^2}{\gamma_1}\right)} \\ &= (\beta_*^T u)^2\left[\frac{\eta^2}{\eta^2 c+\rho^2}\frac{c^2}{1-c} + o\left(\frac{1}{\rho^2}\right) + O\left(\frac{1}{\rho^2 n}\right)\right]\left[\frac{\rho^2}{\eta^2 c+\rho^2} + o\left(\frac{1}{\eta^2\rho^2}\right) + O\left(\frac{1}{n}\right)\right] \\ &\quad + O\left(\frac{1}{n^{1.5}}\right) + O\left(\frac{1}{\eta^2 n}\right) \\ &= (\beta_*^T u)^2\frac{\eta^2\rho^2}{(\eta^2 c+\rho^2)^2}\frac{c^2}{1-c} + o\left(\frac{1}{\eta^2}\right) + O\left(\frac{1}{n}\right).\end{aligned}$$

For the second term, we begin from Theorem 23,

$$\begin{aligned}\mathbb{E}\left[\frac{\eta^4\|t\|^4}{\gamma_1^2}\right] &= \mathbb{E}\left[\frac{\eta^2\|t\|^2}{\gamma_1}\right]^2 + \mathrm{Var}\left(\frac{\eta^2\|t\|^2}{\gamma_1}\right) \\ &= (1-c)^2\,\frac{\rho^4\eta^4}{(\eta^2+\rho^2)^2} + o(1) + O\left(\frac{1}{n^2}\right) + o(\rho^2) + o\left(\frac{1}{n}\right) + O\left(\frac{\rho^2}{n}\right) + O\left(\frac{1}{n}\right) \\ &= (1-c)^2\,\frac{\rho^4\eta^4}{(\eta^2+\rho^2)^2} + o(\rho^2) + O\left(\frac{\rho^2}{n}\right)\end{aligned}$$

Then by Theorem 36,

$$\mathrm{Var}\left(\frac{\eta^4\|t\|^4}{\gamma_1^2}\right) = O\left(\mathbb{E}\left[\frac{\eta^2\|t\|^2}{\gamma_1}\right]^2\mathrm{Var}\left(\frac{\eta^2\|t\|^2}{\gamma_1}\right)\right) = O\left(\frac{\rho^4}{n}\right)$$

Finally, we have that:

$$\begin{aligned}\mathbb{E}\left[\frac{\eta^4\|t\|^4}{\gamma_1^2}(k^T A^\dagger A^{\dagger T}k)(\beta_*^T u)^2\right] &= (\beta_*^T u)^2\mathbb{E}\left[\frac{\eta^4\|t\|^4}{\gamma_1^2}\right]\mathbb{E}\left[k^T A^\dagger A^{\dagger T}k\right] + \sqrt{\mathrm{Var}\left(\frac{\eta^4\|t\|^4}{\gamma_1^2}\right)\mathrm{Var}\left(k^T A^\dagger A^{\dagger T}k\right)} \\ &= (\beta_*^T u)^2\left(\frac{(1-c)^2\,\rho^4\eta^4}{(\eta^2+\rho^2)^2} + o(\rho^2) + O\left(\frac{\rho^2}{n}\right)\right)\left(\frac{c^2}{\rho^4(1-c)^3} + o\left(\frac{1}{\rho^4}\right)\right) \\ &\quad + o(1) + O\left(\frac{1}{n^{1.5}}\right) \\ &= \frac{\eta^4}{(\eta^2+\rho^2)^2}\frac{c^2}{1-c}(\beta_*^T u)^2 + o(1) + O\left(\frac{1}{n^{1.5}}\right).\end{aligned}$$

We now have one term left. Similarly, we will have:

$$\begin{aligned}\mathbb{E}\left[\frac{\eta^3\|t\|^2\xi}{\gamma_1^2}\right] &= \mathbb{E}\left[\frac{\eta^2\|t\|^2}{\gamma_1}\right]\mathbb{E}\left[\frac{\eta\xi}{\gamma_1}\right] + \sqrt{\mathrm{Var}\left(\frac{\eta^2\|t\|^2}{\gamma_1}\right)\mathrm{Var}\left(\frac{\eta\xi}{\gamma_1}\right)} \\ &= \frac{\rho^4\eta^3}{(\eta^2 c+\rho^2)^2}(1-c) + o\left(\frac{\rho^2}{\eta}\right) + O\left(\frac{\rho}{n}\right) + O\left(\frac{\rho^2}{\eta n}\right).\end{aligned}$$

Then by Theorem 36,

$$\operatorname{Var}\left(\frac{\eta^3\|t\|^2\xi}{\gamma_1^2}\right) = O\left(\frac{1}{n}\right) O\left(\rho^4 + \frac{\rho^4}{\eta^2}\right) + o\left(\frac{1}{n}\right) = O\left(\frac{\rho^4}{n}\right).$$

The entire term becomes:

$$\begin{aligned}\mathbb{E}\left[\frac{\eta^3\|t\|^2\xi}{\gamma_1^2}k^T A^\dagger h^T(\beta_*^T u)^2\right] &= (\beta_*^T u)^2\mathbb{E}\left[\frac{\eta^3\|t\|^2\xi}{\gamma_1^2}\right]\mathbb{E}\left[k^T A^\dagger h^T\right] + \sqrt{\operatorname{Var}\left(\frac{\eta^3\|t\|^2\xi}{\gamma_1^2}\right)\operatorname{Var}\left(k^T A^\dagger h^T\right)} \\ &= O\left(\frac{1}{n\rho}\right).\end{aligned}$$

Now we have the expectations and errors for the three terms. Combining them yields the Lemma statement. For $c > 1$, we recall that $hs = 0$,

$$\begin{aligned}\beta_*^T Z(Z+A)^\dagger(Z+A)^{\dagger T} Z\beta_* &= \frac{\eta^2\|h\|^2\xi^2}{\gamma_2^2}(\beta_*^T u)^2 + \frac{\eta^4\|h\|^4\|s\|^2}{\gamma_2^2}(\beta_*^T u)^2 + \frac{2\eta^3\|h\|^2\xi}{\gamma_2^2}\beta_*^T uhsu^T\beta_* \\ &= \left(\frac{\eta^2\|h\|^2(\xi^2+\eta^2\|h\|^2\|s\|^2)}{\gamma_2^2}\right)(\beta_*^T u)^2 \\ &= \left(\frac{\eta^2\|h\|^2\gamma_2}{\gamma_2^2}\right)(\beta_*^T u)^2 \\ &= \frac{\eta^2\|h\|^2}{\gamma_2}(\beta_*^T u)^2 \\ &\overset{E}{=} \frac{\eta^2}{\eta^2+\rho^2}\frac{c}{c-1}(\beta_*^T u)^2 + o\left(\frac{1}{\rho^2}\right) + O\left(\frac{1}{\rho^2 n}\right) \qquad \text{by Lemma 23.}\end{aligned}$$

■

Lemma 30 *In the same setting as Section 2, we have that,*

$$\mathbb{E}\left[\beta_*^T A(Z+A)^\dagger(Z+A)^{\dagger T}A\beta_*\right] = \begin{cases}\|\beta_*\|^2 + \frac{\eta^2(\eta^2+\rho^2)}{(\eta^2c+\rho^2)^2}\frac{c^2}{1-c}(\beta_*^T u)^2 - \frac{2\eta^2 c}{\eta^2c+\rho^2}(\beta_*^T u)^2 + o(1) + O\left(\frac{1}{n}\right), & c < 1 \\ \frac{\|\beta_*\|^2}{c} - \frac{\eta^2}{\eta^2+\rho^2}\left(\frac{\|\beta_*\|^2}{d} - \frac{(\beta_*^T u)^2}{c(c-1)}\right) + o(1) + O\left(\frac{1}{n}\right), & c > 1\end{cases}$$

Proof We use similar expansions that follow from Lemma 11.

$$\begin{aligned}\beta_*^T A(Z+A)^\dagger(Z+A)^{\dagger T}A\beta_* &= \|\beta_*\|^2 + \frac{\eta^2\|h\|^2\xi^2}{\gamma_1^2}(\beta_*^T u)^2 + \frac{\eta^4\|t\|^4}{\gamma_1^2}(k^T A^\dagger A^{\dagger T}k)(\beta_*^T u)^2 \\ &+ \frac{2\eta^3\|t\|^2\xi}{\gamma_1^2}(\beta_*^T u)^2 k^T A^\dagger h^T - \frac{2\eta^2\|t\|^2}{\gamma_1}\beta_*^T uk^T A^\dagger\beta_* - \frac{2\eta\xi}{\gamma_1}\beta_*^T uh\beta_*.\end{aligned}$$

Lemma 29 gives the expectation of the first four terms as,

$$\|\beta_*^2\|^2 + \frac{\eta^2(\eta^2+\rho^2)}{(\eta^2c+\rho^2)^2}\frac{c^2}{1-c}(\beta_*^T u)^2 + o\left(1\right) + O\left(\frac{1}{n}\right).$$

The other expectations follow from Lemmas 23, 18, 19,

$$\mathbb{E}\left[\frac{\eta\xi}{\gamma_1}\beta_*^T uh\beta_*\right] = \mathbb{E}\left[\frac{\eta\xi}{\gamma_1}\right]\mathbb{E}\left[\beta_*^T uh\beta_*\right] + \sqrt{\operatorname{Var}\left(\frac{\eta\xi}{\gamma_1}\right)\operatorname{Var}\left(\beta_*^T uh\beta_*\right)} = O\left(\frac{1}{n\rho}\right).$$

$$\begin{aligned}\mathbb{E}\left[\frac{\eta^2\|t\|^2}{\gamma_1}\beta_*^T uk^T A^\dagger\beta_*\right] &= \mathbb{E}\left[\frac{\eta^2\|t\|^2}{\gamma_1}\right]\mathbb{E}\left[\beta_*^T uk^T A^\dagger\beta_*\right] + \sqrt{\operatorname{Var}\left(\frac{\eta^2\|t\|^2}{\gamma_1}\right)\operatorname{Var}\left(\beta_*^T uk^T A^\dagger\beta_*\right)}\\ &= \left[\frac{\rho^2\eta^2}{\eta^2 c+\rho^2}(1-c) + o(1) + O\left(\frac{1}{n}\right)\right]\left[\frac{c}{\rho^2(1-c)}(\beta_*^T u)^2 + o\left(\frac{1}{\rho^2}\right)\right] + O\left(\frac{1}{n\rho}\right)\\ &= \frac{\eta^2 c}{\eta^2 c+\rho^2} + o(1) + O\left(\frac{1}{n\rho}\right).\end{aligned}$$

Combining these results yields the lemma statement.

For $c > 1$, we let $I_U = AA^\dagger$. With $hs = 0$, $s^T I_U = 0$, $hI_U = h$, we have the following expansion:

$$\begin{aligned}\beta_*^T A(Z+A)^\dagger(Z+A)^{\dagger T}A\beta_* = \beta_*^T I_U\beta_* &+ \frac{\eta^2\|s\|^2\xi^2}{\gamma_2^2}\beta_*^T h^T h\beta_* + \frac{\eta^4\|s\|^4\|h\|^2}{\gamma_2^2}\beta_*^T h^T h\beta_*\\ &+ \frac{\eta^4\|h\|^4\|s\|^2}{\gamma_2^2}\beta_*^T I_U uu^T I_U\beta_* + \frac{\eta^2\|h\|^2\xi^2}{\gamma_2^2}\beta_*^T I_U uu^T I_U\beta_*\\ &- \frac{2\eta^2\|s\|^2}{\gamma_2}\beta_*^T h^T h\beta_* - \frac{2\eta\xi}{\gamma_2}\beta_*^T I_U uh\beta_*\\ &- \frac{2\eta^3\|s\|^2\|h\|^2\xi}{\gamma_2^2}\beta_*^T I_U uh\beta_* + \frac{2\eta^3\|s\|^2\|h\|^2\xi}{\gamma_2^2}\beta_*^T I_U uh\beta_*\end{aligned}$$

We can combine the coefficients as:

$$\frac{\eta^2\|s\|^2\xi^2}{\gamma_2^2} + \frac{\eta^4\|s\|^4\|h\|^2}{\gamma_2^2} - \frac{2\eta^2\|s\|^2}{\gamma_2} = \frac{\eta^2\|s\|^2(\eta^2\|s\|^2\|h\|^2+\xi^2) - 2\eta^2\|s\|^2\gamma_2}{\gamma_2^2} = -\frac{\eta^2\|s\|^2}{\gamma_2},$$

$$\frac{\eta^4\|h\|^4\|s\|^2}{\gamma_2^2} + \frac{\eta^2\|h\|^2\xi^2}{\gamma_2^2} = \frac{\eta^2\|h\|^2(\eta^2\|s\|^2\|h\|^2+\xi^2)}{\gamma_2^2} = \frac{\eta^2\|h\|^2\gamma_2}{\gamma_2^2} = \frac{\eta^2\|h\|^2}{\gamma_2}.$$

Then we have that:

$$\beta_*^T A(Z+A)^\dagger(Z+A)^{\dagger T}A\beta_* = \beta_*^T I_U\beta_* - \frac{\eta^2\|s\|^2}{\gamma_2}\beta_*^T h^T h\beta_* + \frac{\eta^2\|h\|^2}{\gamma_2}\beta_*^T I_U uu^T I_U\beta_* - \frac{2\eta\xi}{\gamma_2}\beta_*^T I_U uh\beta_*.$$

Following Proposition 3, $\mathbb{E}[\beta_*^T I_U\beta_*] = \|\beta_*\|^2/c + o(1)$. Simiarly, we use Lemmas 18, 23, 24, 21 to obtain:

$$\begin{aligned}\mathbb{E}\left[\frac{\eta^2\|s\|^2}{\gamma_2}\beta_*^T h^T h\beta_*\right] &= \mathbb{E}\left[\frac{\eta^2\|s\|^2}{\gamma_2}\right]\mathbb{E}\left[\beta_*^T h^T h\beta_*\right] + \sqrt{\operatorname{Var}\left(\frac{\eta^2\|s\|^2}{\gamma_2}\right)\operatorname{Var}\left(\beta_*^T h^T h\beta_*\right)}\\ &= \left[\left(1-\frac{1}{c}\right)\frac{\rho^2\eta^2}{\eta^2+\rho^2} + o(1) + O\left(\frac{1}{n}\right)\right]\left[\frac{\|\beta_*\|^2}{d}\frac{c}{\rho^2(c-1)} + o\left(\frac{\|\beta_*\|^2}{d\rho^2}\right)\right] + O\left(\frac{1}{\rho^2 n^{1.5}}\right)\\ &= \frac{\|\beta_*\|^2}{d}\frac{\eta^2}{\eta^2+\rho^2} + o\left(\frac{\|\beta_*\|^2}{d}\right) + O\left(\frac{1}{\rho^2 n^{1.5}}\right).\end{aligned}$$

$$\mathbb{E}\left[\frac{\eta^2\|h\|^2}{\gamma_2}\beta_*^T I_U u u^T I_U \beta_*\right] = \frac{\eta^2}{\eta^2+\rho^2}\frac{(\beta_*^T u)^2}{c(c-1)} + o(1) + O\left(\frac{1}{n}\right), \quad \mathbb{E}\left[\frac{\eta\xi}{\gamma_2}\beta_*^T I_U u h \beta_*\right] = O\left(\frac{1}{\rho n^{1.5}}\right).$$

We combine these results. ■

Lemma 31 *In the same setting as Section 2, we have that,*

$$\mathbb{E}\left[\beta_*^T Z(Z+A)^\dagger (Z+A)^{\dagger T} A\beta_*\right] = \begin{cases} -\left(\frac{\eta^2(\eta^2+\rho^2)}{(\eta^2 c+\rho^2)^2}\frac{c^2}{1-c} - \frac{\eta^2 c}{\eta^2 c+\rho^2}\right)(\beta_*^T u)^2 + o(1) + O\left(\frac{1}{n}\right), & c<1 \\ -\frac{\eta^2}{\eta^2+\rho^2}\frac{1}{c-1}(\beta_*^T u)^2 + o(1) + O\left(\frac{1}{n}\right), & c>1 \end{cases}$$

Proof For $c < 1$, we expand using Proposition 2, Theorem 11, and note that all of the relevant terms have been evaluated in the proofs of Lemmas 29, 30,

$$\begin{aligned} \beta_*^T Z(Z+A)^\dagger (Z+A)^{\dagger T} A\beta_* &= \frac{\eta\xi}{\gamma_1}\beta_*^T u h \beta_* + \frac{\eta^2\|t\|^2}{\gamma_1}\beta_*^T u k^T A^\dagger \beta_* - \frac{2\eta^3\|t\|^2\xi}{\gamma_1^2}(\beta_*^T u)^2 h A^{\dagger T} k \\ &\quad - \frac{\eta^4\|t\|^4}{\gamma_1^2}(k^T A^\dagger A^{\dagger T} k)(\beta_*^T u)^2 - \frac{\eta^2\|h\|^2\xi^2}{\gamma_1^2}(\beta_*^T u)^2 \\ &\overset{E}{=} -\left(\frac{\eta^2(\eta^2+\rho^2)}{(\eta^2 c+\rho^2)^2}\frac{c^2}{1-c} - \frac{\eta^2 c}{\eta^2 c+\rho^2}\right)(\beta_*^T u)^2 + o(1) + O\left(\frac{1}{n}\right). \end{aligned}$$

For $c > 1$, $\beta_*^T Z(Z+A)^\dagger (Z+A)^{\dagger T} A\beta_*$ becomes:

$$\begin{aligned} &= \beta_*^T \frac{\eta\xi}{\gamma_2} u h \left(I_U + \frac{\eta\xi}{\gamma_2} s h - \frac{\eta^2\|s\|^2}{\gamma_2} h^T h - \frac{\eta^2\|h\|^2}{\gamma_2} s u^T I_U - \frac{\eta\xi}{\gamma_2} h^T u^T I_U\right)\beta_* \\ &+ \beta_*^T \frac{\eta^2\|h\|^2}{\gamma_2} u s^T \left(I_U + \frac{\eta\xi}{\gamma_2} s h - \frac{\eta^2\|s\|^2}{\gamma_2} h^T h - \frac{\eta^2\|h\|^2}{\gamma_2} s u^T I_U - \frac{\eta\xi}{\gamma_2} h^T u^T I_U\right)\beta_* \\ &= \beta_*^T \left[\frac{\eta\xi}{\gamma_2} u h I_U + \frac{\eta^2\xi^2}{\gamma_2^2} u h s h - \frac{\eta^3\xi\|s\|^2}{\gamma_2^2} u h h^T h - \frac{\eta^3\|h\|^2\xi}{\gamma_2^2} u h s u^T I_U - \frac{\eta^2\xi^2}{\gamma_2^2} u h h^T u^T I_U\right]\beta_* \\ &+ \beta_*^T \left[\frac{\eta^2\|h\|^2}{\gamma_2} u s^T I_U + \frac{\eta^3\|h\|^2\|s\|^2\xi}{\gamma_2^2} u h - \frac{\eta^4\|h\|^2\|s\|^2}{\gamma_2^2} u s^T h^T h - \frac{\eta^4\|h\|^4\|s\|^2}{\gamma_2^2} u u^T I_U - \frac{\eta^3\|h\|^2\xi}{\gamma_2^2} u s^T h^T u^T I_U\right]\beta_* \\ &= \beta_*^T \left[\frac{\eta\xi}{\gamma_2} u h I_U - \frac{\eta^3\xi\|s\|^2\|h\|^2}{\gamma_2^2} u h - \frac{\eta^2\|h\|^2\xi^2}{\gamma_2^2} u u^T I_U\right]\beta_* \\ &+ \beta_*^T \left[\frac{\eta^3\|h\|^2\|s\|^2\xi}{\gamma_2^2} u h - \frac{\eta^4\|h\|^4\|s\|^2}{\gamma_2^2} u u^T I_U\right]\beta_* \\ &= \beta_*^T \left[\frac{\eta\xi}{\gamma_2} u h I_U - \frac{\eta^2\|h\|^2\xi^2}{\gamma_2^2} u u^T I_U - \frac{\eta^4\|h\|^4\|s\|^2}{\gamma_2^2} u u^T I_U\right]\beta_* \\ &= \frac{\eta\xi}{\gamma_2}\beta_*^T u h I_U \beta_* - \frac{\eta^2\|h\|^2}{\gamma_2}\beta_*^T u u^T I_U \beta_* \\ &\overset{E}{=} -\frac{\eta^2}{\eta^2+\rho^2}\frac{1}{c-1}(\beta_*^T u)^2 + o(1) + O\left(\frac{1}{n}\right). \end{aligned}$$

■

Lemma 32 *In the same setting as Section 2, we have that,*

$$\mathbb{E}\left[\varepsilon^T (Z+A)^\dagger (Z+A)^{\dagger T}\varepsilon\right] = \begin{cases} \tau_\varepsilon^2 \left(\frac{cd}{\rho^2(1-c)} - \frac{\eta^2}{\rho^2(\eta^2 c+\rho^2)}\frac{c^2}{1-c}\right) + o\left(\frac{n}{\rho^2}\right) + O\left(\frac{1}{\rho^2 n}\right), & c<1 \\ \tau_\varepsilon^2 \left(\frac{d}{\rho^2(c-1)} - \frac{\eta^2}{\rho^2(\eta^2+\rho^2)}\frac{c}{c-1}\right) + o\left(\frac{n}{\rho^2}\right) + O\left(\frac{1}{\rho^2 n}\right), & c>1 \end{cases}$$

Proof For $c < 1$, we first expand this term using Theorem 8:

$$\begin{aligned}
&\varepsilon^T (Z+A)^\dagger (Z+A)^{\dagger T}\varepsilon \\
&= \varepsilon^T \left(A^\dagger + \frac{\eta}{\xi} t^T k^T A^\dagger - \frac{\xi}{\gamma_1} p_1 q_1^T\right)\left(A^\dagger + \frac{\eta}{\xi} t^T k^T A^\dagger - \frac{\xi}{\gamma_1} p_1 q_1^T\right)^T \varepsilon \\
&= \varepsilon^T A^\dagger A^{\dagger T}\varepsilon + \frac{2\eta}{\xi}\varepsilon^T A^\dagger A^{\dagger T} k t\varepsilon - \frac{2\xi}{\gamma_1}\varepsilon^T A^\dagger q_1 p_1^T \varepsilon + \frac{\eta^2}{\xi^2}\left(k^T A^\dagger A^{\dagger T} k\right)\varepsilon^T t^T t\varepsilon - \frac{2\eta}{\gamma_1}\varepsilon^T t^T k^T A^\dagger q_1 p_1^T\varepsilon + \frac{\xi^2}{\gamma_1^2}\varepsilon^T p_1 q_1^T q_1 p
\end{aligned}$$

Note that Lemma 25 and the fact that $tA^\dagger = 0$ imply that the second term has zero expectation:

$$\mathbb{E}_\varepsilon\left[\frac{2\eta}{\xi}\varepsilon^T A^\dagger A^{\dagger T} k t\varepsilon\right] = \frac{2\eta\tau_\varepsilon^2}{\xi} tA^\dagger A^{\dagger T} k = 0.$$

Simiarly, we will later use:

$$\mathbb{E}_\varepsilon\left[\varepsilon^T A^\dagger h^T t\varepsilon\right] = \tau_\varepsilon^2 tA^\dagger h^T = 0, \quad \mathbb{E}_\varepsilon\left[\varepsilon^T t^T k^T \varepsilon\right] = \tau_\varepsilon^2 Tr(t^T k^T) = \tau_\varepsilon^2 Tr(kt) = 0.$$

Note that these terms do not induce extra variance so that we can directly eliminate them from the following expressions. We now expand the other terms one by one and compute their expectations along the way:

$$\begin{aligned}
-\frac{2\xi}{\gamma_1}\varepsilon^T A^\dagger q_1 p_1^T\varepsilon &= -\frac{2\xi}{\gamma_1}\varepsilon^T A^\dagger \left(\frac{\eta\|t\|^2}{\xi} A^{\dagger T} k + h^T\right)\left(\frac{\eta^2\|k\|^2}{\xi} t + \eta k^T\right)\varepsilon \\
&= -\frac{2\eta^3\|t\|^2\|k\|^2}{\gamma_1\xi}\varepsilon^T A^\dagger A^{\dagger T} k t\varepsilon - \frac{2\eta^2\|t\|^2}{\gamma_1}\varepsilon^T A^\dagger A^{\dagger T} k k^T\varepsilon - \frac{2\eta^2\|k\|^2}{\gamma_1}\varepsilon^T A^\dagger h^T t\varepsilon - \frac{2\eta\xi}{\gamma_1}\varepsilon^T A^\dagger h^T k^T\varepsilon \\
&\overset{E}{=} -\frac{2\eta^2\|t\|^2\tau_\varepsilon^2}{\gamma_1} k^T A^\dagger A^{\dagger T} k - \frac{2\eta\xi\tau_\varepsilon^2}{\gamma_1} k^T A^\dagger h^T
\end{aligned}$$

$$\begin{aligned}
-\frac{2\eta}{\gamma_1}\varepsilon^T t^T k^T A^\dagger q_1 p_1^T\varepsilon &= -\frac{2\eta}{\gamma_1}\varepsilon^T t^T k^T A^\dagger \left(\frac{\eta\|t\|^2}{\xi} A^{\dagger T} k + h^T\right)\left(\frac{\eta^2\|k\|^2}{\xi} t + \eta k^T\right)\varepsilon \\
&= -\frac{2\eta^4\|t\|^2\|k\|^2}{\gamma_1\xi^2}\left(k^T A^\dagger A^{\dagger T} k\right)\varepsilon^T t^T t\varepsilon - \frac{2\eta^3\|t\|^2}{\gamma_1\xi}\left(k^T A^\dagger A^{\dagger T} k\right)\varepsilon^T t^T k^T\varepsilon \\
&\quad - \frac{2\eta^3\|k\|^2}{\gamma_1\xi}(k^T A^\dagger h^T)\varepsilon^T t^T t\varepsilon - \frac{2\eta^2}{\gamma_1}(k^T A^\dagger h^T)\varepsilon^T t^T k^T\varepsilon \\
&\overset{E}{=} -\frac{2\eta^4\|t\|^4\|k\|^2\tau_\varepsilon^2}{\gamma_1\xi^2} k^T A^\dagger A^{\dagger T} k - \frac{2\eta^3\|k\|^2\|t\|^2\tau_\varepsilon^2}{\gamma_1\xi} k^T A^\dagger h^T.
\end{aligned}$$

$$\frac{\eta^2}{\xi^2}\left(k^T A^\dagger A^{\dagger T} k\right)\varepsilon^T t^T t\varepsilon \overset{E}{=} \frac{\eta^2\|t\|^2\tau_\varepsilon^2}{\xi^2} k^T A^\dagger A^{\dagger T} k.$$

By the squared norms in Lemmas 15, 16, and Lemma 25,

$$\begin{aligned}
\frac{\xi^2}{\gamma_1^2}\varepsilon^T p_1 q_1^T q_1 p_1^T \varepsilon &\overset{E}{=} \frac{\xi^2\tau_\varepsilon^2}{\gamma_1^2}\|p_1\|^2\|q_1\|^2 \\
&= \frac{\xi^2\tau_\varepsilon^2}{\gamma_1^2}\left(\frac{\eta^2\|k\|^2}{\xi^2}\gamma_1\right)\left(\frac{\eta^2\|t\|^4}{\xi^2}kA^\dagger A^{\dagger T}k + \frac{2\eta\|t\|^2}{\xi}k^TA^\dagger h^T + \|h\|^2\right) \\
&= \frac{\tau_\varepsilon^2}{\gamma_1}\left(\eta^2\|k\|^2\right)\left(\frac{\eta^2\|t\|^4}{\xi^2}kA^\dagger A^{\dagger T}k + \frac{2\eta\|t\|^2}{\xi}k^TA^\dagger h^T + \|h\|^2\right) \\
&= \tau_\varepsilon^2\left(\frac{\eta^4\|t\|^4\|k\|^2}{\gamma_1\xi^2}kA^\dagger A^{\dagger T}k + \frac{2\eta^3\|t\|^2\|k\|^2}{\gamma_1\xi}k^TA^\dagger h^T + \frac{\eta^2\|k\|^2\|h\|^2}{\gamma_1}\right)
\end{aligned}$$

We combine like terms and simplify the coefficients, which can seem quite complicated at first:

For term $k^TA^\dagger A^{\dagger T}k$,

$$\begin{aligned}
\tau_\varepsilon^2\left(\frac{\eta^4\|t\|^4\|k\|^2}{\gamma_1\xi^2} - \frac{2\eta^4\|t\|^4\|k\|^2}{\gamma_1\xi^2} - \frac{2\eta^2\|t\|^2}{\gamma_1} + \frac{\eta^2\|t\|^2}{\xi^2}\right) &= \tau_\varepsilon^2\eta^2\|t\|^2\left(\frac{\eta^2\|t\|^2\|k\|^2}{\gamma_1\xi^2} - \frac{2\eta^2\|t\|^2\|k\|^2}{\gamma_1\xi^2} - \frac{2}{\gamma_1} + \frac{1}{\xi^2}\right) \\
&= \tau_\varepsilon^2\eta^2\|t\|^2\left(-\frac{\gamma_1-\xi^2}{\gamma_1\xi^2} - \frac{2}{\gamma_1} + \frac{1}{\xi^2}\right) \\
&= \tau_\varepsilon^2\eta^2\|t\|^2\left(-\frac{\gamma_1-\xi^2}{\gamma_1\xi^2} - \frac{2\xi^2}{\gamma_1\xi^2} + \frac{\gamma_1}{\gamma_1\xi^2}\right) \\
&= -\tau_\varepsilon^2\frac{\eta^2\|t\|^2}{\gamma_1}.
\end{aligned}$$

For term $k^TA^\dagger h^T$,

$$\tau_\varepsilon^2\left(\frac{2\eta^3\|t\|^2\|k\|^2}{\gamma_1\xi} - \frac{2\eta^3\|k\|^2\|t\|^2}{\gamma_1\xi} - \frac{2\eta\xi}{\gamma_1}\right) = -\tau_\varepsilon^2\frac{2\eta\xi}{\gamma_1}.$$

Combining these terms together, we have:

$$\varepsilon^T(Z+A)^\dagger(Z+A)^{\dagger T}\varepsilon \overset{E}{=} \varepsilon^TA^\dagger A^{\dagger T}\varepsilon - \frac{\eta^2\|t\|^2\tau_\varepsilon^2}{\gamma_1}k^TA^\dagger A^{\dagger T}k - \frac{2\eta\xi\tau_\varepsilon^2}{\gamma_1}k^TA^\dagger h^T + \frac{\eta^2\|k\|^2\|h\|^2}{\gamma_1}.$$

Similarly, using the relevant lemmas, we have the following:

$$\mathbb{E}\left[\varepsilon^TA^\dagger A^{\dagger T}\varepsilon\right] = \tau_\varepsilon^2\frac{cd}{\rho^2(1-c)} + o\left(\frac{d}{\rho^2}\right),$$

$$\mathbb{E}\left[\frac{\eta^2\|t\|^2}{\gamma_1}k^TA^\dagger A^{\dagger T}k\right] = \frac{\eta^2}{\eta^2c+\rho^2}\frac{c^2}{\rho^2(1-c)^2} + o\left(\frac{1}{\rho^2}\right) + O\left(\frac{1}{\rho^4 n}\right) + O\left(\frac{1}{\rho^2 n^{1.5}}\right),$$

$$\mathbb{E}\left[\frac{\eta\xi}{\gamma_1}k^TA^\dagger h^T\right] = O\left(\frac{1}{\rho^3 n}\right),$$

$$\mathbb{E}\left[\frac{\eta^2\|k\|^2\|h\|^2}{\gamma_1}\right] = \frac{\eta^2}{\eta^2c+\rho^2}\frac{c^3}{\rho^2(1-c)^2} + o\left(\frac{1}{\rho^2}\right) + O\left(\frac{1}{\rho^2 n}\right).$$

After simple algebra, the result follows.

For $c > 1$, we can expand similarly using Theorem 8,

$$
\begin{aligned}
&\varepsilon^T(Z+A)^\dagger(Z+A)^{\dagger T}\varepsilon \\
&= \varepsilon^T\left(A^\dagger + \frac{\eta}{\xi}A^\dagger h^T s^T - \frac{\xi}{\gamma_2}p_2 q_2^T\right)\left(A^{\dagger T} + \frac{\eta}{\xi}shA^{\dagger T} - \frac{\xi}{\gamma_2}q_2 p_2^T\right)\varepsilon \\
&= \varepsilon^T A^\dagger A^{\dagger T}\varepsilon + \frac{2\eta}{\xi}\varepsilon^T \underbrace{A^\dagger s}_{0} hA^{\dagger T}\varepsilon - \frac{2\xi}{\gamma_2}\varepsilon^T A^\dagger q_2 p_2^T\varepsilon \\
&\quad + \frac{\eta^2\|s\|^2}{\xi^2}\varepsilon^T A^\dagger h^T hA^{\dagger T}\varepsilon - \frac{2\eta}{\gamma_2}\varepsilon^T A^\dagger h^T s^T q_2 p_2^T\varepsilon + \frac{\xi^2}{\gamma_2^2}\varepsilon^T p_2 q_2^T q_2 p_2^T\varepsilon
\end{aligned}
$$

We expand the other terms one by one, marking those with zero expectations:

$$
\begin{aligned}
-\frac{2\xi}{\gamma_2}\varepsilon^T A^\dagger q_2 p_2^T\varepsilon &= -\frac{2\xi}{\gamma_2}\varepsilon^T A^\dagger\left(\frac{\eta\|h\|^2}{\xi}s + h^T\right)\left(\frac{\eta^2\|s\|^2}{\xi}hA^{\dagger T} + \eta k^T\right)\varepsilon \\
&= -\frac{2\xi}{\gamma_2}\varepsilon^T A^\dagger h^T\left(\frac{\eta^2\|s\|^2}{\xi}hA^{\dagger T} + \eta k^T\right)\varepsilon \\
&= -\frac{2\eta^2\|s\|^2}{\gamma_2}\varepsilon^T A^\dagger h^T hA^{\dagger T}\varepsilon - \frac{2\eta\xi}{\gamma_2}\varepsilon^T A^\dagger h^T k^T\varepsilon \\
&\overset{E}{=} -\frac{2\eta^2\|s\|^2\tau_\varepsilon^2}{\gamma_2}\|A^\dagger h^T\|^2 - \frac{2\eta\xi\tau_\varepsilon^2}{\gamma_2}k^T A^\dagger h^T
\end{aligned}
$$

$$
\begin{aligned}
-\frac{2\eta}{\gamma_2}\varepsilon^T A^\dagger h^T s^T q_2 p_2^T\varepsilon &= -\frac{2\eta}{\gamma_2}\varepsilon^T A^\dagger h^T s^T\left(\frac{\eta\|h\|^2}{\xi}s + h^T\right)\left(\frac{\eta^2\|s\|^2}{\xi}hA^{\dagger T} + \eta k^T\right)\varepsilon \\
&= -\frac{2\eta}{\gamma_2}\varepsilon^T A^\dagger h^T\left(\frac{\eta\|h\|^2\|s\|^2}{\xi}\right)\left(\frac{\eta^2\|s\|^2}{\xi}hA^{\dagger T} + \eta k^T\right)\varepsilon \\
&= -\frac{2\eta^4\|s\|^4\|h\|^2}{\gamma_2\xi^2}\varepsilon^T A^\dagger h^T hA^{\dagger T}\varepsilon - \frac{2\eta^3\|s\|^2\|h\|^2}{\gamma_2\xi}\varepsilon^T A^\dagger h^T k^T\varepsilon \\
&\overset{E}{=} -\frac{2\eta^4\|s\|^4\|h\|^2\tau_\varepsilon^2}{\gamma_2\xi^2}\|A^\dagger h^T\|^2 - \frac{2\eta^3\|s\|^2\|h\|^2\tau_\varepsilon^2}{\gamma_2\xi}k^T A^\dagger h^T
\end{aligned}
$$

Using the squared norms from Lemmas 15, 16,

$$
\begin{aligned}
\frac{\xi^2}{\gamma_2^2}\varepsilon^T p_2 q_2^T q_2 p_2^T\varepsilon &\overset{E}{=} \frac{\xi^2}{\gamma_2^2}\tau_\varepsilon^2\|p_2\|^2\|q_2\|^2 \\
&= \frac{\xi^2\tau_\varepsilon^2}{\gamma_2^2}\left(\frac{\|h\|^2}{\xi^2}\gamma_2\right)\left(\frac{\eta^4\|s\|^4}{\xi^2}\|A^\dagger h^T\|^2 + \frac{2\eta^3\|s\|^2}{\xi}k^T A^\dagger h^T + \eta^2\|k\|^2\right) \\
&= \tau_\varepsilon^2\left(\frac{\eta^4\|h\|^2\|s\|^4}{\gamma_2\xi^2}\|A^\dagger h^T\|^2 + \frac{2\eta^3\|h\|^2\|s\|^2}{\gamma_2\xi}k^T A^\dagger h^T + \frac{\eta^2\|h\|^2\|k\|^2}{\gamma_2}\right)
\end{aligned}
$$

Similarly, we combine the coefficients: For $\|A^\dagger h^T\|^2$,

$$\begin{aligned}
\tau_\varepsilon^2\left(\frac{\eta^4\|s\|^4\|h\|^2}{\gamma_2\xi^2}-\frac{2\eta^4\|s\|^4\|h\|^2}{\gamma_2\xi^2}-\frac{2\eta^2\|s\|^2}{\gamma_2}+\frac{\eta^2\|s\|^2}{\xi^2}\right) &= \tau_\varepsilon^2\eta^2\|s\|^2\left(\frac{\eta^2\|s\|^2\|h\|^2}{\gamma_2\xi^2}-\frac{2\eta^2\|s\|^2\|h\|^2}{\gamma_2\xi^2}-\frac{2}{\gamma_2}+\frac{1}{\xi^2}\right)\\
&= \tau_\varepsilon^2\eta^2\|s\|^2\left(-\frac{\gamma_2-\xi^2}{\gamma_2\xi^2}-\frac{2}{\gamma_2}+\frac{1}{\xi^2}\right)\\
&= \tau_\varepsilon^2\eta^2\|s\|^2\left(-\frac{\gamma_2-\xi^2}{\gamma_2\xi^2}-\frac{2\xi^2}{\gamma_2\xi^2}+\frac{\gamma_2}{\gamma_2\xi^2}\right)\\
&= -\tau_\varepsilon^2\frac{\eta^2\|s\|^2}{\gamma_2}.
\end{aligned}$$

For term $k^TA^\dagger h^T$,

$$\tau_\varepsilon^2\left(\frac{2\eta^3\|s\|^2\|h\|^2}{\gamma_2\xi}-\frac{2\eta^3\|s\|^2\|h\|^2}{\gamma_2\xi}-\frac{2\eta\xi}{\gamma_2}\right)=-\tau_\varepsilon^2\frac{2\eta\xi}{\gamma_2}.$$

Combining these terms together, we have:

$$\varepsilon^T(Z+A)^\dagger(Z+A)^{\dagger T}\varepsilon \overset{E}{=} \varepsilon^TA^\dagger A^{\dagger T}\varepsilon-\frac{\eta^2\|s\|^2\tau_\varepsilon^2}{\gamma_2}\|A^\dagger h^T\|^2-\frac{2\eta\xi\tau_\varepsilon^2}{\gamma_2}k^TA^\dagger h^T+\frac{\eta^2\|k\|^2\|h\|^2}{\gamma_2}.$$

Similarly, using the relevant lemmas, we have the following:

$$\mathbb{E}\left[\varepsilon^TA^\dagger A^{\dagger T}\varepsilon\right]=\tau_\varepsilon^2\frac{d}{\rho^2(c-1)}+o\left(\frac{n}{\rho^2}\right),$$

$$\mathbb{E}\left[\frac{\eta^2\|s\|^2}{\gamma_2}\|A^\dagger h^T\|^2\right]=\frac{\eta^2}{\eta^2+\rho^2}\frac{c^2}{\rho^2(c-1)^2}+o\left(\frac{1}{\rho^2}\right)+O\left(\frac{1}{\rho^4 n}\right),$$

$$\mathbb{E}\left[\frac{\eta\xi}{\gamma_2}k^TA^\dagger h^T\right]=O\left(\frac{1}{\rho^3 n}\right),$$

$$\mathbb{E}\left[\frac{\eta^2\|k\|^2\|h\|^2}{\gamma_2}\right]=\frac{\eta^2}{\eta^2+\rho^2}\frac{c}{\rho^2(c-1)^2}+o\left(\frac{1}{\rho^2}\right)+O\left(\frac{1}{\rho^2 n}\right).$$

After simple algebra, the result follows.

■

Appendix K. Probability Lemmas

Proposition 3 *If $u, v \in \mathbb{R}^d$ are fixed unit norm vector and $A \in \mathbb{R}^{d\times n}$ is a Gaussian matrix with IID $\mathcal{N}(0,1)$ entries. Then we have that*

$$\mathbb{E}[(u^TAA^\dagger v)^2]=\frac{n}{d(d+2)}\left[(u^Tv)^2(n+2)+\frac{(1-(u^Tv)^2)(d-n)}{d-1}\right]=\frac{1}{c^2}(u^Tv)^2+o(1)$$

Proof Let $P := AA^\dagger$. This is the orthogonal projection matrix onto the column space of A, denoted $C(A) = \text{Range}(A)$. The subspace $C(A)$ is an n-dimensional subspace of $\mathbb{R}^d$. Because the entries A_{ij} are i.i.d. $\mathcal{N}(0,1)$, the distribution of the random subspace $C(A)$ is isotropic (or rotationally invariant). Consequently, the distribution of the random projection matrix P is also rotationally invariant. That is, for any fixed $d \times d$ orthogonal matrix Q, the distribution of QPQ^T is the same as the distribution of P.

We are interested in $\mathbb{E}[(u^TPv)^2]$. Let θ be the angle between u and v, such that $\cos(\theta) = u^Tv$ (since they are unit vectors). Due to the rotational invariance of the distribution of P, we can choose an orthonormal basis without loss of generality. Let Q be an orthogonal matrix such that $u' = Qu = e_1 = (1, 0, \dots, 0)^T$ and $v' = Qv$ lies in the span of e_1 and e_2. Specifically, $v' = \cos(\theta)e_1 + \sin(\theta)e_2$. Let $P' = QPQ^T$. P' has the same distribution as P. Then,

$$u^TPv = (Q^Tu')^TP(Q^Tv') = (u')^T(QPQ^T)v' = (u')^TP'v'$$

Substituting $u' = e_1$ and $v' = \cos(\theta)e_1 + \sin(\theta)e_2$:

$$\begin{aligned} u^TPv &= e_1^TP'(\cos(\theta)e_1 + \sin(\theta)e_2) \\ &= \cos(\theta)(e_1^TP'e_1) + \sin(\theta)(e_1^TP'e_2) \\ &= \cos(\theta)P'_{11} + \sin(\theta)P'_{12} \end{aligned}$$

where P'_{ij} are the elements of P'. Since P' has the same distribution as P, we can drop the prime for calculating expectations involving the elements. Let $X = u^TPv$. We need $\mathbb{E}[X^2]$.

$$\begin{aligned} \mathbb{E}[X^2] &= \mathbb{E}[(\cos(\theta)P_{11} + \sin(\theta)P_{12})^2] \\ &= \mathbb{E}[\cos^2(\theta)P_{11}^2 + \sin^2(\theta)P_{12}^2 + 2\cos(\theta)\sin(\theta)P_{11}P_{12}] \\ &= \cos^2(\theta)\mathbb{E}[P_{11}^2] + \sin^2(\theta)\mathbb{E}[P_{12}^2] + 2\cos(\theta)\sin(\theta)\mathbb{E}[P_{11}P_{12}] \end{aligned}$$

Calculation of Moments We need to compute $\mathbb{E}[P_{11}^2]$, $\mathbb{E}[P_{12}^2]$, and $\mathbb{E}[P_{11}P_{12}]$.

Consider a reflection matrix R that maps e_2 to $-e_2$ and leaves other basis vectors unchanged (i.e., $R = \text{diag}(1, -1, 1, \dots, 1)$). Since the distribution of P is isotropic, it is invariant under reflection. Let $P^* = RPR^T = RPR$. P^* has the same distribution as P. The components are related:

$$P_{11}^* = (RPR)_{11} = R_{11}P_{11}R_{11} = P_{11}$$

and

$$P_{12}^* = (RPR)_{12} = R_{11}P_{12}R_{22} = (1)P_{12}(-1) = -P_{12}.$$

Therefore,

$$\mathbb{E}[P_{11}P_{12}] = \mathbb{E}[P_{11}^*P_{12}^*] = \mathbb{E}[P_{11}(-P_{12})] = -\mathbb{E}[P_{11}P_{12}].$$

This implies $2\mathbb{E}[P_{11}P_{12}] = 0$, so $\mathbb{E}[P_{11}P_{12}] = 0$.

The diagonal element $P_{11} = e_1^TPe_1 = ||Pe_1||_2^2$ represents the squared norm of the projection of the fixed unit vector e_1 onto the random n-dimensional subspace $C(A)$. This variable follows a Beta distribution:

$$P_{11} \sim \text{Beta}\left(\frac{n}{2}, \frac{d-n}{2}\right)$$

The mean and variance of a Beta(α, β) distribution are $\frac{\alpha}{\alpha+\beta}$ and $\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$, respectively. Here, $\alpha = n/2$ and $\beta = (d-n)/2$, so $\alpha + \beta = d/2$.

$$\mathbb{E}[P_{11}] = \frac{n/2}{d/2} = \frac{n}{d}$$

Next

$$\mathrm{Var}(P_{11}) = \frac{(n/2)((d-n)/2)}{(d/2)^2(d/2+1)} = \frac{n(d-n)/4}{(d^2/4)((d+2)/2)} = \frac{n(d-n)\cdot 8}{4d^2(d+2)} = \frac{2n(d-n)}{d^2(d+2)}$$

Now we find $\mathbb{E}[P_{11}^2]$ using $\mathbb{E}[P_{11}^2] = \mathrm{Var}(P_{11}) + (\mathbb{E}[P_{11}])^2$:

$$\begin{aligned}\mathbb{E}[P_{11}^2] &= \frac{2n(d-n)}{d^2(d+2)} + \left(\frac{n}{d}\right)^2 \\ &= \frac{2n(d-n) + n^2(d+2)}{d^2(d+2)} \\ &= \frac{2nd - 2n^2 + n^2 d + 2n^2}{d^2(d+2)} \\ &= \frac{2nd + n^2 d}{d^2(d+2)} = \frac{nd(2+n)}{d^2(d+2)} \\ &= \frac{n(n+2)}{d(d+2)}.\end{aligned}$$

We use the property that P is a projection matrix, so $P^2 = P$. The trace is $\mathrm{Tr}(P) = n$. Also $\mathrm{Tr}(P^2) = \mathrm{Tr}(P) = n$. We can write $\mathrm{Tr}(P^2) = \mathrm{Tr}(PP^T)$ since P is symmetric.

$$\mathrm{Tr}(P^2) = \sum_{i=1}^{d}\sum_{j=1}^{d}(P_{ij})^2$$

Taking the expectation:

$$\mathbb{E}[\mathrm{Tr}(P^2)] = \mathbb{E}\left[\sum_{i,j} P_{ij}^2\right] = \sum_{i,j}\mathbb{E}[P_{ij}^2] = n$$

By rotational symmetry, $\mathbb{E}[P_{ii}^2]$ is the same for all i, and $\mathbb{E}[P_{ij}^2]$ is the same for all $i \neq j$.

$$\sum_{i=1}^{d}\mathbb{E}[P_{ii}^2] + \sum_{i\neq j}\mathbb{E}[P_{ij}^2] = n$$

There are d diagonal terms and $d(d-1)$ off-diagonal terms.

$$d\,\mathbb{E}[P_{11}^2] + d(d-1)\,\mathbb{E}[P_{12}^2] = n$$

Substitute the value for $\mathbb{E}[P_{11}^2]$ (assuming $d > 1$):

$$d\left(\frac{n(n+2)}{d(d+2)}\right) + d(d-1)\,\mathbb{E}[P_{12}^2] = n$$

$$\frac{n(n+2)}{d+2} + d(d-1)\,\mathbb{E}[P_{12}^2] = n$$

$$d(d-1)\,\mathbb{E}[P_{12}^2] = n - \frac{n(n+2)}{d+2} = \frac{n(d+2) - n(n+2)}{d+2} = \frac{nd + 2n - n^2 - 2n}{d+2} = \frac{n(d-n)}{d+2}$$

$$\mathbb{E}[P_{12}^2] = \frac{n(d-n)}{d(d-1)(d+2)}$$

Substitute the moments back into the expression for $\mathbb{E}[X^2]$:

$$\mathbb{E}[X^2] = \cos^2(\theta)\mathbb{E}[P_{11}^2] + \sin^2(\theta)\mathbb{E}[P_{12}^2] + 2\cos(\theta)\sin(\theta)\cdot 0$$

Using $\cos(\theta) = u^T v$, $\cos^2(\theta) = (u^T v)^2$, and $\sin^2(\theta) = 1 - \cos^2(\theta) = 1 - (u^T v)^2$:

$$\begin{aligned}\mathbb{E}[(u^T A A^\dagger v)^2] &= (u^T v)^2 \left(\frac{n(n+2)}{d(d+2)}\right) + (1 - (u^T v)^2)\left(\frac{n(d-n)}{d(d-1)(d+2)}\right) \\ &= \frac{n}{d(d+2)}\left[(u^T v)^2 (n+2) + \frac{(1-(u^T v)^2)(d-n)}{d-1}\right]\end{aligned}$$

■

Lemma 33 *Let $a \neq 0$ be a constant and suppose that $\zeta = a + o(1)$ as $n \to \infty$. Then,*

$$\frac{1}{\zeta} = \frac{1}{a} + o(1).$$

Proof We can write

$$\zeta = a + o(1) = a\left(1 + \frac{o(1)}{a}\right) = a\left[1 + o(1)\right].$$

Taking the reciprocal gives

$$\frac{1}{\zeta} = \frac{1}{a\left(1+o(1)\right)} = \frac{1}{a}\cdot\frac{1}{1+o(1)}.$$

Since for any $u = o(1)$ we have the expansion

$$\frac{1}{1+u} = 1 - u + o(u),$$

it follows that

$$\frac{1}{1+o(1)} = 1 - o(1) = 1 + o(1).$$

Thus,

$$\frac{1}{\zeta} = \frac{1}{a}\left[1 + o(1)\right] = \frac{1}{a} + o(1).$$

■

Lemma 34 *Let X be a random variable satisfying*

$$E[X] = a > 0 \quad \textit{and} \quad \mathrm{Var}(X) = o(1),$$

and assume that X is bounded away from zero with high probability. Then

$$\mathrm{Var}\left(\frac{1}{X}\right) = \frac{1}{a^4}\mathrm{Var}(X) + o\left(\mathrm{Var}(X)\right),$$

so in particular, $\mathrm{Var}(1/X) = o(1)$.

Proof Since $\mathrm{Var}(X) = o(1)$, the random variable X is highly concentrated about its mean a. Consider the function $f(x) = 1/x$, which is differentiable at $x = a$ with derivative

$$f'(a) = -\frac{1}{a^2}.$$

By the first-order Taylor expansion (or the delta method), we have

$$\frac{1}{X} = f(X) \approx f(a) + f'(a)(X - a) = \frac{1}{a} - \frac{1}{a^2}(X - a)$$

for X near a. Since $\mathbb{E}[X - a] = 0$, the mean of $1/X$ is approximately

$$E\left[\frac{1}{X}\right] \approx \frac{1}{a}.$$

Now, the variance of $1/X$ can be approximated by considering the linear term:

$$\mathrm{Var}\left(\frac{1}{X}\right) \approx \mathrm{Var}\left(-\frac{1}{a^2}(X - a)\right) = \frac{1}{a^4}\mathrm{Var}(X).$$

The remainder term in the Taylor expansion, which is of higher order in $(X - a)$, contributes a term that is $o(\mathrm{Var}(X))$. Hence, we obtain

$$\mathrm{Var}\left(\frac{1}{X}\right) = \frac{1}{a^4}\mathrm{Var}(X) + o\left(\mathrm{Var}(X)\right).$$

Since $\mathrm{Var}(X) = o(1)$, it follows that $\mathrm{Var}(1/X) = o(1)$ as well. ■

Lemma 35 *Let A and B be any random variables with finite variances $V(A) = \mathrm{Var}(A)$ and $V(B) = \mathrm{Var}(B)$. Then,*

$$\mathrm{Var}(A + B) \leq \left(\sqrt{V(A)} + \sqrt{V(B)}\right)^2.$$

Proof Recall that

$$\mathrm{Var}(A + B) = \mathrm{Var}(A) + \mathrm{Var}(B) + 2\ \mathrm{Cov}(A, B).$$

By the Cauchy–Schwarz inequality, we have

$$|\mathrm{Cov}(A, B)| \leq \sqrt{V(A)V(B)}.$$

Thus,

$$\mathrm{Var}(A + B) \leq V(A) + V(B) + 2\sqrt{V(A)V(B)} = \left(\sqrt{V(A)} + \sqrt{V(B)}\right)^2.$$

■

Lemma 36 *Let A and B be random variables with finite variances, and denote $a = \mathbb{E}[A], \quad b = \mathbb{E}[B]$, and let $\tilde{A} = A - a, \quad \tilde{B} = B - b$. Then,*

$$\mathrm{Var}(AB) \le \left(|a|\sqrt{\mathrm{Var}(B)} + |b|\sqrt{\mathrm{Var}(A)} + \sqrt{\mathrm{Var}(A)\mathrm{Var}(B)}\right)^2.$$

In particular, if A and B concentrate to a and b respectively (i.e. if $\mathrm{Var}(A),\ \mathrm{Var}(B) \to 0$), then

$$\mathrm{Var}(AB) = a^2\,\mathrm{Var}(B) + b^2\,\mathrm{Var}(A) + o\left(\mathrm{Var}(A) + \mathrm{Var}(B)\right).$$

Proof Write

$$AB = (a + \tilde{A})(b + \tilde{B}) = ab + a\,\tilde{B} + b\,\tilde{A} + \tilde{A}\tilde{B}.$$

Thus,

$$AB - ab = a\,\tilde{B} + b\,\tilde{A} + \tilde{A}\tilde{B}.$$

Taking the L^2 norm (which is the square root of the variance) and applying the triangle inequality yields

$$\sqrt{\mathrm{Var}(AB)} = \|AB - ab\|_2 \le |a|\|\tilde{B}\|_2 + |b|\|\tilde{A}\|_2 + \|\tilde{A}\tilde{B}\|_2.$$

Since $\|\tilde{A}\|_2 = \sqrt{\mathrm{Var}(A)}$ and $\|\tilde{B}\|_2 = \sqrt{\mathrm{Var}(B)}$, and by Cauchy–Schwarz,

$$\|\tilde{A}\tilde{B}\|_2 \le \sqrt{\mathrm{Var}(A)\mathrm{Var}(B)},$$

we have

$$\sqrt{\mathrm{Var}(AB)} \le |a|\sqrt{\mathrm{Var}(B)} + |b|\sqrt{\mathrm{Var}(A)} + \sqrt{\mathrm{Var}(A)\mathrm{Var}(B)}.$$

Squaring both sides gives the stated bound:

$$\mathrm{Var}(AB) \le \left(|a|\sqrt{\mathrm{Var}(B)} + |b|\sqrt{\mathrm{Var}(A)} + \sqrt{\mathrm{Var}(A)\mathrm{Var}(B)}\right)^2.$$

In the situation where $\mathrm{Var}(A)$ and $\mathrm{Var}(B)$ are small, the term $\sqrt{\mathrm{Var}(A)\mathrm{Var}(B)}$ is negligible relative to the linear terms, and hence

$$\mathrm{Var}(AB) = a^2\,\mathrm{Var}(B) + b^2\,\mathrm{Var}(A) + o\left(\mathrm{Var}(A) + \mathrm{Var}(B)\right).$$

■

Lemma 37 *Let $x \in \mathbb{R}^d$ be uniformly distributed on the unit sphere S^{d-1} and let $Q \in \mathbb{R}^{d\times d}$ be a fixed orthogonal matrix. For fixed indices $i, j \in \{1, \ldots, d\}$, define*

$$Y = x_i\,x_j\,(Qx)_i.$$

Then, as $d \to \infty$,

$$\mathrm{Var}(Y) = O\left(\frac{1}{d^3}\right).$$

Proof Since x is uniformly distributed on S^{d-1}, its coordinates satisfy (by symmetry)

$$E[x_k] = 0, \quad E[x_k^2] = \frac{1}{d}, \quad \text{and} \quad E[x_k^4] = O\left(\frac{1}{d^2}\right)$$

for any $k \in \{1, \dots, d\}$. Moreover, for distinct indices $k \neq \ell$ one may show that

$$E[x_k^2 \, x_\ell^2] = O\left(\frac{1}{d^2}\right).$$

Since Q is orthogonal, each row of Q is a unit vector; in particular, for any fixed i we have

$$(Qx)_i = Q_{ii} \, x_i + \sum_{k \neq i} Q_{ik} \, x_k,$$

and a routine calculation shows that

$$\mathrm{Var}\left((Qx)_i\right) = \frac{1}{d}.$$

Thus, each coordinate of Qx is also of order $1/\sqrt{d}$.

We now write

$$Y = x_i \, x_j \, (Qx)_i = Q_{ii} \, x_i^2 x_j + x_i \, x_j \sum_{k \neq i} Q_{ik} \, x_k.$$

For a uniformly random vector on a sphere it is know that

$$x_i^2 x_j = O\left(\frac{1}{d} \cdot \frac{1}{\sqrt{d}}\right) = O\left(\frac{1}{d^{3/2}}\right)$$

and similarly,

$$x_i \, x_j \, x_k = O\left(\frac{1}{d^{3/2}}\right).$$

Thus

$$\mathbb{E}[Y^2] = O\left(\frac{1}{d^3}\right).$$

Since by symmetry $\mathbb{E}[Y] = 0$, we conclude

$$\mathrm{Var}(Y) = E[Y^2] = O\left(\frac{1}{d^3}\right).$$

■