Minimum-Norm Interpolation Under Covariate Shift

Neil Mallinar*
UC San Diego
nmallina@ucsd.edu

Austin Zane*
UC Berkeley
austin.zane@berkeley.edu

Spencer Frei
UC Davis
sfrei@ucdavis.edu

Bin Yu
UC Berkeley
binyu@berkeley.edu

April 2, 2024

Abstract

Transfer learning is a critical part of real-world machine learning deployments and has been extensively studied in experimental works with overparameterized neural networks. However, even in the simplest setting of linear regression a notable gap still exists in the theoretical understanding of transfer learning. In-distribution research on high-dimensional linear regression has led to the identification of a phenomenon known as *benign overfitting*, in which linear interpolators overfit to noisy training labels and yet still generalize well. This behavior occurs under specific conditions on the source covariance matrix and input data dimension. Therefore, it is natural to wonder how such high-dimensional linear models behave under transfer learning. We prove the first non-asymptotic excess risk bounds for benignly-overfit linear interpolators in the transfer learning setting. From our analysis, we propose a taxonomy of *beneficial* and *malignant* covariate shifts based on the degree of overparameterization. We follow our analysis with empirical studies that show these beneficial and malignant covariate shifts for linear interpolators on real image data, and for fully-connected neural networks in settings where the input data dimension is larger than the training sample size.

1 Introduction

Practical deployments of machine learning models are almost always in a transfer learning setting, where models trained on a *source data distribution* with noisy labels are expected to perform well on a different *target data distribution*, referred to as the "out-of-distribution" (OOD) dataset [Ogl+22; DAm+22]. There have been many experimental works on transfer learning with complex models and datasets [Rec+19; Koh+21; Mil+21; Hen+21; Wen+22; Lia+23], but remarkably fewer attempts to study it theoretically, even in the simplest case of linear models which have been of great interest in recent years [Dwi+20; Bar+20; Has+22; TB23].

There has been an extensive "in-distribution" (ID) theoretical interest in high-dimensional linear regression

*Equal contribution.

and specifically *interpolation*, meaning a model reaches zero training loss. Frameworks such as "benign overfitting", or "harmless interpolation" [Bar+20; Mut+20] emerged as an attempt to explain why interpolating neural networks often do not overfit catastrophically [Zha+17]. They found that, in specific cases, overfitting can be "benign", meaning that a model interpolates noisy training labels and yet has vanishing excess risk. In linear regression, this occurs if and only if the training (source) covariance matrix satisfies very specific conditions. Under these conditions, the minimum-norm interpolator (MNI) approximately acts like a ridge regression solution.

This sparked an initial wave of in-distribution theoretical research into benign overfitting in high-dimensional linear models [CLB22; TB23; CL23], kernel regression [RZ19; Haa+23; BS23], and even some shallow neural networks [FCB22; Kou+23; KYS23; Xu+24]. Although these works were motivated by a desire to understand overfitting in modern deep learning, recent works have shown that in many practical settings of interest, overfitting is not benign [Mal+22; Haa+23; Lai+23]. Thus, deeper investigations into the generalization behavior of overfit models are warranted.

Given the increasing prevalence of overparameterized models, it is natural to ask how such models perform in the transfer learning setting. There have been some efforts to answer this in the theoretically tractable cases of linear regression and random feature and kernel regression [PMW22; Wan23]. However, these works either provide asymptotic bounds that require the training sample size and data dimension to go to infinity at the same rate [TAP21], study minimax settings which only considers worst-case risk [LHL21], or focus on augmented gradient-based training algorithms, like importance weighting [Wan+22].

Summary of contributions. In this paper, we investigate the generalization behavior of the minimum ℓ_2-norm linear interpolator (MNI) under distribution shifts when the source distribution satisfies the conditions necessary for benign overfitting. We summarize our main contributions as follows.

- We provide the first non-asymptotic, instance-wise risk bounds for covariate shifts in interpolating linear regression when the source covariance matrix satisfies benign overfitting conditions and commutes with the target covariance matrix.
- We use our risk bounds to propose a taxonomy of covariate shifts for the MNI. We show how the ratio of target eigenvalues to source eigenvalues and the degree of overparameterization affect whether a shift is *beneficial* or *malignant*, meaning OOD risk is better or worse than ID risk, respectively.
- We empirically show that our taxonomy of shifts holds: (1) for the MNI on real image data under natural shifts like blur (a beneficial shift) and noise (a malignant shift), underscoring the significance of our findings beyond the idealized source and target covariances for which our theory is applicable; (2) for neural networks in settings where the input data dimension is larger than the training sample size, showing that our findings for the MNI are also reflective of the behavior of more complex models.

1.1 Prior work and comparisons to this work

Excess risk analysis under distribution shifts: Tripuraneni, Adlam, and Pennington [TAP21] give an asymptotic analysis of high-dimensional random feature regression in covariate shift. They require the number of samples, n, data dimension, p, and random feature dimension to go to ∞ at the same rate. In contrast, our non-asymptotic analysis considers finite sample cases and differing rates. This allows us to draw new conclusions about how the *degree of overparameterization* changes the way in which interpolating linear models exhibit out-of-distribution (OOD) generalization. Additionally, our bounds let us analyze any

sequence of eigenvalues for the target feature covariance matrix, which is not possible within their framework.

Lei, Hu, and Lee [LHL21] study linear regression under distribution shifts in the minimax setting. Their minimax bounds consider the worst-case risk over an ℓ_2-ball of target models, whereas we compute risk bounds specific to any model instantiation, with no restriction on the target model class. Furthermore, their experimental results only consider the underparameterized regime.

Several other works study OOD generalization in more distant settings. Wang et al. [Wan+22] study linear interpolators for classification, when trained with gradient descent and importance weighting, whereas we consider the closed-form MNI for linear regression. Simchowitz et al. [Sim+23] study covariate shifts when the target function class is the sum of two other function classes, and shifts are defined with regard to metric entropy between classes, whereas we focus on well-specified linear models. Pathak, Ma, and Wainwright [PMW22], Ma, Pathak, and Wainwright [MPW23], and Feng et al. [Fen+23] consider covariate shift in kernel regression based on likelihood ("importance") ratios between source and target distributions while we consider source and target eigenvalue ratios which offer granular insights into feature scale changes whereas likelihood ratios capture shifts that affect the global data distribution. Pathak, Ma, and Wainwright [PMW22] and Ma, Pathak, and Wainwright [MPW23] also analyze worst-case, minimax risk for nonparametric function classes. Finally, we note that risk bounds in these prior works do not sufficiently account for the behavior of the high-rank covariance tail that benign overfitting requires.

Experimental work on distribution shifts: Hendrycks and Dietterich [HD19] propose the CIFAR-10C dataset as an OOD counterpart to CIFAR-10, featuring test set images corrupted by visual filters like blurs and noises. Koh et al. [Koh+21] present benchmarks on more realistic datasets with modern models that can be seen "in-the-wild". Miller et al. [Mil+21] experimentally show a linear relationship between ID accuracy and OOD accuracy for a wide range of modern neural networks and datasets, though their results show ID accuracy is almost always better than OOD accuracy. On a subset of CIFAR-10C, we find settings in which OOD accuracy is *better* than ID accuracy for linear interpolators.

Benign overfitting "in-distribution": Bartlett et al. [Bar+20] propose benign overfitting, give a non-asymptotic analysis of the MNI, and show specific, necessary conditions under which the MNI achieves zero excess risk in-distribution. Tsigler and Bartlet [TB23] extend this work by considering benign overfitting in the case of ridge regression. Our proof techniques follow most closely to the ideas presented in these two papers for the in-distribution setting. Frei, Chatterji, and Bartlett [FCB22] show benign overfitting in shallow non-linear MLPs trained with gradient descent on the logistic loss if the data dimension grows faster than the number of training samples. Mallinar et al. [Mal+22] experimentally show that interpolating neural networks do not benignly overfit due to the low input data dimension. Our experiments build on this by looking at settings in which $p > n$ and $n < p$ where n is the training sample size and p is the input data dimension. Other works study benign overfitting under a variety of conditions [Kou+23; CL23; Fre+23].

2 Preliminaries

We extend notations in Bartlett et al. [Bar+20] and Tsigler and Bartlet [TB23] to the transfer learning setting with OOD generalization risk as our performance metric. Appendix A formalizes our setting of linear regression under distribution shift, and we provide necessary details here.

2.1 Covariate assumptions

Consider source and target distributions, $\mathcal{D}_\mathsf{s}$ and $\mathcal{D}_\mathsf{t}$, both over $(x, y) \in \mathbb{R}^p \times \mathbb{R}$. The source design matrix, $X \in \mathbb{R}^{n\times p}$, has rows $x \overset{iid}{\sim} \mathcal{D}_\mathsf{s}$ such that $\mathbb{E}_{\mathcal{D}_\mathsf{s}}[x] = 0$. We consider the overparameterized regime with $n < p$ (see e.g. Bartlett et al. [Bar+20]).

Denote the source and target population covariance matrices by Σ_s and $\Sigma_\mathsf{t} \in \mathbb{R}^{p\times p}$. We assume that there exists an orthonormal basis in which both matrices are diagonal, but show in Section 4 with experiments that our results hold even when this is violated. Formally,

$$\Sigma_\mathsf{s} = \mathop{\mathbb{E}}_{\mathcal{D}_\mathsf{s}}[xx^T] = \mathrm{diag}(\lambda_1, ..., \lambda_p),$$
$$\Sigma_\mathsf{t} = \mathop{\mathbb{E}}_{\mathcal{D}_\mathsf{t}}[xx^T] = \mathrm{diag}(\tilde{\lambda}_1, ..., \tilde{\lambda}_p),$$

where $\lambda_1 \geq ... \geq \lambda_p > 0$, $\tilde{\lambda}_i \geq 0$ for all i, and $\sum_i \lambda_i \tilde{\lambda}_i < \infty$.

Assume further that the rows of the whitened data matrix $Z := X\Sigma_\mathsf{s}^{-1/2}$ are mean-zero i.i.d. σ_x^2-subgaussian random vectors with independent components. Subgaussianity is a common assumption in statistical learning theory that encompasses a wide array of distributions of interest [Ver18].

2.2 Linear regression models for source and target data

Denote the source response vector by $\boldsymbol{y}_\mathsf{s} \in \mathbb{R}^n$. Assume a linear regression model $\boldsymbol{y}_\mathsf{s} = X\theta^*_\mathsf{s} + \boldsymbol{\varepsilon}_\mathsf{s}$, where the noise vector $\boldsymbol{\varepsilon}_\mathsf{s}$ has independent components with mean 0 and variance $v^2_{\boldsymbol{\varepsilon}_\mathsf{s}}$. For an observation pair $(x, y) \sim \mathcal{D}_\mathsf{t}$, the target responses are defined as $y = x^T\theta^*_\mathsf{t} + \varepsilon_\mathsf{t}$, where $\theta^*_\mathsf{t} \in \mathbb{R}^p$ and the noise vector ε_t has mean 0 and variance $v^2_{\varepsilon_\mathsf{t}}$. Note that we use the same (x, y) for source and target data, but will differentiate between the two by explicitly denoting the distribution from which the pair is drawn.

2.3 Minimum-norm interpolator and target excess risk

Given a source data matrix X, the minimum-norm interpolator (MNI) for any vector $\xi \in \mathbb{R}^n$ is defined as

$$\begin{aligned}\widehat{\theta}(\xi) &:= \mathrm{argmin}\left\{\|\theta\|^2 : X\theta = \xi\right\} \\ &= X^T(XX^T)^{-1}\xi.\end{aligned}$$

If we consider $\xi = \boldsymbol{y}_\mathsf{s}$, then we recover the MNI for the labels given by the response model, but our analysis will also involve implicit MNIs for different label vectors in $\mathbb{R}^n$.

The quantity that we seek to bound is the excess risk on the target distribution, which we define for an estimator $\theta \in \mathbb{R}^p$ as,

$$R(\theta, \mathcal{D}_\mathsf{t}) := \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}_\mathsf{t}}\left[\left(y - x^T\theta\right)^2 - \left(y - x^T\theta^*_\mathsf{t}\right)^2\right]. \tag{1}$$

We now derive bounds for the target excess risk and its expectation over the source response noise. The proof of the following can be found in Appendix C.

Theorem 2.1. *(Target excess risk decomposition) The excess risk of the MNI trained on the source data, when evaluated on the target distribution, satisfies*

$$R(\widehat{\theta}(\boldsymbol{y}_\mathsf{s}), \mathcal{D}_\mathsf{t}) \leq 4B_1 + 4B_2 + 2V_{\boldsymbol{\varepsilon}_\mathsf{s}}, \tag{2}$$

and

$$\mathop{\mathbb{E}}_{\varepsilon_{\mathsf{s}}} R(\widehat{\theta}(\boldsymbol{y}_{\mathsf{s}}), \mathcal{D}_{\mathsf{t}}) = B_1 + B_2 + \mathop{\mathbb{E}}_{\varepsilon_{\mathsf{s}}} V_{\varepsilon_{\mathsf{s}}} + 2(\theta^*_{\mathsf{t}} - \theta^*_{\mathsf{s}})^\top \Sigma_{\mathsf{t}} (\theta^*_{\mathsf{s}} - \widehat{\theta}(X\theta^*_{\mathsf{s}})),$$

where we define

$$B_1 := \|\theta^*_{\mathsf{s}} - \theta^*_{\mathsf{t}}\|^2_{\Sigma_{\mathsf{t}}}, \tag{3}$$

$$B_2 := \|\theta^*_{\mathsf{s}} - \widehat{\theta}(X\theta^*_{\mathsf{s}})\|^2_{\Sigma_{\mathsf{t}}}, \tag{4}$$

$$V_{\varepsilon_{\mathsf{s}}} := \|\widehat{\theta}(\varepsilon_{\mathsf{s}})\|^2_{\Sigma_{\mathsf{t}}}, \tag{5}$$

and $\|x\|^2_M := x^\top M x$.

We observe that B_1 is a deterministic model shift term and that no further analysis can improve its dependency on θ^*_{s}, θ^*_{t}, or Σ_{t}. The cross-term, $(\theta^*_{\mathsf{t}} - \theta^*_{\mathsf{s}})^\top \Sigma_{\mathsf{t}} (\theta^*_{\mathsf{s}} - \widehat{\theta}(X\theta^*_{\mathsf{s}}))$, is dominated by the bias and variance as evidenced by the upper bound. Therefore we focus our analysis on B_2 and $V_{\varepsilon_{\mathsf{s}}}$. A useful normalized version of $V_{\varepsilon_{\mathsf{s}}}$ is defined by

$$V = \mathop{\mathbb{E}}_{\varepsilon_{\mathsf{s}}} \left[V_{\varepsilon_{\mathsf{s}}} / v^2_{\varepsilon_{\mathsf{s}}} \right]. \tag{6}$$

Note that B_2, V are reminiscent of the ID bias and variance in prior work [Bar+20; TB23].

2.4 Separation of components and effective ranks

For an index k, we define the following quantities related to the effective rank of the tail of Σ_{s} [TB23]:

$$\rho_k = \frac{\sum_{i>k} \lambda_i}{n\lambda_{k+1}}, \qquad R_k = \frac{(\sum_{i>k} \lambda_i)^2}{\sum_{i>k} \lambda_i^2}.$$

ρ_k measures the ratio of the energy of the source covariance tail to the number of training data observations, after normalizing the tail eigenvalues. R_k measures the quantity of noisy features and how evenly distributed their eigenvalues are. It is minimized when there is only one nonzero eigenvalue and maximized when there are many equal eigenvalues.

Benign overfitting occurs if the MNI is overfit to noisy training labels and yet ID excess risk decays to zero. The central finding of Bartlett et al. [Bar+20] is that the only way benign overfitting happens for the MNI is if the following occurs: (1) there exists a $k^* = \min\{k \ : \ \rho_k \geq b\}$ for a universal constant $b > 1$, meaning that the last $p - k^*$ components of Σ_{s} have a high effective rank relative to the number of training samples, n; (2) the magnitudes of the bottom $p - k^*$ eigenvalues are small relative to the top k^*; and (3) $k^* \ll n$. More formally, consider quantities $p = p(n)$, a sequence of source covariance matrices $\Sigma_n = \mathrm{diag}(\lambda_1, \cdots, \lambda_p)$, $k^* = k^*(n)$ as defined above, $R_{k^*} = R_{k*}(\Sigma_n)$, and $\rho_k = \rho_k(\Sigma_n)$. A sufficient condition for benign overfitting is,

$$\lim_{n\to\infty} \rho_0 = \lim_{n\to\infty} \frac{k^*}{n} = \lim_{n\to\infty} \frac{n}{R_{k^*}} = 0. \tag{7}$$

If this occurs, then the MNI behaves similarly to an estimator with two components. One component has variance similar to the ordinary least squares (OLS) estimator in k^* dimensions and bias similar to the ridge

regression solution with ridge parameter proportional to $\sum_{i>k} \lambda_i$, a sort of data-induced regularization. The other component is a high-dimensional component, which has vanishing variance when the data is sufficiently high-dimensional and a bias which is proportional to $\sum_{i>k} \lambda_i (\theta^*_{\mathsf{s}})_i^2$ [TB23]. From these conditions, we see that the top k^* components are like "signal" components of the data and the bottom $p - k^*$ components are "noise" components.

2.5 Spiked covariance models

We will consider a special case of the (k, ϵ)-spike model, a canonical covariance structure that exhibits benign overfitting for the MNI [CLB22; CL23], to experimentally show properties of interest.

Definition 1 ((k, δ, ϵ)-spike model)**.** *For a source distribution $\mathcal{D}_{\mathsf{s}}$, $\delta > 0$ and $\epsilon > 0$ such that $\delta \gg \epsilon$, let*

$$\mathop{\mathbb{E}}_{x \sim \mathcal{D}_{\mathsf{s}}}[xx^T] = \operatorname{diag}(\underbrace{\lambda_1, \cdots, \lambda_k}_{=\delta}, \underbrace{\lambda_{k+1}, \cdots, \lambda_p}_{=\epsilon}).$$

In this simplified setting, there are k high-energy "signal" directions and $p-k$ low-energy "noise" directions. For a target distribution $\mathcal{D}_{\mathsf{t}}$, we use different hyperparameters $\tilde{k}, \tilde{\delta}, \tilde{\epsilon}$ to similarly characterize a shifted covariance matrix.

3 Main Theorems

This section provides upper and lower bounds for the variance and bias terms in Equation 6 and Equation 4, respectively. We start with the bounds for the variance term. Appendix D gives a high-level overview of our proof techniques. Subsequent appendices provide complete proofs. Appendix E contains a proof of the following theorem.

Theorem 3.1. *(Upper and lower bounds for the variance term) There exist universal constants $b, c_1 > 1$ given in Lemma B.1, a universal constant c_2 given in Lemma B.4 and a constant $c > 1$ that only depends on σ_x, c_1, c_2, such that for $k \in (0, n/c)$, with probability at least $1 - 10e^{-n/c}$,*

$$V \geq \frac{1}{cn} \sum_{i=1}^{p} \frac{\tilde{\lambda}_i}{\lambda_i} \min\left(1, \frac{\lambda_i^2}{\lambda_{k+1}^2 (\rho_k + 1)^2}\right) := \underline{V}. \tag{8}$$

If in addition $\rho_k \geq b$, with probability $1 - 7e^{-n/c}$,

$$V/c \leq \frac{1}{n} \sum_{i=1}^{k} \frac{\tilde{\lambda}_i}{\lambda_i} + n \frac{\sum_{i>k} \tilde{\lambda}_i \lambda_i}{(\sum_{i>k} \lambda_i)^2} := \overline{V}. \tag{9}$$

We first note that the variance lower bound does not depend on $\rho_k \geq b$ and so it holds for any interpolating linear model, even when benign source conditions are not satisfied. However, we will see that if $\rho_k \geq b$ for some k, then the upper and lower bounds are tight. In the case where $\Sigma_{\mathsf{t}} = \Sigma_{\mathsf{s}}$, these bounds reduce to their in-distribution counterparts [Bar+20]. Our variance bounds show that the excess risk contribution of each feature is scaled by the ratio of the target and source eigenvalues, $\tilde{\lambda}_i / \lambda_i$. We immediately see that scaling down the target eigenvalues will lessen the overall contribution to variance and that scaling up the target eigenvalues will increase the contribution. We investigate these scaling factors and the separation of the first k components and last $p - k$ components in Section 3.1.

We now state upper and lower bounds for the bias term, B_2, given in Equation 4. The proof of the following theorem can be found in Appendix F.

Theorem 3.2. *(Upper and lower bounds for the bias term) For the lower bound only, assume that random models $\bar{\theta}$ are obtained from the underlying θ^*_{s} as $(\bar{\theta})_i = \gamma_i(\theta^*_{\mathsf{s}})_i$, where each γ_i is an independent Rademacher random variable. There exists a universal constant $b > 1$, constants c, C that depend only on b and σ_x, and $k < n/C$ such that if $\rho_k \geq b$, then with probability at least $1 - 10e^{-n/c}$,*

$$\mathop{\mathbb{E}}_{\bar{\theta}}[B_2] \geq \frac{1}{c}\left(\sum_{i=1}^{k} \frac{\tilde{\lambda}_i}{\lambda_i} \frac{\lambda_i (\theta^*_{\mathsf{s}})_i^2}{(1 + \frac{\lambda_i}{\lambda_{k+1}\rho_k})^2} + \sum_{i>k} \tilde{\lambda}_i (\theta^*_{\mathsf{s}})_i^2\right) := \underline{B_2}.$$

If we assume that p is at most exponential in n, then with probability $1 - 5e^{-n/c}$,

$$B_2/c \leq \|\theta^*_{\mathsf{s}}\|^2 \sum_{i=1}^{p} \frac{\tilde{\lambda}_i}{\lambda_i} \frac{\lambda_i}{\left(1 + \frac{\lambda_i}{\lambda_{k+1}\rho_k}\right)} := \overline{B_2}.$$

Note that while the lower bound is in expectation over the random models $\bar{\theta}$, the resulting expression only depends on the ground-truth θ^*_{s}. This Bayesian approach also appears in prior work, i.e. Tsigler and Bartlet [TB23]. In studying the bias lower bound, we observe a similar separation of signal and noise components and depence on eigenvalue ratios as in the variance bounds.

To show tightness of our bounds, we assume there exists a k such that $\rho_k \geq b$ for some universal constant $b > 1$. When this condition is satisfied, the variance bounds are tight up to constant factors. The bias bounds leave a model-dependent and source covariance-dependent gap, which we discuss in the proof overview in Appendix D and in the complete proof found in Appendix G.

Theorem 3.3. *(Tightness of variance and bias bounds) Let the lower bound and upper bound of V be given by $\underline{V}$ and $\overline{V}$, respectively. There exists a universal constant $b \geq 1$, and constant c as defined in Theorem 3.1, and $k \in (0, n/c)$ such that if $\rho_k \geq b$, then*

$$\underline{V}/\overline{V} \in \left[b^{-2}(1+b)^{-2}/c^2, 1\right].$$

Let the lower bound and upper bound of B_2 be given by $\underline{B_2}$ and $\overline{B_2}$, respectively, and the assumptions of Theorem 3.2 be satisfied. Then

$$\underline{B_2}/\overline{B_2} \in \left[\frac{\min_i \left\{(\theta^*_{\mathsf{s}})_i^2 \,:\, (\theta^*_{\mathsf{s}})_i \neq 0\right\}}{\|\theta^*_{\mathsf{s}}\|^2 \left(1 + b^{-1}\frac{\lambda_1}{\lambda_{k+1}}\right)}, 1\right].$$

Note that the gap between our bias upper and lower bounds is independent of the target distribution.

3.1 A Taxonomy of Shifts

We now present a taxonomy of covariate shifts on the target distribution inspired by our prior analysis. We first consider OOD generalization and formally categorize shifts as *beneficial* or *malignant*.

Definition 2 (Beneficial and Malignant shifts)**.** *For a source distribution, $\mathcal{D}_s$, a target distribution, $\mathcal{D}_t$, excess risk, R, and MNI, $\widehat{\theta}$, we say that a shift is*

1. ***beneficial*** *if $R(\widehat{\theta}, \mathcal{D}_s) > R(\widehat{\theta}, \mathcal{D}_t)$,*
2. ***malignant*** *if $R(\widehat{\theta}, \mathcal{D}_s) < R(\widehat{\theta}, \mathcal{D}_t)$.*

We define these shifts for excess risk and note in Appendix J.1 that, empirically, the variance is the primary contributor to excess risk and the bias contributions are negligible when $\Sigma_{\mathtt{s}}$ satisfies benign overfitting conditions. This is in keeping with prior literature that focuses on studying variance in interpolating methods [Bar+20]. We will thus focus on variance in the following discussion.

Prior work shows that if $n, p \to \infty$ at the same rate, $\mathrm{tr}(\Sigma_{\mathtt{s}}) < \mathrm{tr}(\Sigma_{\mathtt{t}})$ results in malignant shifts on excess risk and $\mathrm{tr}(\Sigma_{\mathtt{s}}) > \mathrm{tr}(\Sigma_{\mathtt{t}})$ results in beneficial shifts on excess risk [TAP21]. In this section we generalize these conditions by considering differing rates of $n, p \to \infty$ and measuring overparameterization by the modified "effective rank" measure R_k/n rather than p. This leads us to a novel characterization of the role of overparameterization in covariate shifts. For completeness, we describe their trace conditions in terms of our shifts in Appendix H.1.

We first consider separate multiplicative shifts on the signal components and noise components. Let $\Sigma_{\mathtt{s}}$ be a covariance matrix that satisfes benign source conditions. Define $\Sigma_{\mathtt{t}}$ by, $\tilde{\lambda}_i = \alpha\lambda_i$ for $i \leq k$, and $\tilde{\lambda}_i = \beta\lambda_i$ for $i > k$ with $\alpha, \beta \geq 0$. While these are simple multiplicative shifts, they are instructive with regard to understanding the dynamics of overparameterization and covariate shift. In Appendix H.3 we generalize this analysis to allow for arbitrary multiplicative shifts in every direction.

By Theorem 3.1, up to constants,

$$V_{ood} \approx \alpha\frac{k}{n} + \beta\frac{n}{R_k} \tag{10}$$

where $R_k = (\sum_{i>k}\lambda_i)^2/(\sum_{i>k}\lambda_i^2)$.

It is clear that if $V_{ood} - V_{id} > 0$ then we have a malignant shift on the variance, and if $V_{ood} - V_{id} < 0$ then we have a beneficial shift on the variance. Observe that,

$$V_{ood} - V_{id} \approx (\alpha - 1)\frac{k}{n} + (\beta - 1)\frac{n}{R_k}. \tag{11}$$

In this expression, we see that the scales of signal and noise shifts, α and β, are important, as is the relationship between k/n (the "classical" rate) and n/R_k (the "high-dimensional" rate). The quantity n/R_k can be interpreted as an inverse measure of overparameterization, where smaller values correspond to higher levels of overparameterization. The rate of overparameterization relative to the classical rate of k/n determines whether the shift on the first k components (α) or the shift on the last $p - k$ components (β) contributes more to the difference in excess risk.

Based on this intuition, we define two regimes of overparameterization: *mild* and *severe*.

Definition 3 (Mild and severe overparameterization for multiplicative shifts)**.** *Let $\Sigma_{\mathtt{s}}$ be a source covariance that satisfies benign source conditions and let $k \leq n$. Define $\Sigma_{\mathtt{t}}$ as, $\tilde{\lambda}_i = \alpha\lambda_i$ for $i \leq k$ and $\tilde{\lambda}_i = \beta\lambda_i$ for $i > k$, with $\alpha, \beta \geq 0$. Let $C_{\alpha\beta} := \left|\frac{\alpha-1}{1-\beta}\right|$.*

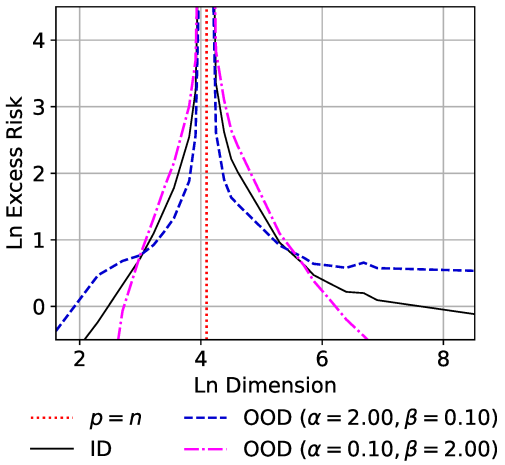


Figure 1: We experiment with the (k, δ, ϵ) spiked covariance models and examine conditions for beneficial and malignant shifts as given in Theorem 3.4. We take $n = 60, k = 10, \delta = 1.0, \epsilon = 1e^{-6}, \tilde{\delta} = 2.0, \tilde{\epsilon} = 1e^{-7}$, and vary p. We see a cross-over from mild to severe overparameterization on the right side of $p = n$ where both OOD shifts swap between beneficial and malignant. For both ID and OOD curves, we observe that excess risk is a decreasing function if input dimension. Curves are averaged over 100 independent runs.

*We are in the **mildly overparameterized** regime if*

$$\frac{n}{R_k} = \omega\left(C_{\alpha\beta} \cdot \frac{k}{n}\right). \tag{12}$$

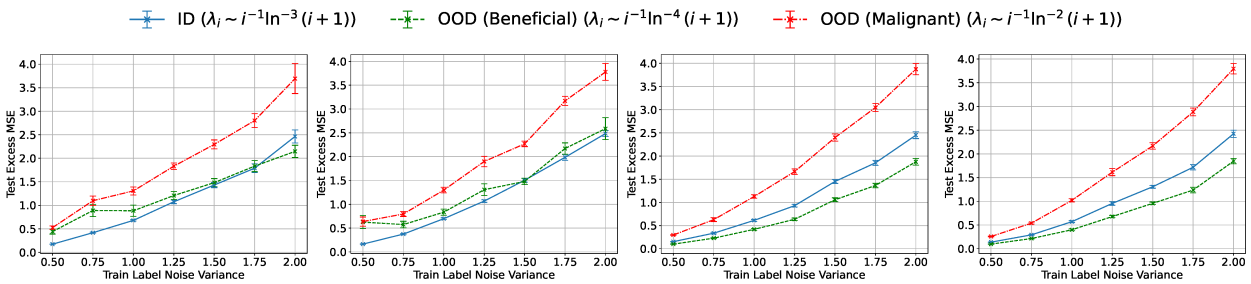
*We are in the **severely overparameterized** regime if*

$$\frac{n}{R_k} = o\left(C_{\alpha\beta} \cdot \frac{k}{n}\right). \tag{13}$$

For $\beta = 1$ we define $C_{\alpha\beta} = \infty$ and thus are effectively in the severely overparameterized regime with regard to the types of shift we observe.

It is clear that k is important in defining regimes of overparameterization. The aforementioned definitions hold for any $k < n$, however we derive our taxonomy of shifts in the case in which $\exists\ k < n$ such that $\rho_k \geq b$ for a universal constant $b > 1$. We note that even for non-linear models or settings that do not exhibit benign overfitting we can still think about a notion of a "k" akin to the intrinsic dimension of the data. We empirically show in Figures 6 and 7 that our taxonomy of shifts is reflective of shift behavior in realistic settings by heuristically taking k small enough to sufficiently capture the low-dimensional signal in the data.

In Equation 11, we see that the limit of the severe overparameterization regime would take $R_k \to \infty$ first, while holding other problem parameters fixed. In this case, we are only left with α shifts on the top k components, as any shift on the bottom components is suppressed by the high rank covariance tail. This leads to behaviors akin to classical intuitions for an underparameterized linear regression estimator where $k = p < n$. In this case, $\alpha > 1$ leads to more variance and thus harder learning, whereas $\alpha < 1$ leads



(a) $n = 200, p = 20, h = 512$ (b) $n = 200, p = 20, h = 2048$ (c) $n = 200, p = 2k, h = 512$ (d) $n = 200, p = 2k, h = 2048$

Figure 2: We train 3 layer ReLU dense neural networks with hidden width, h, on n samples from p-dimensional Gaussians. ID test data is sampled from the same distribution and OOD test sets are constructed based on beneficial and malignant covariate shifts in our theory. Ground truth models are sampled as $\theta_{\mathsf{s}}^* \sim \mathcal{S}^{p-1}$, no model shift is invoked. For training data, X, train labels are given by $\boldsymbol{y}_{\mathsf{s}} = X\theta_{\mathsf{s}}^* + \boldsymbol{\varepsilon}_{\mathsf{s}}$ with label noise $\boldsymbol{\varepsilon}_{\mathsf{s}} \sim \mathcal{N}(0, \sigma^2)$. All runs reach train loss $< 5e^{-6}$. Points are averaged over 20 independent runs with standard error bars reported.

to less variance and thus easier learning. These notions of hard vs. easy learning naturally correspond to $\mathrm{tr}(\Sigma_{\mathsf{t}}) > \mathrm{tr}(\Sigma_{\mathsf{s}})$ and $\mathrm{tr}(\Sigma_{\mathsf{t}}) < \mathrm{tr}(\Sigma_{\mathsf{s}})$, respectively. This is shown experimentally in Figs. 1 and 7 by looking at the left and right sides of the figures.

On the other hand, in the mildly overparameterized regime covariance tail shifts are not sufficiently suppressed and lead to non-negligible interactions with shifts on the signal components. An increase in energy in the signal components can be counteracted by a decrease in energy in the noise components, effectively increasing the contrast between signal and noise in favor of the signal. Similarly, a decrease in energy in the signal components can be harmfully counteracted by an increase in energy in the noise components. This is visible in Figs. 1 and 7 just above the threshold of interpolation. Interestingly, in the mildly overparameterized regime we can also define settings in which $\mathrm{tr}(\Sigma_{\mathsf{t}}) > \mathrm{tr}(\Sigma_{\mathsf{s}})$ and yet still obtain a beneficial shift, and settings in which $\mathrm{tr}(\Sigma_{\mathsf{t}}) < \mathrm{tr}(\Sigma_{\mathsf{s}})$ and yet still obtain malignant shifts.

We formalize these observations in the following theorem, the proof of which can be found in Appendix H.2.

Theorem 3.4. *(Beneficial and Malignant Multiplicative Shifts on Variance) Let Σ_{s} be a source covariance that satisfies benign source conditions. That is, $\exists\, k$ such that $\rho_k \geq b$ for a universal constant $b > 1$. Define Σ_{t} as $\tilde{\lambda}_i = \alpha\lambda_i$ for $i \leq k$ and $\tilde{\lambda}_i = \beta\lambda_i$ for $i > k$, with $\alpha, \beta \geq 0$.*

1. *If $\alpha < 1, \beta \leq 1$ or $\alpha \leq 1, \beta < 1$ then we obtain a beneficial shift in variance.*
2. *If $\alpha > 1, \beta \geq 1$ or $\alpha \geq 1, \beta > 1$ then we obtain a malignant shift in variance.*
3. *If we are in the mildly overparameterized regime:*
   - *$\alpha > 1$ and $\beta < 1$ leads to beneficial shifts;*
   - *$\alpha < 1$ and $\beta > 1$ leads to malignant shifts.*
4. *If we are in the severely overparameterized regime:*

- $\alpha > 1$ *and* $\beta < 1$ *leads to malignant shifts;*
- $\alpha < 1$ *and* $\beta > 1$ *leads to beneficial shifts.*

Figs. 1 and 5 demonstrate the relationship between the n/R_k and k/n rates in the case of $C_{\alpha\beta} = 1.11, C_{\alpha\beta} = 1$, respectively, for spiked covariance models. In both, we clearly see a cross-over from beneficial to malignant shifts when we transition from mild to severely overparameterized.

Overparameterization improves OOD robustness Focusing on just the target excess risk, let $\alpha = \alpha(n)$ and $\beta = \beta(n, p)$. We say that the benignly-overfit MNI is robust if its excess risk decays to zero despite the presence of multiplicative covariate shifts. In order for the variance upper bound to decay to 0, it is sufficient to have the shifts in the signal and noise components satisfy, $\alpha = o(n/k),\ \beta = o(R_k/n)$. The condition $\beta = o(R_k/n)$ allows the shift strength to increase at a rate determined by the level of overparameterization, so we conclude that increasing the amount of overparameterization improves robustness to multiplicative distribution shifts. Note that α has no dependence on R_k and so robustness to shifts on the signal components is independent of the degree of overparameterization.

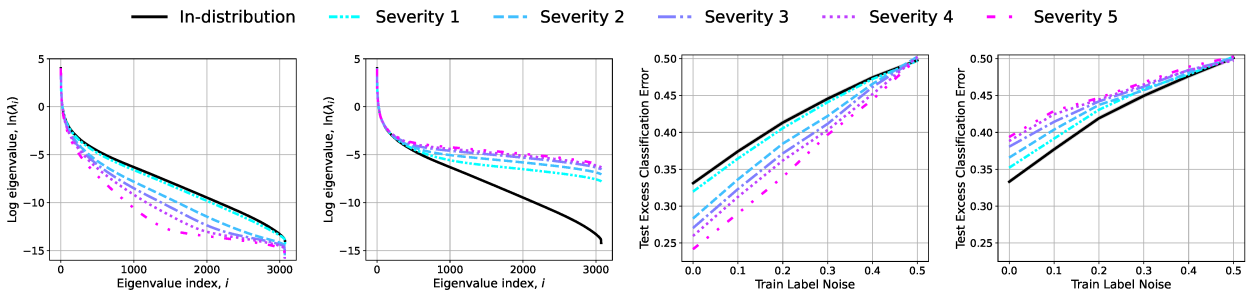
4 Experiments

Our theoretical results have provided insight into distribution shifts in high-dimensional linear regression. We now present experiments with linear models and neural networks, relaxing many of the assumptions used for theoretical results. Specifically, we empirically: (1) observe beneficial and malignant shifts on synthetic and real data for linear models (benignly overfit and otherwise) and even high-dimensional dense neural networks; (2) show the benefit of overparameterization in covariate shift for interpolating linear estimators; (3) validate that our findings hold when the source and target covariance matrices are not simultaneously diagonalizable, as well as under model misspecification; (4) provide experimental insight that high-dimensional neural networks, i.e. when the input data dimension is large relative to the training sample size, act similarly to the MNI whereas low-dimensional neural networks do not, regardless of the level of overparameterization. Details of experimental setup, data, and models are given in Appendix I. We now discuss key observations and takeaways from the experiments.

4.1 Synthetic Data Experiments

Fig. 1 shows excess risk vs. input dimension for data sampled from the (k, δ, ϵ)-spike covariance model with $k = 10$, $\delta = 1.0$, and $\epsilon = 1e^{-6}$. Beneficial and malignant shifts are seen in the setting of Theorem 3.4 with $\alpha = 2.0, \beta = 0.1$. That is, we see *two* cross-over points: one in the underparameterized regime and one in the overparameterized regime (going from mild to severe). This suggests that non-negligible covariance tail effects are a property of shifting when a model is in a region around the double descent peak. The further we are from the double descent peak, the more "classical" our behavior is, in that the top k components are the only ones that influence shift and the bottom $p - k$ components either don't exist or have negligible effects. Appendix J.1 explores this setup for different values of α, β and interpolating linear models for eigendecay rates that lead to harmful interpolation, i.e. *tempered* or *catastrophic* overfitting [Mal+22].

Figs. 2 and 8 show similar results for 3-layer dense ReLU neural networks trained until near-interpolation (train MSE $< 5e^{-6}$) on synthetic data with benign overfitting eigendecay rates [Bar+20]. For the neural network, we consider p to be the dimension of the input data, rather than the number of network parameters. From Fig. 2 we observe similar trends predicted by our theory for beneficial and malignant shifts when $p > n$, indicating that while our theory is developed for linear models we are able to extrapolate to more



(a) Blur Covariance (b) Noise Covariance (c) MNI tested on Blur (d) MNI tested on Noise

Figure 3: We fit the MNI to binary CIFAR-10 (dog vs. truck) and test on binary CIFAR-10C under Gaussian blur and noise corruptions. In (a), (b) we plot the eigenvalues of the covariance matrices for ID test data and on test sets for each severity. To ensure $p > n$ we subsample the training set to $n = 1k$ and average curves over 50 independent runs. We evaluate the MNI against all 5 corruption severities and plot excess classification error vs training label noise, which is class label flip probability. We see that the eigenspectra of the OOD datasets is directly correlated to the OOD performance of the MNI.

complex models. In both the $p > n$ and $p < n$ experiments, our results are agnostic to the hidden width of the network, further suggesting that overparameterization is qualitatively different from high-dimensionality. When $p > n$, a neural network appears to act like the interpolating MNI under distribution shifts. For $p < n$ the interpolating dense net does not exhibit the properties of an interpolating MNI under distribution shift and the ID excess error is better than both "beneficial" and "malignant" OOD excess errors. However, the relative difference between beneficial and malignant shifts is still preserved. Note that we observe the exact same behavior in Fig. 10 when training ResNets to interpolation on CIFAR-10 and testing on CIFAR-10C blur and noise corruptions [HD19].

4.2 CIFAR-10 Experiments

Next, we consider experiments with linear interpolators on a binarized CIFAR-10 and CIFAR-10C with Gaussian noise and blur corruptions at varying levels of corruption severity. For details, see Appendix I. This experiment breaks the assumption of simultaneous diagonalizability, and the well-specified assumption as the labels for CIFAR are not obtained by a ground-truth linear model.

Fig. 3 shows empirical results on the MNI fit to this problem. We plot the eigenspectra of the blur and noise covariances from CIFAR-10C compared to the eigenspectra of CIFAR-10 in Figs. 3a, 3b. We identified these two shifts due to their eigenspectra reflecting what we expect to lead to beneficial and malignant shifts based on Theorem 3.4. We observe a tight relationship between changes in the eigenspectra of the target data and excess classification error when evaluated by the MNI. We notice that blurs reduce covariance energy with increased blurring, like a "denoising"-style operation. Experimentally this leads to improved OOD accuracy. In contrast, noise corruptions add energy to the covariance tail and lead to worsened OOD accuracy. Fig. 9 also shows that further overparameterization in this setting leads to improved behavior of the MNI on both corruptions.

5 Conclusion and future work

Our work provides the first finite-sample, instance-wise analysis of the MNI under transfer learning with high-dimensional linear models. We show a taxonomy of beneficial and malignant covariate shifts depending on whether we are in a *mild* or *severely* overparameterized regime. In the mildly overparameterized regime, variance contributions on the top k components interact with that of the bottom $p-k$ components in non-negligible ways, leading to non-standard shifts. In the severely overparameterized regime, the high-rank covariance tail suppresses variance contributions in the bottom $p-k$ components and so OOD generalization acts more "classical", akin to underparameterized linear regression where $k = p < n$.

Benign overfitting literature commonly claims to be motivated by "overparameterized" neural networks, referring to the number of parameters in the network rather than the dimension of the data. However recent works have challenged this, suggesting the role of the ambient dimension and source covariance is more important than parameter count in determining whether overfitting is benign or catastrophic in neural networks [FCB23; KYS23]. Prior work has also shown that gradient descent on 2-layer neural networks has an implicit bias towards linear decision boundaries when $p \gg n$, independent of the degree of overparameterization [Fre+22].

Our experiments further support the view that high-dimensional neural networks behave similarly to high-dimensional linear models, whereas low-dimensional neural networks do not. They provide a new and important perspective on the difference between high-dimensionality and overparameterization in neural networks in the case of distribution shift, which has yet to be appreciated in the literature. While dimensionality and degree of overparameterization are inextricably linked in linear regression, practical deep learning tends to operate in the overparameterized setting, not the high-dimensional one.

An important future direction is to investigate the extent to which our results hold for distribution shifts on more complex high-dimensional datasets. It is also of interest to extend our finite-sample theoretical analysis to shallow ReLU neural networks, other nonlinear models, and learning algorithms that overfit in a *tempered* manner [Mal+22]. Finally, future work might seek to extend our understanding of neural networks by carefully studying the interplay between the data dimension, number of network parameters, number of training samples, and the optimization algorithm and loss function, and how this interplay can affect ID & OOD generalization.

Acknowledgements

The authors thank Alexander Tsigler, Peter Bartlett, Emmanuel Abbe and Libin Zhu for useful discussion and feedback on the manuscript.

We gratefully acknowledge partial support from NSF grants DMS-2209975, 2015341, NSF grant 2023505 on Collaborative Research: Foundations of Data Science Institute (FODSI), the NSF and the Simons Foundation for the Collaboration on the Theoretical Foundations of Deep Learning through awards DMS-2031883 and 814639, and NSF grant MC2378 to the Institute for Artificial CyberThreat Intelligence and OperatioN (ACTION). NM gratefully acknowledges funding and support for this research from the Eric and Wendy Schmidt Center at The Broad Institute of MIT & Harvard. AZ additionally acknowledges support from NSF RTG Grant #1745640.

This work used Delta GPU compute nodes at NCSA and HPE and Expanse GPU compute nodes at Dell

and SDSC through allocation CIS220009 from the Advanced Cyberinfrastructure Coordination Ecosystem: Services & Support (ACCESS) program, which is supported by National Science Foundation grants #2138259, #2138286, #2138307, #2137603, and #2138296 [Boe+23].

References

[Bar+20] Peter L. Bartlett, Philip M. Long, Gábor Lugosi, and Alexander Tsigler. "Benign Overfitting in Linear Regression". In: *Proceedings of the National Academy of Sciences* 117.48 (2020), pp. 30063–30070 (Cited on pages 1–6, 8, 11, 17–19, 21–23, 25, 26, 28, 29, 45, 48).

[BS23] Daniel Barzilai and Ohad Shamir. "Generalization in Kernel Regression Under Realistic Assumptions". In: *Preprint, arXiv:2312.15995* (2023) (Cited on page 2).

[Boe+23] Timothy J. Boerner, Stephen Deems, Thomas R. Furlani, Shelley L. Knuth, and John Towns. "ACCESS: Advancing Innovation: NSF's Advanced Cyberinfrastructure Coordination Ecosystem: Services & Support". In: *In Practice and Experience in Advanced Research Computing (PEARC '23)*. Portland, OR, USA. ACM, New York, NY, USA, July 2023. DOI: 10.1145/3569951.3597559 (Cited on page 14).

[CL23] Niladri S. Chatterji and Philip M. Long. "Deep Linear Networks can Benignly Overfit when Shallow Ones Do". In: *Journal of Machine Learning Research* 24.117 (2023), pp. 1–39 (Cited on pages 2, 3, 6).

[CLB22] Niladri S. Chatterji, Philip M. Long, and Peter L. Bartlett. "The Interplay Between Implicit Bias and Benign Overfitting in Two-Layer Linear Networks". In: *Journal of Machine Learning Research* 23.263 (2022), pp. 1–48 (Cited on pages 2, 6).

[DAm+22] Alexander D'Amour et al. "Underspecification Presents Challenges for Credibility in Modern Machine Learning". In: *Journal of Machine Learning Research* 23.226 (2022), pp. 1–61 (Cited on page 1).

[Dwi+20] Raaz Dwivedi, Chandan Singh, Bin Yu, and Martin J. Wainwright. "Revisiting minimum description length complexity in overparameterized models". In: *Preprint, arXiv:2006.10189* (2020) (Cited on page 1).

[Fen+23] Xingdong Feng, Xin He, Caixing Wang, Chao Wang, and Jingnan Zhang. "Towards a Unified Analysis of Kernel-based Methods Under Covariate Shift". In: *Conference on Neural Information Processing Systems (NeurIPS)*. 2023 (Cited on page 3).

[FCB23] Spencer Frei, Niladri S. Chatterji, and Peter L. Bartlett. "Random Feature Amplification: Feature Learning and Generalization in Neural Networks". In: *Journal of Machine Learning Research* 24.303 (2023), pp. 1–49 (Cited on page 13).

[FCB22] Spencer Frei, Niladri S Chatterji, and Peter Bartlett. "Benign Overfitting without Linearity: Neural Network Classifiers Trained by Gradient Descent for Noisy Linear Data". In: *Conference on Learning Theory (COLT)*. 2022 (Cited on pages 2, 3).

[Fre+23] Spencer Frei, Gal Vardi, Peter L. Bartlett, and Nathan Srebro. "Benign Overfitting in Linear Classifiers and Leaky ReLU Networks from KKT Conditions for Margin Maximization". In: *Conference on Learning Theory (COLT)*. 2023 (Cited on page 3).

[Fre+22] Spencer Frei, Gal Vardi, Peter L. Bartlett, Nathan Srebro, and Wei Hu. "Implicit Bias in Leaky ReLU Networks Trained on High-Dimensional Data". In: *International Conference on Learning Representations (ICLR)*. 2022 (Cited on page 13).

[Haa+23] Moritz Haas, David Holzmüller, Ulrike von Luxburg, and Ingo Steinwart. "Mind the spikes: Benign overfitting of kernels and neural networks in fixed dimension". In: *Conference on Neural Information Processing Systems (NeurIPS)*. 2023 (Cited on page 2).

[Has+22] Trevor Hastie, Andrea Montanari, Saharon Rosset, and Ryan J. Tibshirani. “Surprises in High-Dimensional Ridgeless Least Squares Interpolation”. In: *Annals of Statistics* 50.2 (2022), pp. 949–986 (Cited on page 1).

[Hen+21] Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, Dawn Song, Jacob Steinhardt, and Justin Gilmer. “The Many Faces of Robustness: A Critical Analysis of Out-of-Distribution Generalization”. In: *International Conference on Computer Vision (ICCV)*. 2021 (Cited on page 1).

[HD19] Dan Hendrycks and Thomas G. Dietterich. “Benchmarking neural network robustness to common corruptions and perturbations”. In: *International Conference on Learning Representations (ICLR)*. 2019 (Cited on pages 3, 12, 46).

[Koh+21] Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard L. Phillips, Ian Gao, et al. “Wilds: A benchmark of in-the-wild distribution shifts”. In: *International Conference on Machine Learning (ICML)*. 2021 (Cited on pages 1, 3).

[KYS23] Guy Kornowski, Gilad Yehudai, and Ohad Shamir. “From Tempered to Benign Overfitting in ReLU Neural Networks”. In: *Conference on Neural Information Processing Systems (NeurIPS)*. 2023 (Cited on pages 2, 13).

[Kou+23] Yiwen Kou, Zixiang Chen, Yuanzhou Chen, and Quanquan Gu. “Benign Overfitting in Two-layer ReLU Convolutional Neural Networks”. In: *International Conference on Machine Learning (ICML)*. 2023 (Cited on pages 2, 3).

[Lai+23] Jianfa Lai, Zixiong Yu, Songtao Tian, and Qian Lin. “Generalization Ability of Wide Residual Networks”. In: *Preprint, arXiv:2305.18506* (2023) (Cited on page 2).

[LHL21] Qi Lei, Wei Hu, and Jason D. Lee. “Near-Optimal Linear Regression under Distribution Shift”. In: *International Conference on Machine Learning (ICML)*. 2021 (Cited on pages 2, 3).

[Lia+23] Weixin Liang, Yining Mao, Yongchan Kwon, Xinyu Yang, and James Y. Zou. “Accuracy on the Curve: On the Nonlinear Correlation of ML Performance Between Data Subpopulations”. In: *International Conference on Machine Learning (ICML)*. 2023 (Cited on page 1).

[MPW23] Cong Ma, Reese Pathak, and Martin J. Wainwright. “Optimally tackling covariate shift in RKHS-based nonparametric regression”. In: *Annals of Statistics* 51.2 (2023), pp. 738–761 (Cited on page 3).

[Mal+22] Neil Mallinar, James Simon, Amirhesam Abedsoltan, Parthe Pandit, Mikhail Belkin, and Preetum Nakkiran. “Benign, Tempered, or Catastrophic: Toward a Refined Taxonomy of Overfitting”. In: *Conference on Neural Information Processing Systems (NeurIPS)*. 2022 (Cited on pages 2, 3, 11, 13, 48).

[Mil+21] John P Miller, Rohan Taori, Aditi Raghunathan, Shiori Sagawa, Pang Wei Koh, Vaishaal Shankar, Percy Liang, Yair Carmon, and Ludwig Schmidt. “Accuracy on the Line: on the Strong Correlation Between Out-of-Distribution and In-Distribution Generalization”. In: *International Conference on Machine Learning (ICML)*. 2021 (Cited on pages 1, 3).

[Mut+20] Vidya Muthukumar, Kailas Vodrahalli, Vignesh Subramanian, and Anant Sahai. “Harmless interpolation of noisy data in regression”. In: *IEEE Journal on Selected Areas in Information Theory* 1.1 (2020), pp. 67–83 (Cited on page 2).

[Ogl+22] Dino Oglic, Zoran Cvetkovic, Peter Sollich, Steve Renals, and Bin Yu. “Towards Robust Waveform-Based Acoustic Models”. In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 30 (2022), pp. 1977–1992 (Cited on page 1).

[PMW22] Reese Pathak, Cong Ma, and Martin Wainwright. "A new similarity measure for covariate shift with applications to nonparametric regression". In: *International Conference on Machine Learning (ICML)*. 2022 (Cited on pages 2, 3).

[RZ19] Alexander Rakhlin and Xiyu Zhai. "Consistency of Interpolation with Laplace Kernels is a High-Dimensional Phenomenon". In: *Conference on Learning Theory (COLT)*. 2019 (Cited on page 2).

[Rec+19] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. "Do ImageNet Classifiers Generalize to ImageNet?" In: *International Conference on Machine Learning (ICML)*. 2019 (Cited on page 1).

[Sim+23] Max Simchowitz, Anurag Ajay, Pulkit Agrawal, and Akshay Krishnamurthy. "Statistical Learning under Heterogeneous Distribution Shift". In: *International Conference on Machine Learning (ICML)*. 2023 (Cited on page 3).

[TAP21] Nilesh Tripuraneni, Ben Adlam, and Jeffrey Pennington. "Overparameterization Improves Robustness to Covariate Shift in High Dimensions". In: *Conference on Neural Information Processing Systems (NeurIPS)*. 2021 (Cited on pages 2, 8).

[TB23] Alexander Tsigler and Peter L. Bartlet. "Benign overfitting in ridge regression". In: *Journal of Machine Learning Research* 24.123 (2023), pp. 1–76 (Cited on pages 1–3, 5–7, 17, 19, 25, 28, 35).

[Ver18] Roman Vershynin. *High-dimensional probability*. Cambridge University Press, 2018 (Cited on page 4).

[Wan23] Kaizheng Wang. "Pseudo-Labeling for Kernel Ridge Regression under Covariate Shift". In: *Preprint, arXiv:2302.10160* (2023) (Cited on page 2).

[Wan+22] Ke Alexander Wang, Niladri S Chatterji, Saminul Haque, and Tatsunori Hashimoto. "Is Importance Weighting Incompatible with Interpolating Classifiers?" In: *International Conference on Learning Representations (ICLR)*. 2022 (Cited on pages 2, 3).

[Wen+22] F. Wenzel, Andrea Dittadi, Peter Gehler, Carl-Johann Simon-Gabriel, Max Horn, Dominik Zietlow, David Kernert, Chris Russell, Thomas Brox, Bernt Schiele, Bernhard Scholkopf, and Francesco Locatello. "Assaying Out-Of-Distribution Generalization in Transfer Learning". In: *Conference on Neural Information Processing Systems (NeurIPS)*. 2022 (Cited on page 1).

[Xu+24] Zhiwei Xu, Yutong Wang, Spencer Frei, Gal Vardi, and Wei Hu. "Benign Overfitting and Grokking in ReLU Networks for XOR Cluster Data". In: *International Conference on Learning Representations (ICLR)*. 2024 (Cited on page 2).

[Zha+17] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. "Understanding deep learning requires rethinking generalization". In: *International Conference on Learning Representations (ICLR)*. 2017 (Cited on page 2).

A Formal assumptions

Definition 4 (Linear regression under distribution shift)**.** *We consider a training dataset comprised of n i.i.d. pairs $(\boldsymbol{x}^i, y^i)_{i=1}^n \sim \mathcal{D}_{\mathsf{s}}^n$ concatenated into a data matrix $X \in \mathbb{R}^{n \times p}$ and a response vector $\boldsymbol{y}_{\mathsf{s}} \in \mathbb{R}^n$. In the overparameterized setting, $p > n$ meaning we have more input features than training samples.*

We define

1. *the covariance matrix $\Sigma_{\mathsf{s}} = \mathbb{E}_{\mathcal{D}_{\mathsf{s}}}[\boldsymbol{x}\boldsymbol{x}^\top]$,*
2. *the optimal parameter vector $\theta^*_{src} \in \mathbb{R}^p$, satisfying*

$$\mathop{\mathbb{E}}_{\mathcal{D}_{\mathsf{s}}}\left[(y - x^\top \theta^*_{src})^2\right] = min_\theta \mathop{\mathbb{E}}_{\mathcal{D}_{\mathsf{s}}}\left[(y - x^\top \theta)^2\right].$$

We test on the distribution $\mathcal{D}_{\mathsf{t}}$ with Σ_{t} and θ^*_{t} defined in the same way. We assume

1. *(centered rows) $\mathbb{E}_{\mathcal{D}_{\mathsf{s}}}[\boldsymbol{x}] = 0$;*
2. *(well-specified - source) For $(X, \boldsymbol{y}) \subseteq \mathcal{D}_{\mathsf{s}}$, $\boldsymbol{y} = X\theta^*_{\mathsf{s}} + \boldsymbol{\varepsilon}_{\mathsf{s}}$. We assume that the components of the source noise vector $\boldsymbol{\varepsilon}_{\mathsf{s}}$ are i.i.d. centered random variables with positive variance $v_{\varepsilon_s^2}$ and that $\mathbb{E}_{\mathcal{D}_{\mathsf{s}}}[y|x] = x^T\theta^*_{\mathsf{s}}$;*
3. *(well-specified - target) For $(X, \boldsymbol{y}) \subseteq \mathcal{D}_{\mathsf{t}}$, $\boldsymbol{y} = X\theta^*_{\mathsf{t}} + \boldsymbol{\varepsilon}_{\mathsf{t}}$. We assume that the components of the target noise vector $\boldsymbol{\varepsilon}_{\mathsf{t}}$ are i.i.d. centered random variables with noise variance, $v_{\varepsilon_t^2}$, and that $\mathbb{E}_{\mathcal{D}_{\mathsf{t}}}[y|x] = x^T\theta^*_{\mathsf{t}}$;*
4. *(simultaneously diagonalizability) Σ_{s} and Σ_{t} commute; that is, there exists an orthogonal matrix $V \in \mathbb{R}^p$ such that $V^\top \Sigma_{\mathsf{s}} V$ and $V^\top \Sigma_{\mathsf{t}} V$ are both diagonal. This allows us to fix an orthonormal basis in which we can express the covariance matrices as*

$$\Sigma_{\mathsf{s}} = \mathop{\mathbb{E}}_{x \sim \mathcal{D}_{\mathsf{s}}}[xx^T] = diag(\lambda_1, \lambda_2, ..., \lambda_p),$$
$$\Sigma_{\mathsf{t}} = \mathop{\mathbb{E}}_{x \sim \mathcal{D}_{\mathsf{t}}}[xx^T] = diag(\tilde{\lambda}_1, \tilde{\lambda}_2, ..., \tilde{\lambda}_p),$$

   where the source eigenvalues are a non-increasing sequence, $\lambda_1 \geq \lambda_2 \cdots \geq \lambda_p$. Note that we do not require the target eigenvalues to be a non-increasing sequence, however we require that $\tilde{\lambda}_i \lambda_i \geq 0$ for all i;
5. *(subgaussianity) the whitened data matrix, denoted $Z = X\Sigma_{\mathsf{s}}^{-1/2}$, has centered i.i.d. row vectors with independent coordinates. We assume that the rows are subgaussian with subgaussian norm σ_x; that is, for all $\gamma \in \mathbb{R}^p$,*

$$\mathbb{E}[\exp(\gamma^\top z)] \leq \exp(\sigma_x^2 ||\gamma||^2/2).$$

B Key results from prior work and technical lemmas

For ease on the reader, we replicate some key lemma statements from Bartlett et al. [Bar+20] and Tsigler and Bartlet [TB23] and provide new lemmas and corollaries that we use in our work.

Recall that $\rho_k = \frac{1}{n\lambda_{k+1}} \sum_{i>k} \lambda_i$, $X \in \mathbb{R}^{n\times p}$, and $\Sigma_{\mathsf{s}} \in \mathbb{R}^{p\times p} = \text{diag}(\lambda_1, \cdots, \lambda_p)$. Let $X_{0:k} \in \mathbb{R}^{n\times k}$ denote the matrix comprised of the first k feature columns. Similarly, $X_{k:p} \in \mathbb{R}^{n\times(p-k)}$ denote the matrix of the last $p-k$ feature columns. The Gram matrix of the data, denoted here by

$$A = XX^T,$$

plays a central role in the investigation of high-dimensional linear regression. Analogous to the above, we express $A_{0:k} = X_{0:k}X_{0:k}^T \in \mathbb{R}^{n\times n}$ and similarly for $A_{k:p} \in \mathbb{R}^{n\times n}$.

Letting $Z = X\Sigma_{\mathsf{s}}^{-1/2} \in \mathbb{R}^{n\times p}$ and denoting the independent column vectors of Z by $z_i \in \mathbb{R}^n$, we have the following expressions:

$$A = \sum_i \lambda_i z_i z_i^T, \qquad A_{-i} = \sum_{j\neq i} \lambda_j z_j z_j^T, \qquad A_k = \sum_{i>k} \lambda_i z_i z_i^T.$$

The following lemma from Bartlett et al. [Bar+20] is key in controlling the largest and smallest eigenvalues of the data Gram matrix and its variants A_{-i} and A_k. Importantly, it also shows that if the energy in the bottom $p-k$ components of the covariance matrix is sufficiently large (ρ_k is lower bounded by a constant), then the largest and smallest eigenvalues of A_k are equal up to constants.

Lemma B.1 (Lemma 5 from Bartlett et al. [Bar+20])**.** *There are constants $b, c \geq 1$ such that for any $k \geq 0$, with probability at least $1 - 2e^{-n/c}$,*

1. *for all $i \geq 1$,*
$$\mu_{k+1}(A_{-i}) \leq \mu_{k+1}(A) \leq \mu_1(A_k) \leq c\left(\sum_{j>k} \lambda_j + \lambda_{k+1}n\right),$$
2. *for all $1 \leq i \leq k$,*
$$\mu_n(A) \geq \mu_n(A_{-i}) \geq \mu_n(A_k) \geq \frac{1}{c}\sum_{j>k} \lambda_j - c\lambda_{k+1}n,$$
3. *if $\rho_k \geq b$, then*
$$\frac{1}{c}\lambda_{k+1}\rho_k n \leq \mu_n(A_k) \leq \mu_1(A_k) \leq c\lambda_{k+1}\rho_k n.$$

A consequence of the prior eigenvalue bounds is that when ρ_k is lower bounded by a constant, the condition number of A_k is upper bounded by a constant. Therefore even as problem parameters such as training sample size and input dimension grow to ∞, A_k is still well-conditioned. This is important as non-benign overfitting occurs when the condition number bound on A_k grows with problem parameters. This would happen if the lower bound on the smallest eigenvalue of A_k decays to zero too quickly which would cause the condition number of A_k to diverge. If this occurs then the excess risk of the MNI would be lower bounded. This is shown for the in-distribution case in Bartlett et al. [Bar+20].

Corollary B.2. *Following from Lemma B.1, there are constants $b, c \geq 1$ such that for any $k \geq 0$, with probability at least $1 - 2e^{-n/c}$, if $\rho_k \geq b$ then*

$$\frac{\mu_1(A_k)}{\mu_n(A_k)} \leq c^2 \tag{14}$$

which is the equivalent of the assumption CondNum$(k, 2e^{-n/c}, c^2)$ as defined in Tsigler and Bartlet [TB23].

The following definition and lemma omit all references to *NonCritReg* and the ridge parameter in Tsigler and Bartlet [TB23].

Definition 5 (StableLowerEig(k, δ, L) from Tsigler and Bartlet [TB23])**.** *Assume that for any $j \in \{1, 2, \cdots, p\}$ with probability (separate for every j) at least $1 - \delta$,*

$$\mu_n(A_{-j}) \geq \mu_n(\mathbb{E}\, A_k)/L = (\sum_{i>k} \lambda_i)/L. \tag{15}$$

We now state key assumptions that are necessary in order to obtain an explicit bias lower bound. Exchangeable coordinates (*ExchCoord*) is a weaker assumption than independent components of the data vector. It is used in Tsigler and Bartlet [TB23] instead of independent components. We assume that components of Z are independent and so we immediately satisfy the *ExchCoord*, which we define here.

Definition 6 (ExchCoord)**.** *Assume the sequence of coordinates of $\Sigma_{\mathsf{s}}^{-1/2}x$, for any $x \in X$, is exchangable (any deterministic permutation of the coordinates of whitened data vectors doesn't change their distribution).*

The *PriorSigns* assumption is necessary to obtain lower bounds on the bias term. It allows us to use bounds on the expectation of a quadratic form, $\mathbb{E}_v[v^T M v]$, in order to separately analyze the contributions of v and M. As the bias takes the form $\theta_{\mathsf{s}}^{*T}(I - X^T A^{-1} X)\Sigma_{\mathsf{t}}(I - X^T A^{-1} X)\theta_{\mathsf{s}}^*$ we see that such a bound would separate the contributions of the model from that of data-dependent matrix expressions.

Definition 7 (PriorSigns)**.** *Assume that θ^* is sampled from a prior distribution in the following way: one starts with vector $\overline{\theta}$ and flips signs of all its coordinates with probability 0.5 independently.*

Under *PriorSigns*, the random model vector is obtained by flipping signs on the components of the ground-truth model vector. This does not affect our bounds as we see in Theorem 3.2 that our bias lower bound only relies on squared components of the random model vector which are equivalent to the squared components of the ground truth model.

An important consequence of having a bounded condition number and independent coordinates is that with high probability the smallest eigenvalue of A_{-i} for all $i \geq 1$ is lower bounded by $n\lambda_{k+1}\rho_k$ up to constants. These assumptions allow Bartlett et al. [Bar+20] to prove Lemma B.1, which in turn allows us to derive the *StableLowerEig* condition. This is a simple consequence of B.1 and we provide details here for completeness.

Corollary B.3 (Our variant of StableLowerEig from Tsigler and Bartlet [TB23])**.** *For all $i \geq 1$, with probability at least $1 - 2e^{-n/c_2}$*

$$\mu_n(A_{-i}) \geq \frac{1}{c_2}\mu_n(\mathbb{E}\, A_k) = \frac{1}{c_2}\sum_{j>k} \lambda_j.$$

Proof. By Lemma B.1, for some absolute constant $c_1 \geq 1$ with probability at least $1 - 2e^{-n/c_1}$

$$\mu_n(A_k) \geq \frac{1}{c_1}\sum_{i>k} \lambda_i - c_1\lambda_{k+1}n.$$

The assumption $\rho_k \geq b$ for some $b \geq 1$ gives us

$$\begin{aligned}\frac{1}{c_1}\sum_{i>k}\lambda_i - c_1\lambda_{k+1}n &= \frac{1}{c_1}\lambda_{k+1}n\rho_k - c_1\lambda_{k+1}n \\ &\geq \left(\frac{1}{c_1} - \frac{c_1}{b}\right)\lambda_{k+1}n\rho_k \\ &= \left(\frac{1}{c_1} - \frac{c_1}{b}\right)\sum_{i>k}\lambda_i.\end{aligned}$$

Choosing $b > c_1^2$ and $c_2 = \max\{c_1, (1/c_1 - c_1/b)^{-1}\}$, we get that with probability at least $1 - 2e^{-n/c_2}$

$$\mu_n(A_k) \geq \frac{1}{c_2}\sum_{i>k}\lambda_i.$$

The next step is to extend this result to A_{-i} for all i.

For $i \leq k$, observe that $A_{-i} \succeq A_k$ gives us $\mu_n(A_{-i}) \geq \mu_n(A_k)$. For the case of $i > k$, we have

$$\begin{aligned}A_{-i} &= \sum_{j\neq i}\lambda_j z_j z_j^\top \\ &= \sum_{j\leq k}\lambda_j z_j z_j^\top + \sum_{j>k, j\neq i}\lambda_j z_j z_j^\top \\ &\succeq \lambda_1 z_1 z_1^\top + \sum_{j>k, j\neq i}\lambda_j z_j z_j^\top \\ &\succeq \lambda_i z_1 z_1^\top + \sum_{j>k, j\neq i}\lambda_j z_j z_j^\top.\end{aligned}$$

We assume that the features are independent and z_i is centered and whitened, so $\lambda_i z_1 z_1^\top + \sum_{j>k, j\neq i}\lambda_j z_j z_j^\top$ has the same distribution as $A_k = \sum_{j>k}\lambda_j z_j z_j^\top$. Therefore,

$$\begin{aligned}&\mathbb{P}\left(\mu_n(A_{-i}) \geq \frac{1}{c_2}\sum_{i>k}\lambda_i\right) \\ \geq&\mathbb{P}\left(\mu_n\left(\lambda_i z_1 z_1^\top + \sum_{j>k, j\neq i}\lambda_j z_j z_j^\top\right) \geq \frac{1}{c_2}\sum_{i>k}\lambda_i\right) \\ =&\mathbb{P}\left(\mu_n(A_k) \geq \frac{1}{c_2}\sum_{i>k}\lambda_i\right) \\ \geq&1 - 2e^{-n/c}.\end{aligned}$$

□

The following corollaries provide high-probability bounds on random subgaussian vectors with independent coordinates.

Corollary B.4 (Corollary 1 from Bartlett et al. [Bar+20])**.** *There is a universal constant c such that for any centered random vector $z \in \mathbb{R}^n$ with independent σ^2-subgaussian coordinates with unit variances, any random subspace $\mathscr{L}$ of $\mathbb{R}^n$ of codimension k that is independent of z, and any $t > 0$, with probability at least $1 - 3e^{-t}$,*

$$\|z\|^2 \leq n + c\sigma^2(t + \sqrt{nt}),$$
$$\|\Pi_{\mathscr{L}} z\|^2 \geq n - c\sigma^2(k + t + \sqrt{nt}),$$

where $\Pi_{\mathscr{L}}$ is the orthogonal projection on $\mathscr{L}$.

In our proofs, we will need to control the norm of z_i for all $i \leq p$ on the same high-probability event. In these cases we need to apply a union bound over the events in the summation. The following corollary shows how to invoke a union bound over ℓ of these events in such a way that the probability over the union of all such events holds with high probability that depends n.

Corollary B.5. *There is a universal constant c as defined in Corollary B.4. Let $z \in \mathbb{R}^n$ be a centered random vector with σ^2-subgaussian coordinates and unit variances. Let $\mathscr{L}$ be a random subspace of $\mathbb{R}^n$ of codimension k that is independent of z.*

For $0 < t < n/c_0$ and $k \in (0, n/c_1)$ for $c_1 > c_0$ with c_0 sufficiently large, with probability $1 - 3e^{-t}$,

$$\|z\|^2 \leq c_2 n$$
$$\|\Pi_{\mathscr{L}} z\|^2 \geq n/c_3$$

where c_2, c_3 only depends on c, c_0, σ.

We obtain a union bound over the intersection of ℓ of these events so long as $\ln(\ell) \leq n/c_0 \Rightarrow \ell \leq e^{n/c_0}$. Then for $k \in (0, n/c_1)$ for $c_1 > c_0$ with c_0 sufficiently large, if $\ell \leq e^{n/c_0}$, with probability at least $1 - 3e^{-n/c_0}$, ℓ of the above events independently hold.

Proof. Let Corollary B.4 hold with universal constant c. Then, with probability $1 - 3e^{-t}$ for $t > 0$,

$$\|z\|^2 \leq n + c\sigma^2(t + \sqrt{nt})$$
$$\|\Pi_{\mathscr{L}} z\|^2 \geq n - c\sigma^2(k + t + \sqrt{nt}).$$

Let $t \leq \frac{n}{c_0}$. Then we have that,

$$-\frac{n}{c_0} \leq -t \qquad -\frac{n}{\sqrt{c_0}} \leq -\sqrt{nt}.$$

Plugging in for $\|z\|^2$,

$$\begin{aligned}
\|z\|^2 &\leq n + c\sigma^2(t + \sqrt{nt}) \\
&\leq n + c\sigma^2(\frac{n}{c_0} + \frac{n}{\sqrt{c_0}}) \\
&= n(1 + c\sigma^2(c_0^{-1} + c_0^{-1/2})) \\
&= c_1 n
\end{aligned}$$

for c_1 only dependent on c, c_0, σ. Now, plugging in for $\|\Pi_{\mathscr{L}} z\|^2$,

$$\begin{aligned}\|\Pi_{\mathscr{L}} z\|^2 &\geq n - c\sigma^2(k + t + \sqrt{nt}) \\ &\geq n - c\sigma^2(k + \frac{n}{c_0} + \frac{n}{\sqrt{c_0}}) \\ &= n(1 - c\sigma^2(\frac{k}{n} + c_0^{-1} + c_0^{-1/2})).\end{aligned}$$

Let $k < \frac{n}{c_2}$ for $c_2 > c_0$. Then it is clear that $-\frac{k}{n} > -\frac{1}{c_2}$ and,

$$\begin{aligned}n(1 - c\sigma^2(\frac{k}{n} + c_0^{-1} + c_0^{-1/2})) &\geq n(1 - c\sigma^2(c_2^{-1} + c_0^{-1} + c_0^{-1/2})) \\ &= n/c_3\end{aligned}$$

for constant c_3 that only depends on c, σ^2, c_0. We finally require that $1 - c\sigma^2(c_1^{-1} + c_0^{-1} + c_0^{-1/2}) > 0$ which we can achieve by taking c_0 sufficiently large.

We now proceed to bound the union of ℓ of the complement events, in order to obtain a bound over the intersection of ℓ of these events.

For multiple z_i's, define by A_i the events shown above, that $\|z_i\|^2 \leq c_2 n$ and $\|\Pi_{\mathscr{L}_i} z_i\|^2 \geq n/c_3$ where z_i and $\mathscr{L}_i$ are defined analogous to $z, \mathscr{L}$ above. Then

$$\begin{aligned}P(\cup_{i=1}^{\ell}(A_i)^c) &\leq \sum_{i=1}^{\ell} P((A_i)^c) \\ &\leq \sum_{i=1}^{\ell} 3e^{-t} \\ &= 3\ell e^{-t}.\end{aligned}$$

Then $P(\cap_{i=1}^{\ell} A_i) \geq 1 - 3\ell e^{-t}$. Observing that $3\ell e^{-t} = 3e^{\ln(\ell)}e^{-t} = 3e^{-t+\ln(\ell)} = 3e^{-(t-\ln(\ell))}$ we can set the per-event t accordingly and obtain the necessary bound. We want $0 < t - \ln(\ell) \leq n/c_0$ to complete the bound. Therefore, we need that, per-event, $\ln(\ell) < t \leq n/c_0 + \ln(\ell)$. If $\ln(\ell) \leq n/c_0$ then this reduces to needing $\ln(\ell) < t \leq 2n/c_0$. Since each event is defined for $t \in (0, n/c_0]$ the union bound proof is complete by taking $t = n/c_0$ and requiring that $\ln(\ell) \leq n/c_0$.

□

The following lemma is necessary in order to extend a summation over random variables, each lower bounded by a real number with equal probability, to a unified lower bound over the entire summation.

Lemma B.6 (Lemma 9 from Bartlett et al. [Bar+20])**.** *Suppose $n \leq \infty$ and $\{\eta_i\}_{i=1}^n$ is a sequence of non-negative random variables, $\{t_i\}_{i=1}^n$ is a sequence of non-negative real numbers (at least one of which is strictly positive) such that for some $\delta \in (0, 1)$ and any $i \leq n$, $P(\eta_i > t_i) \geq 1 - \delta$. Then,*

$$P\left(\sum_{i=1}^{n} \eta_i \geq \frac{1}{2}\sum_{i=1}^{n} t_i\right) \geq 1 - 2\delta.$$

We now provide a minor generalization of Corollary S.6 in Bartlett et al. [Bar+20] that comes from replacing a_1 in a non-increasing sequence of non-negative numbers $\{a_i\}_{i=1}^p$ with $\max_i a_i$ and only requiring that $\{a_i\}_{i=1}^p$ is a sequence of non-negative numbers.

Corollary B.7. *There is a universal constant c such that for any sequence $\{a_i\}_{i=1}^p$ of non-negative numbers such that $\sum_{i=1}^p a_i < \infty$, and any independent, centered, σ-subexponential random variables $\{\xi_i\}_{i=1}^p$, and any $x > 0$, with probability at least $1 - 2e^{-cx}$,*

$$|\sum_i a_i \xi_i| \leq \sigma \max\left(x \max_i a_i, \sqrt{x \sum_{i=1}^p a_i^2}\right).$$

Lastly, the following identity will allow us to use the *PriorSigns* assumption to derive a new form for the bias term, which will be used for the proof of the lower bound.

Lemma B.8 (Identity for expectation of a quadratic form)**.** *Assume $M \in \mathbb{R}^{p\times p}$ is a symmetric matrix. For a random vector $x \in \mathbb{R}^p$ with mean $\mathbb{E}[x]$ and covariance $\mathrm{Cov}(x)$,*

$$\mathop{\mathbb{E}}_{x}[x^T M x] = \mathbb{E}[x]^T M\, \mathbb{E}[x] + \mathrm{tr}(M\mathit{Cov}(x)).$$

Proof.

$$\begin{aligned}
\mathop{\mathbb{E}}_{x}[x^T M x] &= \mathbb{E}[\mathrm{tr}(x^T M x)] \\
&= \mathbb{E}[\mathrm{tr}(M x x^T)] \\
&= \mathrm{tr}(M\, \mathbb{E}[x x^T]) \\
&= \mathrm{tr}(M\mathrm{Cov}(x) + M\, \mathbb{E}[x]\, \mathbb{E}[x]^T) \\
&= \mathrm{tr}(M\mathrm{Cov}(x)) + \mathrm{tr}(\mathbb{E}[x] M\, \mathbb{E}[x]^T) \\
&= \mathrm{tr}(M\mathrm{Cov}(x)) + \mathbb{E}[x] M\, \mathbb{E}[x]^T.
\end{aligned}$$

□

C Proof of excess risk bound

We start by restating Theorem 2.1.

Theorem 2.1. *(Target excess risk decomposition) The excess risk of the MNI trained on the source data, when evaluated on the target distribution, satisfies*

$$R(\widehat{\theta}(\boldsymbol{y}_{\mathsf{s}}), \mathcal{D}_{\mathsf{t}}) \leq 4B_1 + 4B_2 + 2V_{\boldsymbol{\varepsilon}_{\mathsf{s}}}, \tag{2}$$

and

$$\mathop{\mathbb{E}}_{\boldsymbol{\varepsilon}_{\mathsf{s}}} R(\widehat{\theta}(\boldsymbol{y}_{\mathsf{s}}), \mathcal{D}_{\mathsf{t}}) = B_1 + B_2 + \mathop{\mathbb{E}}_{\boldsymbol{\varepsilon}_{\mathsf{s}}} V_{\boldsymbol{\varepsilon}_{\mathsf{s}}} + 2(\theta^*_{\mathsf{t}} - \theta^*_{\mathsf{s}})^\top \Sigma_{\mathsf{t}} (\theta^*_{\mathsf{s}} - \widehat{\theta}(X\theta^*_{\mathsf{s}})),$$

where we define

$$B_1 := \|\theta_{\mathsf{s}}^* - \theta_{\mathsf{t}}^*\|_{\Sigma_{\mathsf{t}}}^2, \tag{3}$$

$$B_2 := \|\theta_{\mathsf{s}}^* - \widehat{\theta}(X\theta_{\mathsf{s}}^*)\|_{\Sigma_{\mathsf{t}}}^2, \tag{4}$$

$$V_{\boldsymbol{\varepsilon}_{\mathsf{s}}} := \|\widehat{\theta}(\boldsymbol{\varepsilon}_{\mathsf{s}})\|_{\Sigma_{\mathsf{t}}}^2, \tag{5}$$

and $\|x\|_M^2 := x^\top M x$.

Proof. Let us begin by noting that the excess risk of any θ is given by,

$$\begin{aligned}
R(\theta) &= \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}_{\mathsf{t}}}\left[\left(y - x^\top\theta\right)^2\right] - \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}_{\mathsf{t}}}\left[\left(y - x^\top\theta_{\mathsf{t}}^*\right)^2\right] \\
&= \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}_{\mathsf{t}}}\left[\left(y - x^\top\theta_{\mathsf{t}}^* + x^\top\theta_{\mathsf{t}}^* - x^\top\theta\right)^2\right] - \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}_{\mathsf{t}}}\left[\left(y - x^\top\theta_{\mathsf{t}}^*\right)^2\right] \\
&= \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}_{\mathsf{t}}}\left[\left(x^\top\theta_{\mathsf{t}}^* - x^\top\theta\right)^2\right] + 2\mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}_{\mathsf{t}}}\left[\left(y - x^\top\theta_{\mathsf{t}}^*\right)\left(x^\top\theta_{\mathsf{t}}^* - x^\top\theta\right)\right] \\
&\overset{(i)}{=} \mathop{\mathbb{E}}_{x\sim\mathcal{D}_{\mathsf{t}}}\left[\left(x^\top\theta_{\mathsf{t}}^* - x^\top\theta\right)^2\right].
\end{aligned} \tag{16}$$

Equality (i) uses that, conditional on x, $y - x^\top\theta_{\mathsf{t}}^*|x$ is mean-zero, which is given in Assumption 3 (*well-specified - target*). So that

$$\mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}_{\mathsf{t}}}\left[\left(y - x^\top\theta_{\mathsf{t}}^*\right)\left(x^\top\theta_{\mathsf{t}}^* - x^\top\theta\right)\right] = \mathbb{E}\left[\left(x^\top\theta_{\mathsf{t}}^* - x^\top\theta\right)\mathbb{E}\left[\left(y - x^\top\theta_{\mathsf{t}}^*\right)|x\right]\right] = 0.$$

We now note that the source-data MNI can be decomposed as follows,

$$\begin{aligned}
\widehat{\theta}(\boldsymbol{y}_{\mathsf{s}}) &= X^\top(X_{\mathsf{s}}X_{\mathsf{s}}^\top)^{-1}\boldsymbol{y}_{\mathsf{s}} \\
&= X^\top(X_{\mathsf{s}}X_{\mathsf{s}}^\top)^{-1}(X\theta_{\mathsf{s}}^* + \boldsymbol{\varepsilon}_{\mathsf{s}}) \\
&= \widehat{\theta}(X\theta_{\mathsf{s}}^*) + \widehat{\theta}(\boldsymbol{\varepsilon}_{\mathsf{s}})
\end{aligned}$$

We can thus continue from (16) to characterize the excess risk of the source-data MNI as

$$\begin{aligned}
R(\widehat{\theta}(\boldsymbol{y}_{\mathsf{s}})) &= \mathop{\mathbb{E}}_{x\sim\mathcal{D}_{\mathsf{t}}}\left[\left(x^\top\theta_{\mathsf{t}}^* - x^\top\widehat{\theta}(\boldsymbol{y}_{\mathsf{s}})\right)^2\right] \\
&= \mathop{\mathbb{E}}_{x\sim\mathcal{D}_{\mathsf{t}}}\left[\left(x^\top\theta_{\mathsf{t}}^* - x^\top(\widehat{\theta}(X\theta_{\mathsf{s}}^*) + \widehat{\theta}(\boldsymbol{\varepsilon}_{\mathsf{s}}))\right)^2\right] \\
&= \mathop{\mathbb{E}}_{x\sim\mathcal{D}_{\mathsf{t}}}\left[\left(x^\top(\theta_{\mathsf{t}}^* - \widehat{\theta}(X\theta_{\mathsf{s}}^*)) - x^\top\widehat{\theta}(\boldsymbol{\varepsilon}_{\mathsf{s}})\right)^2\right] \\
&\overset{(i)}{\leq} \mathop{\mathbb{E}}_{x\sim\mathcal{D}_{\mathsf{t}}}\left[2\left(x^\top(\theta_{\mathsf{t}}^* - \widehat{\theta}(X\theta_{\mathsf{s}}^*))\right)^2 + 2\left(x^\top\widehat{\theta}(\boldsymbol{\varepsilon}_{\mathsf{s}})\right)^2\right] \\
&= \mathop{\mathbb{E}}_{x\sim\mathcal{D}_{\mathsf{t}}}\left[2\left(x^\top(\theta_{\mathsf{t}}^* - \theta_{\mathsf{s}}^* + \theta_{\mathsf{s}}^* - \widehat{\theta}(X\theta_{\mathsf{s}}^*))\right)^2 + 2\left(x^\top\widehat{\theta}(\boldsymbol{\varepsilon}_{\mathsf{s}})\right)^2\right] \\
&\overset{(ii)}{\leq} \mathop{\mathbb{E}}_{x\sim\mathcal{D}_{\mathsf{t}}}\left[4\left(x^\top(\theta_{\mathsf{t}}^* - \theta_{\mathsf{s}}^*)\right)^2 + 4\left(x^\top(\theta_{\mathsf{s}}^* - \widehat{\theta}(X\theta_{\mathsf{s}}^*))\right)^2 + 2\left(x^\top\widehat{\theta}(\boldsymbol{\varepsilon}_{\mathsf{s}})\right)^2\right].
\end{aligned} \tag{17}$$

In inequalities (i) and (ii), we have used Young's inequality, which implies $(a-b)^2 \le 2(a-c)^2 + 2(b-c)^2$ for any $a, b, c \in \mathbb{R}$. Recalling that

$$\|x\|_M^2 := x^\top M x,$$

it is apparent that the first term is just the weighted distance between the source and target vectors,

$$\mathop{\mathbb{E}}_{x\sim\mathcal{D}_\mathsf{t}}\left[\left(x^\top(\theta_\mathsf{t}^* - \theta_\mathsf{s}^*)\right)^2\right] = (\theta_\mathsf{t}^* - \theta_\mathsf{s}^*)^\top \mathop{\mathbb{E}}_{x\sim\mathcal{D}_\mathsf{t}}\left[xx^\top\right](\theta_\mathsf{t}^* - \theta_\mathsf{s}^*) = \|\theta_\mathsf{s}^* - \theta_\mathsf{t}^*\|_{\Sigma_\mathsf{t}}^2. \tag{18}$$

The second term looks quite similar to the bias term, B, in Bartlett et al. [Bar+20] and Tsigler and Bartlet [TB23].

$$\begin{aligned}
&\mathop{\mathbb{E}}_{x\sim\mathcal{D}_\mathsf{t}}\left[\left(x^\top(\theta_\mathsf{s}^* - \widehat{\theta}(X\theta_\mathsf{s}^*))\right)^2\right] \\
&\quad= \left(\theta_\mathsf{s}^* - \widehat{\theta}\,(X\theta_\mathsf{s}^*)\right) \mathop{\mathbb{E}}_{x\sim\mathcal{D}_\mathsf{t}}[xx^\top] \left(\theta_\mathsf{s}^* - \widehat{\theta}\,(X\theta_\mathsf{s}^*)\right) \\
&\quad= \|\theta_\mathsf{s}^* - \widehat{\theta}\,(X\theta_\mathsf{s}^*)\,\|_{\Sigma_\mathsf{t}}^2.
\end{aligned} \tag{19}$$

The key difference with the standard supervised setting is that now the quantitiy in the middle is Σ_t, not Σ_s. Equivalently, the norm on $\theta_\mathsf{s}^* - \widehat{\theta}(X\theta_\mathsf{s}^*)$ is induced by Σ_t rather than Σ_s.

And finally, the third term is similar to the variance term, C, in Bartlett et al. [Bar+20]:

$$\begin{aligned}
\mathop{\mathbb{E}}_{x\sim\mathcal{D}_\mathsf{t}}\left[\left(x^\top\widehat{\theta}(\boldsymbol{\varepsilon}_\mathsf{s})\right)^2\right] &= \widehat{\theta}(\boldsymbol{\varepsilon}_\mathsf{s})^\top \mathop{\mathbb{E}}_{x\sim\mathcal{D}_\mathsf{t}}[xx^\top]\widehat{\theta}(\boldsymbol{\varepsilon}_\mathsf{s}) \\
&= \widehat{\theta}(\boldsymbol{\varepsilon}_\mathsf{s})^\top \Sigma_\mathsf{t} \widehat{\theta}(\boldsymbol{\varepsilon}_\mathsf{s}) \\
&= \|\widehat{\theta}(\boldsymbol{\varepsilon}_\mathsf{s})\|_{\Sigma_\mathsf{t}}^2.
\end{aligned} \tag{20}$$

As in the bias term, the only difference is that the middle term is Σ_t rather than Σ_s. Equivalently, the norm on $\widehat{\theta}(\boldsymbol{\varepsilon}_\mathsf{s})$ is induced by Σ_t rather than Σ_s.

Putting it all together, we get the following upper bound for the excess risk of the minimum-norm interpolator on the training data,

$$R(\widehat{\theta}(\boldsymbol{y}_\mathsf{s})) \le 4\|\theta_\mathsf{s}^* - \theta_\mathsf{t}^*\|_{\Sigma_\mathsf{t}}^2 + 4\|\theta_\mathsf{s}^* - \widehat{\theta}(X\theta_\mathsf{s}^*)\|_{\Sigma_\mathsf{t}}^2 + 2\|\widehat{\theta}(\boldsymbol{\varepsilon}_\mathsf{s})\|_{\Sigma_\mathsf{t}}^2.$$

This completes the upper bound for the risk.

For the lower bound, we have

$$\begin{aligned}
\mathop{\mathbb{E}}_{\boldsymbol{\varepsilon}_\mathsf{s}} R(\widehat{\theta}(\boldsymbol{y}_\mathsf{s})) &= \mathop{\mathbb{E}}_{\boldsymbol{\varepsilon}_\mathsf{s}, x\sim\mathcal{D}_\mathsf{t}}\left[\left(x^\top\theta_\mathsf{t}^* - x^\top\widehat{\theta}(\boldsymbol{y}_\mathsf{s})\right)^2\right] \\
&= \mathop{\mathbb{E}}_{\boldsymbol{\varepsilon}_\mathsf{s}, x\sim\mathcal{D}_\mathsf{t}}\left[\left(x^\top(\theta_\mathsf{t}^* - \widehat{\theta}(X\theta_\mathsf{s}^*)) - x^\top\widehat{\theta}(\boldsymbol{\varepsilon}_\mathsf{s})\right)^2\right] \\
&= \mathop{\mathbb{E}}_{\boldsymbol{\varepsilon}_\mathsf{s}, x\sim\mathcal{D}_\mathsf{t}}\Bigg[\left(x^\top(\theta_\mathsf{t}^* - \widehat{\theta}(X\theta_\mathsf{s}^*))\right)^2 - 2\cdot x^\top(\theta_\mathsf{t}^* - \widehat{\theta}(X\theta_\mathsf{s}^*))\cdot x^\top\widehat{\theta}(\boldsymbol{\varepsilon}_\mathsf{s}) \\
&\qquad + \left(x^\top\widehat{\theta}(\boldsymbol{\varepsilon}_\mathsf{s})\right)^2\Bigg] \\
&\overset{(i)}{=} \mathop{\mathbb{E}}_{x\sim\mathcal{D}_\mathsf{t}}\left[\left(x^\top(\theta_\mathsf{t}^* - \widehat{\theta}(X\theta_\mathsf{s}^*))\right)^2\right] + \mathop{\mathbb{E}}_{\boldsymbol{\varepsilon}_\mathsf{s}, x\sim\mathcal{D}_\mathsf{t}}\left[\left(x^\top\widehat{\theta}(\boldsymbol{\varepsilon}_\mathsf{s})\right)^2\right]
\end{aligned}$$

The equality (i) uses that, conditional on X, ε_{s} is zero-mean. Note that the second term above is just $\mathbb{E}_{\varepsilon_{\mathsf{s}}} \|\widehat{\theta}(\varepsilon_{\mathsf{s}})\|^2_{\Sigma_{\mathsf{t}}}$, so we need only deal with the first term. Adding and subtracting θ^*_{s} inside the square and expanding, we have

$$\begin{aligned}
&\mathop{\mathbb{E}}_{x\sim\mathcal{D}_{\mathsf{t}}}\left[\left(x^\top(\theta^*_{\mathsf{t}} - \widehat{\theta}(X\theta^*_{\mathsf{s}}))\right)^2\right] \\
&= \mathop{\mathbb{E}}_{x\sim\mathcal{D}_{\mathsf{t}}}\left[\left(x^\top(\theta^*_{\mathsf{t}} - \theta^*_{\mathsf{s}})\right)^2\right] + \mathop{\mathbb{E}}_{x\sim\mathcal{D}_{\mathsf{t}}}\left[\left(x^\top(\theta^*_{\mathsf{s}} - \widehat{\theta}(X\theta^*_{\mathsf{s}}))\right)^2\right] \\
&\quad + 2\mathop{\mathbb{E}}_{x\sim\mathcal{D}_{\mathsf{t}}}\left[(\theta^*_{\mathsf{t}} - \theta^*_{\mathsf{s}})^\top xx^\top(\theta^*_{\mathsf{s}} - \widehat{\theta}(X\theta^*_{\mathsf{s}}))\right] \\
&= \|\theta^*_{\mathsf{t}} - \theta^*_{\mathsf{s}}\|^2_{\Sigma_{\mathsf{t}}} + \|\theta^*_{\mathsf{s}} - \widehat{\theta}(X\theta^*_{\mathsf{s}})\|^2_{\Sigma_{\mathsf{t}}} + 2(\theta^*_{\mathsf{t}} - \theta^*_{\mathsf{s}})^\top\Sigma_{\mathsf{t}}(\theta^*_{\mathsf{s}} - \widehat{\theta}(X\theta^*_{\mathsf{s}})).
\end{aligned}$$

□

D Overview of variance and bias proof techniques

The central pillar of both proofs is controlling the eigenvalues of A_k, which in turn provides certain bounds on the eigenvalues of A and A_{-i}. A key finding of Bartlett et al. [Bar+20] is that once ρ_k is large enough, all eigenvalues of A_k are identical up to a constant factor. Specifically,

$$z^T Az \approx n^2\lambda_{k+1}\rho_k, \qquad\qquad z^T A^{-1}z \approx n(n\lambda_{k+1}\rho_k)^{-1}.$$

D.1 Variance

Due to independence between the components of ε_{s}, the variance term from Eqn. 6 can be expressed as

$$\begin{aligned}
V &= \mathop{\mathbb{E}}_{\varepsilon_{\mathsf{s}}}[V_{\varepsilon_{\mathsf{s}}}/v_{\varepsilon}^2] \\
&= \operatorname{tr}(A^{-1}X\tilde{\Sigma}X^\top A^{-1}) \\
&= \sum_{i=1}^{p}\tilde{\lambda}_i\lambda_i z_i^T A^{-2}z_i.
\end{aligned}$$

Now that we are dealing with a sum of quadratic forms, we consider the first k^* signal and last $p-k^*$ noise components separately. Using the Sherman-Morrison formula the former can be written as

$$\begin{aligned}
\sum_{i\le k^*}\tilde{\lambda}_i\lambda_i z_i^T A^{-2}z_i &= \sum_{i\le k^*}\frac{\tilde{\lambda}_i}{\lambda_i}\frac{\lambda_i^2 z_i^T A_{-i}^{-2}z_i}{(1+\lambda_i z_i^T A_{-i}^{-1}z_i)^2} \\
&\approx \sum_{i\le k^*}\frac{\tilde{\lambda}_i}{\lambda_i}\frac{\lambda_i^2 n(n\lambda_{k+1}\rho_k)^{-2}}{\lambda_i^2 n^2(n\lambda_{k+1}\rho_k)^{-2}} \\
&= \sum_{i\le k^*}\frac{\tilde{\lambda}_i}{\lambda_i}\frac{1}{n},
\end{aligned}$$

where $\lambda_i z_i^T A_{-i}^{-1}z_i$ dominates 1 for $i \le k^*$. For the sum over the noise components the 1 in the denominator dominates the other term and so we directly analyze the tail contributions as,

$$\sum_{i>k^*}\frac{\tilde{\lambda}_i}{\lambda_i}\lambda_i^2 z_i^T A^{-2}z_i \approx \sum_{i>k^*}\frac{\tilde{\lambda}_i}{\lambda_i}\lambda_i^2 n(n\lambda_{k+1}\rho_k)^{-2}.$$

The result is that the variance term is upper and lower bounded by

$$\frac{1}{n}\sum_{i=1}^{k}\frac{\tilde{\lambda}_i}{\lambda_i}+\sum_{i>k}\frac{\tilde{\lambda}_i}{\lambda_i}\left(\frac{\lambda_i^2}{n\lambda_{k+1}^2\rho_k^2}\right)$$

times constant factors.

D.2 Bias

As in Eqn. 4, the bias term is given by

$$\begin{aligned}
B_2 &= \|\theta_{\mathsf{s}}^* - X^TA^{-1}X\theta_{\mathsf{s}}^*\|_{\Sigma_{\mathsf{t}}}^2 \\
&= \mathrm{tr}(\theta_{\mathsf{s}}^{*T}(I-X^TA^{-1}X)\Sigma_{\mathsf{t}}(I-X^TA^{-1}X)\theta_{\mathsf{s}}^*) \\
&\leq \mathrm{tr}(\theta_{\mathsf{s}}^*\theta_{\mathsf{s}}^{*T})\cdot\mathrm{tr}((I-X^TA^{-1}X)\Sigma_{\mathsf{t}}(I-X^TA^{-1}X)) \\
&= \|\theta_{\mathsf{s}}^*\|^2\sum_{i=1}^{p}\frac{\tilde{\lambda}_i}{\lambda_i}\sum_{j=1}^{p}\left(e_i[j]-\sqrt{\lambda_i\lambda_j}z_i^\top A^{-1}z_j\right)^2,
\end{aligned}$$

where we use the Cauchy-Scharwz inequality to separate the parameter vector from the quadratic form. A quick application of the Sherman-Morrison formula allows us to write

$$B_2 \leq \|\theta_{\mathsf{s}}^*\|^2\sum_{i=1}^{p}\tilde{\lambda}_i\frac{1}{1+\lambda_i z_i^TA_{-i}^{-1}z_i}.$$

From here, we once again exert control over the eigenvalues of A_{-i} to get

$$\frac{1}{1+\lambda_i z_i^TA_{-i}^{-1}z_i}\approx\frac{1}{1+\frac{\lambda_i}{\lambda_{k+1}\rho_k}},$$

which completes the upper bound proof sketch.

Note that the looseness of the bias bounds largely stems from the application of the Cauchy-Schwarz inequality. The only situations in which the bound becomes an equality are when

$$c\theta_{\mathsf{s}}^* = (I-X^TA^{-1}X)\Sigma_{\mathsf{t}}^{1/2}$$

for some scalar $c\in\mathbb{R}$ or when θ_{s}^* is the zero vector.

Between the upper and lower bounds, the latter is likely tighter due to the use of the *PriorSigns* assumption. As detailed in Appendix F.2, it allows us to write

$$B \geq \theta_{\mathsf{s}}^{*T}(I-\mathrm{diag}(X^TA^{-1}X))\Sigma_{\mathsf{t}}(I-\mathrm{diag}(X^TA^{-1}X))\theta_{\mathsf{s}}^*,$$

where for a matrix $Q\in\mathbb{R}^{m\times m}$, we use $\mathrm{diag}(Q)\in\mathbb{R}^{m\times m}$ to denote zeroed off-diagonal entries. The contribution of the off-diagonal entries is non-negative and dominated by the diagonals, so they can be dropped in the lower bound while preserving tightness under the *PriorSigns* assumption. In general, non-negative terms cannot be discarded in the proof of an upper bound, so we resort to the Cauchy-Schwarz inequality in order to avoid addressing the off-diagonals directly. However, decoupling the model vector θ_{s}^* from the matrix $(I-X^TA^{-1}X)\Sigma_{\mathsf{t}}^{1/2}$ introduces another degree of looseness, contributing to the gap between our bounds. Improving our upper bound will require controlling the off-diagonals of this matrix product with a technique more appropriate than Cauchy-Schwarz.

E Proof of variance bounds

Theorem 3.1. *(Upper and lower bounds for the variance term) There exist universal constants $b, c_1 > 1$ given in Lemma B.1, a universal constant c_2 given in Lemma B.4 and a constant $c > 1$ that only depends on σ_x, c_1, c_2, such that for $k \in (0, n/c)$, with probability at least $1 - 10e^{-n/c}$,*

$$V \geq \frac{1}{cn} \sum_{i=1}^{p} \frac{\tilde{\lambda}_i}{\lambda_i} \min\left(1, \frac{\lambda_i^2}{\lambda_{k+1}^2(\rho_k + 1)^2}\right) := \underline{V}. \tag{8}$$

If in addition $\rho_k \geq b$, with probability $1 - 7e^{-n/c}$,

$$V/c \leq \frac{1}{n} \sum_{i=1}^{k} \frac{\tilde{\lambda}_i}{\lambda_i} + n \frac{\sum_{i>k} \tilde{\lambda}_i \lambda_i}{(\sum_{i>k} \lambda_i)^2} := \overline{V}. \tag{9}$$

Proof. We derive the variance terms necessary here and finish the proof of the upper bound in Appendix E.1 and the lower bound in Appendix E.2.

We follow the proof techniques in Bartlett et al. [Bar+20] and Tsigler and Bartlet [TB23]. Observe that we can express the variance term as follows,

$$\begin{aligned} V &= \mathop{\mathbb{E}}_{\varepsilon_{\mathsf{s}}}[V_{\varepsilon_{\mathsf{s}}}/v_\varepsilon^2] \\ &= \mathop{\mathbb{E}}_{\varepsilon_{\mathsf{s}}}[\|X^\top (XX^\top)^{-1} \varepsilon_{\mathsf{s}}\|_{\Sigma_{\mathsf{t}}}^2 / v_\varepsilon^2]. \end{aligned}$$

Defining $A = XX^\top$,

$$\begin{aligned} V &= \mathop{\mathbb{E}}_{\varepsilon_{\mathsf{s}}}[\|X^\top A^{-1} \varepsilon_{\mathsf{s}}\|_{\Sigma_{\mathsf{t}}}^2 / v_\varepsilon^2] \\ &= \mathop{\mathbb{E}}_{\varepsilon_{\mathsf{s}}}[(\varepsilon_{\mathsf{s}}^\top A^{-1} X \Sigma_{\mathsf{t}} X^\top A^{-1} \varepsilon_{\mathsf{s}}) / v_\varepsilon^2] \\ &= \mathop{\mathbb{E}}_{\varepsilon_{\mathsf{s}}}[\mathrm{tr}(\varepsilon_{\mathsf{s}}^\top A^{-1} X \Sigma_{\mathsf{t}} X^\top A^{-1} \varepsilon_{\mathsf{s}}) / v_\varepsilon^2]. \end{aligned}$$

Using the trace trick,

$$\begin{aligned} V &= \mathrm{tr}(A^{-1} X \Sigma_{\mathsf{t}} X^\top A^{-1} \mathop{\mathbb{E}}_{\varepsilon_{\mathsf{s}}}[\varepsilon_{\mathsf{s}} \varepsilon_{\mathsf{s}}^\top]) / v_\epsilon^2 \\ &= \mathrm{tr}(A^{-1} X \Sigma_{\mathsf{t}} X^\top A^{-1} v_\epsilon^2 I_n) / v_\epsilon^2 \\ &= \mathrm{tr}(A^{-1} X \Sigma_{\mathsf{t}} X^\top A^{-1}) \\ &= \mathrm{tr}(X \Sigma_{\mathsf{t}} X^\top A^{-2}) \\ &= \mathrm{tr}((\sum_{i=1}^{p} \tilde{\lambda}_i x^i (x^i)^\top) A^{-2}) \\ &= \mathrm{tr}((\sum_{i=1}^{p} \tilde{\lambda}_i \lambda_i z_i z_i^\top) A^{-2}) \end{aligned}$$

where $x^i \in \mathbb{R}^n$ and $\frac{x^i}{\sqrt{\lambda_i}} = z_i \in \mathbb{R}^n$ are columns of $X \in \mathbb{R}^{n\times p}$ and $X\Sigma_{\mathsf{s}}^{-1/2} \in \mathbb{R}^{n\times p}$, respectively. Continuing the calculation, we have that

$$\begin{aligned} V &= \sum_{i=1}^{p} \tilde{\lambda}_i \lambda_i \mathrm{tr}(z_i^T A^{-2} z_i) \\ &= \sum_{i=1}^{p} \tilde{\lambda}_i \lambda_i \mathrm{tr}(z_i^T (A_{-i} + z_i z_i^T \lambda_i)^{-2} z_i) \\ &= \sum_{i=1}^{p} \tilde{\lambda}_i \lambda_i \frac{z_i^T A_{-i}^{-2} z_i}{(1+\lambda_i z_i^T A_{-i}^{-1} z_i)^2} \end{aligned}$$

where $A_{-i} = XX^T - \lambda_i z_i z_i^T = \sum_{j\neq i} \lambda_j z_j z_j^T$. This expression will serve as the starting point for the variance term, which we will now proceed to upper and lower bound.

□

E.1 Upper bound

After isolating the contribution of $\frac{\tilde{\lambda}_i}{\lambda_i}$, most of the components of this proof are as given in the proof of Lemma 6 in Bartlett et al. [Bar+20]. For completeness, we replicate them here and refer the reader to their paper for further details and intuitions.

We start by separating the variance term into the top k components and the bottom $p-k$ components as follows,

$$V = \sum_{i=1}^{k} \frac{\tilde{\lambda}_i}{\lambda_i} \frac{\lambda_i^2 z_i^T A_{-i}^{-2} z_i}{(1+\lambda_i z_i^T A_{-i}^{-1} z_i)^2} + \sum_{i>k} \frac{\tilde{\lambda}_i}{\lambda_i} (\lambda_i^2 z_i^T A^{-2} z_i).$$

Fix constants $b, c_1 \geq 1$ as defined in Lemma B.1. Then, with probability $1 - 2e^{-n/c_1}$, if $\rho_k \geq b$ then for all $z \in \mathbb{R}^n$ and $i \in [1,k]$,

$$\begin{aligned} z_i^T A_{-i}^{-2} z_i &\leq \mu_1(A_{-i}^{-2}) \|z_i\|^2 \\ &\leq \mu_n(A_{-i})^{-2} \|z_i\|^2 \\ &\leq \frac{c_1^2 \|z_i\|^2}{(n\lambda_{k+1}\rho_k)^2} \end{aligned}$$

and on the same event,

$$\begin{aligned} z_i^T A_{-i}^{-1} z_i &\geq (\Pi_{\mathcal{L}_i} z_i)^T A_{-i}^{-1} (\Pi_{\mathcal{L}_i} z_i) \\ &\geq \mu_n(A_{-i}^{-1}) \|\Pi_{\mathcal{L}_i} z_i\|^2 \\ &\geq \mu_{k+1}(A_{-i})^{-1} \|\Pi_{\mathcal{L}_i} z_i\|^2 \\ &\geq \frac{\|\Pi_{\mathcal{L}_i} z_i\|^2}{nc_1 \lambda_{k+1} \rho_k} \end{aligned}$$

where $\Pi_{\mathcal{L}_i}$ is the orthogonal projection onto the span of the bottom $n-k$ eigenvectors of A_{-i}. It is important to use the projection onto the bottom eigenvectors of A_{-i} in lower bounding the denominator term because we have to use the fact that $\mu_n(A_{-i}^{-1}) \geq \mu_1(A_{-i})^{-1}$. When we don't do the projection, then z_i is affected by all of A_{-i} and so the largest eigenvalue that affects this expression is $\mu_1(A_{-i}) = \lambda_1$. After doing this projection, we no longer have contributions from the top k eigenvectors / eigenvalues in the summation of $z_i^T A_{-i}^{-1} z_i$. Therefore, the largest eigenvalue that affects this summation is now λ_{k+1} instead of λ_1, and so we can use this in our lower bound instead, as desired.

Putting it together, for $i \leq k$,

$$\begin{aligned}\frac{\lambda_i^2 z_i^T A_{-i}^{-2} z_i}{(1+\lambda_i z_i^T A_{-i}^{-1} z_i)^2} &\leq \frac{z_i^T A_{-i}^{-2} z_i}{(z_i^T A_{-i}^{-1} z_i)^2} \\ &\leq c_1^4 \frac{\|z_i\|^2}{\|\Pi_{\mathcal{L}_i} z_i\|^4}.\end{aligned}$$

We now invoke Corollary B.5 with a union bound over k events. Let $t < n/c_0$ and $k \in (0, n/c)$ for $c > c_0$ and c_0 sufficiently large. Since $k < n/c$ we also satisfy the union bound condition that $\ln(k) < n/c$. Then, with probability at least $1 - 3e^{-t}$,

$$\begin{aligned}\|z_i\|^2 &\leq c_2 n \\ \|\Pi_{\mathcal{L}_i} z_i\|^2 &\geq n/c_3\end{aligned}$$

for constants c_2, c_3 that only depend on σ_x, c_0, and a universal constant c as defined in Corollary B.4.

Altogether, with probability $1 - 5e^{-n/c_0}$ for c_0 sufficiently large,

$$\begin{aligned}\sum_{i=1}^k \frac{\tilde{\lambda}_i}{\lambda_i}\left(\frac{\lambda_i^2 z_i^T A_{-i}^{-2} z_i}{(1+\lambda_i z_i^T A_{-i}^{-1} z_i)^2}\right) &\leq \sum_{i=1}^k \frac{\tilde{\lambda}_i}{\lambda_i} c_1^4 \frac{\|z_i\|^2}{\|\Pi_{\mathcal{L}_i} z_i\|^4} \\ &\leq \sum_{i=1}^k \frac{\tilde{\lambda}_i}{\lambda_i} c_1^4 \frac{c_2 c_3^2}{n} \\ &= c_4 \sum_{i=1}^k \frac{\tilde{\lambda}_i}{\lambda_i} \frac{1}{n}.\end{aligned}$$

On the same event we use to bound $\mu_{k+1}(A_{-i})$ via Lemma B.1, we also have that $\mu_1(A^{-2}) \leq \mu_n(A)^{-2}$. As such,

$$\sum_{i>k} \frac{\tilde{\lambda}_i}{\lambda_i}(\lambda_i^2 z_i^T A^{-2} z_i) \leq \frac{c_1^2 \sum_{i>k} \frac{\tilde{\lambda}_i}{\lambda_i} \lambda_i^2 \|z_i\|^2}{(n\lambda_{k+1}\rho_k)^2}.$$

Then by Corollary B.7, there is a universal constant a such that with probability at least $1 - 2e^{-t}$ for $t < n/c_0$

and $c_0 > a^{-1}$,

$$\begin{aligned}
\sum_{i>k} \frac{\tilde{\lambda}_i}{\lambda_i} \lambda_i^2 \|z_i\|^2 &\leq \sigma_x^2 \max(t \max_{i>k}(\tilde{\lambda}_i \lambda_i), \sqrt{t \sum_{i>k} (\tilde{\lambda}_i \lambda_i)^2}) \\
&\leq n \sum_{i>k} \tilde{\lambda}_i \lambda_i + \sigma_x^2 \max(t \max_{i>k}(\tilde{\lambda}_i \lambda_i), \sqrt{tn \sum_{i>k} (\tilde{\lambda}_i \lambda_i)^2}) \\
&\leq n \sum_{i>k} \tilde{\lambda}_i \lambda_i + \sigma_x^2 \max(t \sum_{i>k} \tilde{\lambda}_i \lambda_i, \sqrt{tn} \sum_{i>k} \tilde{\lambda}_i \lambda_i) \\
&\leq c_5 n \sum_{i>k} \tilde{\lambda}_i \lambda_i \\
&= c_5 n \sum_{i>k} \frac{\tilde{\lambda}_i}{\lambda_i} \lambda_i^2.
\end{aligned}$$

Altogether,

$$\begin{aligned}
\sum_{i>k} \frac{\tilde{\lambda}_i}{\lambda_i} (\lambda_i^2 z_i^T A^{-2} z_i) &\leq \frac{c_1^2 \sum_{i>k} \frac{\tilde{\lambda}_i}{\lambda_i} \lambda_i^2 \|z_i\|^2}{(n \lambda_{k+1} \rho_k)^2} \\
&\leq \frac{c_1^2 c_5 n}{(n \lambda_{k+1} \rho_k)^2} \sum_{i>k} \frac{\tilde{\lambda}_i}{\lambda_i} \lambda_i^2 \\
&= c_6 \sum_{i>k} \frac{\tilde{\lambda}_i}{\lambda_i} \left(\frac{\lambda_i^2}{n \lambda_{k+1}^2 \rho_k^2} \right).
\end{aligned}$$

By taking $c > \max(c_0, c_4, c_6)$ we have that with probability $1 - 7e^{-n/c}$,

$$\begin{aligned}
V &\leq c \left(\sum_{i=1}^{k} \frac{\tilde{\lambda}_i}{\lambda_i} \frac{1}{n} + \sum_{i>k} \frac{\tilde{\lambda}_i}{\lambda_i} \left(\frac{\lambda_i^2}{n \lambda_{k+1}^2 \rho_k^2} \right) \right) \\
&= \frac{1}{n} \sum_{i=1}^{k} \frac{\tilde{\lambda}_i}{\lambda_i} + n \frac{\sum_{i>k} \tilde{\lambda}_i \lambda_i}{(\sum_{i>k} \lambda_i)^2}.
\end{aligned}$$

E.2 Lower bound

Recall that the variance takes the form,

$$V = \sum_{i=1}^{p} \tilde{\lambda}_i \lambda_i \frac{z_i^T A_{-i}^{-2} z_i}{(1 + \lambda_i z_i^T A_{-i}^{-1} z_i)^2}.$$

By Cauchy-Schwartz,

$$(z_i^T A_{-i}^{-1} z_i)^2 = |\langle z_i, A_{-i}^{-1} z_i \rangle|^2 \leq \|z_i\|^2 \cdot (z_i^T A_{-i}^{-2} z_i).$$

We plug this identity into our lower bound, and further multiply by $\frac{\lambda_i}{\lambda_i}$, resulting in

$$\begin{aligned}
V &= \sum_{i=1}^{p} \tilde{\lambda}_i \lambda_i \frac{z_i^T A_{-i}^{-2} z_i}{(1+\lambda_i z_i^T A_{-i}^{-1} z_i)^2} \\
&= \sum_{i=1}^{p} (\frac{\tilde{\lambda}_i}{\lambda_i}) \frac{\lambda_i^2 z_i^T A_{-i}^{-2} z_i}{(1+\lambda_i z_i^T A_{-i}^{-1} z_i)^2} \\
&\geq \sum_{i=1}^{p} (\frac{\tilde{\lambda}_i}{\lambda_i}) \frac{\lambda_i^2 (z_i^T A_{-i}^{-1} z_i)^2}{||z_i||^2 (1+\lambda_i z_i^T A_{-i}^{-1} z_i)^2} \\
&= \sum_{i=1}^{p} (\frac{\tilde{\lambda}_i}{\lambda_i}) \frac{1}{||z_i||^2 (1+\lambda_i z_i^T A_{-i}^{-1} z_i)^2 (\lambda_i z_i^T A_{-i}^{-1} z_i)^{-2}} \\
&= \sum_{i=1}^{p} (\frac{\tilde{\lambda}_i}{\lambda_i}) \frac{1}{||z_i||^2 (1+(\lambda_i z_i^T A_{-i}^{-1} z_i)^{-1})^2}.
\end{aligned}$$

Then, let $k \in (0, n)$ and $\mathscr{L}_i$ be the span of the bottom $n-k$ eigenvectors of A_{-i} and $\Pi_{\mathscr{L}_i}$ be the projection onto the orthogonal complement of $\mathscr{L}_i$. We have that

$$\begin{aligned}
z_i^T A_{-i}^{-1} z_i &\geq (\Pi_{\mathscr{L}_i} z_i)^T A_{-i}^{-1} (\Pi_{\mathscr{L}_i} z_i) \\
&\geq \|\Pi_{\mathscr{L}_i} z_i\|^2 \mu_{k+1}(A_{-i})^{-1}.
\end{aligned}$$

From Lemma B.1, there is a constant $c_1 \geq 1$, such that for any $k \geq 0$, with probability $1 - 2e^{-n/c_1}$, $\mu_{k+1}(A_{-i}) \leq c_1(\sum_{j>k} \lambda_j + \lambda_{k+1} n)$. Additionally, by Corollary B.5, let $t < n/c_3$ and $k \in (0, n/c)$ for $c > c_3$ and c_3 sufficiently large. Then, with probability at least $1 - 3e^{-t}$

$$\|\Pi_{\mathscr{L}_i} z_i\|^2 \geq n/c_4$$

where c_4 only depends on c_3, σ_x and the universal constant given in Corollary B.4.

Then, for $c \geq \max\{c_1, c_3\}$, with probability $1 - 5e^{-n/c}$,

$$\begin{aligned}
z_i^T A_{-i}^{-1} z_i &\geq \|\Pi_{\mathscr{L}_i} z_i\|^2 \mu_{k+1}(A_{-i})^{-1} \\
&\geq \frac{n}{c_4(\sum_{j>k} \lambda_j + \lambda_{k+1} n)}.
\end{aligned}$$

By again applying Corollary B.5 on the same event we have

$$\|z_i\|^2 \leq c_5 n.$$

where c_5 has the same dependencies as c_4.

Altogether, we have for each i, with probability $1 - 5e^{-n/c}$,

$$\begin{aligned}\frac{1}{||z_i||^2(1+(\lambda_i z_i^T A_{-i}^{-1} z_i)^{-1})^2} &\geq \frac{1}{c_5 n(1+(\frac{c_4(\sum_{j>k}\lambda_j+\lambda_{k+1}n)}{\lambda_i n}))^2} \\ &= \frac{1}{c_5 n(1+\frac{c_4\lambda_{k+1}}{\lambda_i}(\frac{\sum_{j>k}\lambda_j}{\lambda_{k+1}n}+1))^2} \\ &= \frac{1}{c_5 c_4^2 n(1/c_4+\frac{\lambda_{k+1}}{\lambda_i}(\rho_k+1))^2} \\ &\geq \frac{1}{c_6 n(1+\frac{\lambda_{k+1}}{\lambda_i}(\rho_k+1))^2}\end{aligned}$$

where $c_6 = c_5 c_4^2$ and $c > \max\{c_1, c_3\}$ as defined above.

Finally, we invoke Lemma B.6 and that $1/(a+b)^2 \geq \min(a^{-2}, b^{-2})/4$ to get that, with probability $1 - 10e^{-n/c}$,

$$V \geq \frac{1}{8c_6 n}\sum_{i=1}^{p}\frac{\tilde{\lambda}_i}{\lambda_i}\min(1, \frac{\lambda_i^2}{\lambda_{k+1}^2(\rho_k+1)^2}).$$

For $c_7 \geq \max\{8c_6, c\}$ we have that with probability $1 - 10e^{-n/c_7}$,

$$V \geq \frac{1}{c_7 n}\sum_{i=1}^{p}\frac{\tilde{\lambda}_i}{\lambda_i}\min(1, \frac{\lambda_i^2}{\lambda_{k+1}^2(\rho_k+1)^2}).$$

F Proof of bias bounds

Theorem 3.2. *(Upper and lower bounds for the bias term) For the lower bound only, assume that random models $\overline{\theta}$ are obtained from the underlying θ_{s}^* as $(\overline{\theta})_i = \gamma_i(\theta_{\mathsf{s}}^*)_i$, where each γ_i is an independent Rademacher random variable. There exists a universal constant $b > 1$, constants c, C that depend only on b and σ_x, and $k < n/C$ such that if $\rho_k \geq b$, then with probability at least $1 - 10e^{-n/c}$,*

$$\mathop{\mathbb{E}}_{\bar{\theta}}[B_2] \geq \frac{1}{c}\left(\sum_{i=1}^{k}\frac{\tilde{\lambda}_i}{\lambda_i}\frac{\lambda_i(\theta_{\mathsf{s}}^*)_i^2}{(1+\frac{\lambda_i}{\lambda_{k+1}\rho_k})^2} + \sum_{i>k}\tilde{\lambda}_i(\theta_{\mathsf{s}}^*)_i^2\right) := \underline{B_2}.$$

If we assume that p is at most exponential in n, then with probability $1 - 5e^{-n/c}$,

$$B_2/c \leq \|\theta_{\mathsf{s}}^*\|^2\sum_{i=1}^{p}\frac{\tilde{\lambda}_i}{\lambda_i}\frac{\lambda_i}{\left(1+\frac{\lambda_i}{\lambda_{k+1}\rho_k}\right)} := \overline{B_2}.$$

F.1 Upper bound

Proof. As defined in Eqn. 4,

$$\begin{aligned}B_2 &= \|\theta_{\mathsf{s}}^* - \widehat{\theta}(X\theta_{\mathsf{s}}^*)\|_{\Sigma_{\mathsf{t}}}^2 \\ &= \|\theta_{\mathsf{s}}^* - X^T A^{-1} X\theta_{\mathsf{s}}^*\|_{\Sigma_{\mathsf{t}}}^2 \\ &= \theta_{\mathsf{s}}^{*T}(I - X^T A^{-1} X)\Sigma_{\mathsf{t}}(I - X^T A^{-1} X)\theta_{\mathsf{s}}^*. \end{aligned} \tag{21}$$

The i^{th} row of $I_p - X^T A^{-1} X$ is given by $e_i - \sqrt{\lambda_i} z_i^T A^{-1} X$. It follows that

$$
\begin{aligned}
(\theta^*)^T M \theta^* &= \begin{pmatrix} \vdots \\ \sum_{j=1}^p \theta_j (e_i[j] - \sqrt{\lambda_i \lambda_j} z_i^\top A^{-1} z_j) \\ \vdots \end{pmatrix}^\top \Sigma_{\mathsf{t}} \begin{pmatrix} \vdots \\ \sum_{j=1}^p \theta_j (e_i[j] - \sqrt{\lambda_i \lambda_j} z_i^\top A^{-1} z_j) \\ \vdots \end{pmatrix} \quad i^{th} \text{ row shown} \\
&= \sum_{i=1}^p \tilde{\lambda}_i \Big(\sum_{j=1}^p \theta_j (e_i[j] - \sqrt{\lambda_i \lambda_j} z_i^\top A^{-1} z_j) \Big)^2 \\
&\leq \sum_{i=1}^p \tilde{\lambda}_i \Big(\sum_{j=1}^p \theta_j^2 \Big) \sum_{j=1}^p \Big(e_i[j] - \sqrt{\lambda_i \lambda_j} z_i^\top A^{-1} z_j \Big)^2 \\
&= \|\theta^*\|^2 \sum_{i=1}^p \tilde{\lambda}_i \sum_{j=1}^p \Big(e_i[j] - \sqrt{\lambda_i \lambda_j} z_i^\top A^{-1} z_j \Big)^2 .
\end{aligned}
$$

Next we look at i^{th} term in the outer sum.

$$
\begin{aligned}
\tilde{\lambda}_i \sum_{j=1}^p (e_i[j] - \sqrt{\lambda_i \lambda_j} z_i^\top A^{-1} z_j)^2 &= \tilde{\lambda}_i (1 - \lambda_i z_i^\top A^{-1} z_i)^2 + \tilde{\lambda}_i \sum_{j \neq i} \lambda_i \lambda_j (z_i^\top A^{-1} z_j)^2 \\
&= \tilde{\lambda}_i (1 - 2\lambda_i z_i^T A^{-1} z_i + \lambda_i^2 (z_i^T A^{-1} z_i)^2 + \sum_{j \neq i} \lambda_i \lambda_j (z_i^\top A^{-1} z_j)^2) \\
&= \tilde{\lambda}_i (1 - 2\lambda_i z_i^T A^{-1} z_i + \sum_{i=1}^p \lambda_i \lambda_j (z_i^\top A^{-1} z_j)^2) \\
&= \tilde{\lambda}_i (1 - 2\lambda_i z_i^T A^{-1} z_i + \lambda_i z_i^\top A^{-1} \Big(\sum_{i=1}^p \lambda_j z_j z_j^T \Big) A^{-1} z_i) \\
&= \tilde{\lambda}_i \left(1 - 2\lambda_i z_i^T A^{-1} z_i + \lambda_i z_i^\top A^{-1} A A^{-1} z_i \right) \\
&= \tilde{\lambda}_i \left(1 - 2\lambda_i z_i^T A^{-1} z_i + \lambda_i z_i^\top A^{-1} z_i \right) \\
&= \tilde{\lambda}_i \left(1 - \lambda_i z_i^T A^{-1} z_i \right) .
\end{aligned}
$$

Using the Sherman-Morrison formula, we get that

$$
\begin{aligned}
1 - \lambda_i z_i^T A^{-1} z_i &= 1 - \lambda_i z_i^T \left(A_{-i} + \lambda_i z_i z_i^T \right)^{-1} z_i \\
&= 1 - \lambda_i z_i^T \left(A_{-i}^{-1} - \lambda_i A_{-i}^{-1} z_i (1 + \lambda_i z_i^T A_{-i}^{-1} z_i)^{-1} z_i^T A_{-i}^{-1} \right) z_i \\
&= 1 - \lambda_i z_i^T A_{-i}^{-1} z_i + \frac{(\lambda_i z_i^T A_{-i}^{-1} z_i)^2}{1 + \lambda_i z_i^T A_{-i}^{-1} z_i} \\
&= \frac{1}{1 + \lambda_i z_i^T A_{-i}^{-1} z_i} .
\end{aligned}
$$

We now provide an upper bound for the remaining term. Let $\Pi_{\mathcal{L}_i}$ be the orthogonal projection onto the bottom $n - k$ eigenvectors of A_{-i}. By Lemma B.1, there exist constants $b, c_0 \geq 1$ such that if $\rho_k \geq b$, then

with probability at least $1 - 2e^{-n/c_0}$,

$$\mu_{k+1}(A_{-i}) \leq c_0 \lambda_{k+1} \rho_k n,$$

so we get

$$\begin{aligned} 1 + \lambda_i z_i^T A_{-i}^{-1} z_i &\geq 1 + \lambda_i (\Pi_{\mathcal{L}_i} z_i)^T A_{-i}^{-1} (\Pi_{\mathcal{L}_i} z_i) \\ &\geq 1 + \frac{\lambda_i \left\| \Pi_{\mathcal{L}_i} z_i \right\|^2}{c_0 \lambda_{k+1} n \rho_k}. \end{aligned}$$

By Corollary B.5, there exist constants c_1 and c_2 with $c_2 > c_1$ and c_1 sufficiently large such that for $0 < k < n/c_2$, we have with probability at least $1 - 3e^{-n/c_1}$,

$$\left\| \Pi_{\mathcal{L}_i} z_i \right\|^2 \geq n/c_3,$$

where c_3 depends only on c_1 and σ.

Plugging these in gives us with probability at least $1 - 5e^{-n/c_4}$,

$$\begin{aligned} \tilde{\lambda}_i \left(1 - \lambda_i z_i^T A^{-1} z_i\right) &\leq \frac{\tilde{\lambda}_i}{\left(1 + \frac{c_5^2 \lambda_i}{\lambda_{k+1} \rho_k}\right)} \\ &= \frac{\tilde{\lambda}_i}{\lambda_i} \frac{\lambda_i}{\left(1 + \frac{c_5^2 \lambda_i}{\lambda_{k+1} \rho_k}\right)}, \end{aligned}$$

where $c_4 = \max(c_0, c_1)$ and $c_5 = \min(c_0, c_3)$.

Therefore by union bound over the application of Corollary B.5,

$$\begin{aligned} B &\leq \|\theta^*\|^2 \sum_{i=1}^{p} \frac{\tilde{\lambda}_i}{\lambda_i} \frac{\lambda_i}{\left(1 + \frac{c_5^2 \lambda_i}{\lambda_{k+1} \rho_k}\right)} \\ &\leq \frac{1}{c_6} \|\theta^*\|^2 \sum_{i=1}^{p} \frac{\tilde{\lambda}_i}{\lambda_i} \frac{\lambda_i}{\left(1 + \frac{\lambda_i}{\lambda_{k+1} \rho_k}\right)}, \end{aligned}$$

where $c_6 = \min(c_5^2, 1)$. Taking $c = \max(c_6^{-1}, c_4)$ gives us the result.

□

F.2 Lower bound

After isolating the contribution of $\frac{\tilde{\lambda}_i}{\lambda_i}$, many of the components of this proof are as given in Tsigler and Bartlet [TB23]. For completeness, we replicate them here.

Proof. Assume that the vector θ_{s}^* is randomly distributed according to the PriorSigns($\overline{\theta}_{\mathsf{s}}$) assumption. Using

Lemma B.8, the bias term can be rewritten as

$$\begin{aligned}
B &= \mathop{\mathbb{E}}_{\theta^*_{\mathsf{s}}}[B_{\theta^*_{\mathsf{s}}}] \\
&= \mathop{\mathbb{E}}_{\theta^*_{\mathsf{s}}}[\|\theta^*_{\mathsf{s}} - \widehat{\theta}(X\theta^*_{\mathsf{s}})\|^2_{\Sigma_{\mathsf{t}}}] \\
&= \mathop{\mathbb{E}}_{\theta^*_{\mathsf{s}}}[(\theta^*_{\mathsf{s}})^T(I_p - X^T(XX^T)^{-1}X)\Sigma_{\mathsf{t}}(I_p - X^T(XX^T)^{-1}X)\theta^*_{\mathsf{s}}] \\
&= \mathop{\mathbb{E}}_{\theta^*_{\mathsf{s}}}[(\theta^*_{\mathsf{s}})^T M\theta^*_{\mathsf{s}}] \\
&= \mathop{\mathbb{E}}_{\theta^*_{\mathsf{s}}}[\theta^*_{\mathsf{s}}]^T M \mathop{\mathbb{E}}_{\theta^*_{\mathsf{s}}}[\theta^*_{\mathsf{s}}] + \mathrm{tr}(M\mathrm{Cov}(\theta^*_{\mathsf{s}})) \\
&= \mathrm{tr}(M\mathrm{Cov}(\theta^*_{\mathsf{s}})).
\end{aligned}$$

where $M = (I_p - X^T(XX^T)^{-1}X)\Sigma_{\mathsf{t}}(I_p - X^T(XX^T)^{-1}X)$. The last equality follows from the assumption $\mathbb{E}_{\theta^*_{\mathsf{s}}}[(\theta^*_{\mathsf{s}})] = 0$. The diagonal elements of $\mathrm{Cov}(\theta^*_{\mathsf{s}})$ are the component-wise variances of θ^*_{s}, which are given by $(\theta^*_{\mathsf{s}})^2_i = (\overline{\theta}_{\mathsf{s}})^2_i$. The off-diagonal elements are 0 since the components of θ^*_{s} are independent. As such, we need only consider the diagonal elements of M.

Note that the i^{th} row of $I_p - X^T(XX^T)^{-1}X$ is equal to $e_i - \sqrt{\lambda_i}z_i^T(XX^T)^{-1}X$, where e_i is the i^{th} vector of the standard orthonormal basis. It follows that the i^{th} diagonal element of M is given by

$$\begin{aligned}
M_{ii} &= \sum_{j=1}^{p} \tilde{\lambda}_j(e_i[j] - \sqrt{\lambda_i\lambda_j}z_i^T A^{-1}z_j)^2 \\
&= \tilde{\lambda}_i(1 - \lambda_i z_i^T A^{-1}z_i)^2 + \sum_{j\neq i} \tilde{\lambda}_j\lambda_i\lambda_j(z_i^T A^{-1}z_j)^2.
\end{aligned}$$

Hence, we can express the bias term as

$$\begin{aligned}
B &= \sum_{i=1}^{p} (\overline{\theta}_{\mathsf{s}})^2_i \left[\tilde{\lambda}_i(1 - \lambda_i z_i^T A^{-1}z_i)^2 + \sum_{j\neq i} \tilde{\lambda}_j\lambda_i\lambda_j(z_i^T A^{-1}z_j)^2\right] \\
&\geq \sum_{i=1}^{p} \frac{\tilde{\lambda}_i}{\lambda_i}\lambda_i(\overline{\theta}_{\mathsf{s}})^2_i(1 - \lambda_i z_i^T A^{-1}z_i)^2.
\end{aligned}$$

We are able to eliminate the second term because it is non-negative. Substituting $A = A_{-i} + \lambda_i z_i z_i^T$ and using the Sherman-Morrison identity, we have that $1 - \lambda_i z_i^T A^{-1}z_i = \frac{1}{1+\lambda_i z_i^T A_{-i}^{-1}z_i}$ (see proof of bias upper bound in Appendix F.1). Then,

$$B \geq \sum_{i=1}^{p} \frac{\tilde{\lambda}_i}{\lambda_i} \frac{\lambda_i(\overline{\theta}_{\mathsf{s}})^2_i}{(1 + \lambda_i z_i^T A_{-i}^{-1}z_i)^2}$$

Let's bound each term in that sum from below with high probability. By Corollary B.3, there exist constants $b, c_0 \geq 1$ such that for any $i \geq 0$ with probability at least $1 - 2e^{-n/c_0}$, if $\rho_k \geq b$, then

$$\mu_n(A_{-i}) \geq \frac{1}{c_0} n\lambda_{k+1}\rho_k.$$

Next,

$$\frac{\lambda_i}{(1+\lambda_i z_i^T A_{-i}^{-1} z_i)^2} \geq \frac{\lambda_i}{(1+\lambda_i \mu_n (A_{-i})^{-1} \|z_i\|^2)^2}.$$

By Corollary B.5, for constants c_1, c_2 such that $k < n/c_2$ with $c_2 > c_1$ for sufficiently large c_1 with probability at least $1 - 3e^{-n/c_1}$ we have $\|z_i\|^2 \leq c_3 n$, where c_3 depends only on c_1 and σ.

We obtain that w.p. at least $1 - 5e^{-n/c_4}$,

$$\frac{\lambda_i \bar{\theta}_i^2}{(1+\lambda_i z_i^T A_{-i}^{-1} z_i)^2} \geq \frac{\lambda_i \bar{\theta}_i^2}{\left(1 + \frac{c_4^2 \lambda_i}{\lambda_{k+1}\rho_k}\right)^2},$$

where $c_4 = \max(c_0, c_1, c_3)$. All the terms are non-negative so Lemma B.6 provides a lower bound on their sum. With probability at least $1 - 10e^{-n/c_4}$,

$$\begin{aligned} B &\geq \frac{1}{2} \sum_{i=1}^{p} \frac{\tilde{\lambda}_i}{\lambda_i} \frac{\lambda_i \bar{\theta}_i^2}{(1 + \frac{c_4^2 \lambda_i}{\lambda_{k+1}\rho_k})^2} \\ &\geq \frac{1}{c_5} \sum_{i=1}^{p} \frac{\tilde{\lambda}_i}{\lambda_i} \frac{\lambda_i \bar{\theta}_i^2}{(1 + \frac{\lambda_i}{\lambda_{k+1}\rho_k})^2}, \end{aligned}$$

where $c_5 = 2\max(c_4^2, 1)$.

Finally, we notice that on $i > k$ we have $\rho_k \geq b > 1$ and $\lambda_i \leq \lambda_{k+1}$ giving us,

$$\begin{aligned} \sum_{i>k} \frac{\tilde{\lambda}_i}{\lambda_i} \frac{\lambda_i \bar{\theta}_i^2}{(1 + \frac{\lambda_i}{\lambda_{k+1}\rho_k})^2} &\geq \sum_{i>k} \frac{\tilde{\lambda}_i}{\lambda_i} \frac{\lambda_i \bar{\theta}_i^2}{(1 + \frac{\lambda_i}{\lambda k+1})^2} \\ &\geq \sum_{i>k} \frac{\tilde{\lambda}_i}{\lambda_i} \frac{\lambda_i \bar{\theta}_i^2}{4} \\ &= \frac{1}{4} \sum_{i>k} \tilde{\lambda}_i \bar{\theta}_i^2. \end{aligned}$$

Letting $c = 4\max(c_4, c_5)$ gives us the result.

□

G Proof of tightness of bounds

Theorem 3.3. *(Tightness of variance and bias bounds) Let the lower bound and upper bound of V be given by $\underline{V}$ and $\overline{V}$, respectively. There exists a universal constant $b \geq 1$, and constant c as defined in Theorem 3.1, and $k \in (0, n/c)$ such that if $\rho_k \geq b$, then*

$$\underline{V}/\overline{V} \in \left[b^{-2}(1+b)^{-2}/c^2, 1\right].$$

Let the lower bound and upper bound of B_2 be given by $\underline{B_2}$ and $\overline{B_2}$, respectively, and the assumptions of Theorem 3.2 be satisfied. Then

$$\underline{B_2}/\overline{B_2} \in \left[\frac{\min_i \left\{(\theta^*_{\mathsf{s}})_i^2 \ : \ (\theta^*_{\mathsf{s}})_i \neq 0\right\}}{\|\theta^*_{\mathsf{s}}\|^2 \left(1 + b^{-1}\frac{\lambda_1}{\lambda_{k+1}}\right)}, 1\right].$$

Proof. We split the proof into the variance proof in Appendix G.1 and the bias proof in Appendix G.2. □

G.1 Variance Proof

Proof. Recall that

$$\underline{V} = \frac{1}{8c_6 n}\sum_{i=1}^{p}\frac{\tilde{\lambda}_i}{\lambda_i}\min\left(1, \frac{\lambda_i^2}{\lambda_{k+1}^2(\rho_k+1)^2}\right)$$
$$\overline{V} = c\left(\sum_{i=1}^{k}\frac{\tilde{\lambda}_i}{\lambda_i}\frac{1}{n} + \sum_{i>k}\frac{\tilde{\lambda}_i}{\lambda_i}\left(\frac{\lambda_i^2}{n\lambda_{k+1}^2\rho_k^2}\right)\right).$$

Since k is the smallest ℓ such that $\rho_\ell \geq b$, it is clear by definition that $\rho_{k-1} < b$. Then we observe that

$$\rho_{k-1} = \frac{1}{n\lambda_k}\sum_{j>k-1}\lambda_j = \frac{\lambda_k + \sum_{j>k}\lambda_j}{n\lambda_k} = \frac{\lambda_k + n\lambda_{k+1}\rho_k}{n\lambda_k} < b$$
$$\therefore \lambda_k + n\lambda_{k+1}\rho_k < nb\lambda_k \Rightarrow \lambda_k > \frac{\lambda_k + n\lambda_{k+1}\rho_k}{nb} > \frac{n\lambda_{k+1}\rho_k}{nb} = \frac{\lambda_{k+1}\rho_k}{b}.$$

On $i \leq k$,

$$\underline{V} : \overline{V} = \frac{1}{8c_6 n}\sum_{i=1}^{k}\frac{\tilde{\lambda}_i}{\lambda_i}\min\left(1, \frac{\lambda_i^2}{\lambda_{k+1}^2(\rho_k+1)^2}\right) : \frac{\tilde{\lambda}_i}{\lambda_i}\frac{c}{n}$$
$$\geq \frac{1}{8c_6 c}\sum_{i=1}^{k}\min\left(1, \frac{\lambda_i^2}{\lambda_{k+1}^2(\rho_k+1)^2}\right) : 1.$$

If the min is 1 then we are okay otherwise, using the identity above and that fact that $\lambda_i \geq \lambda_k$, we have that

$$\frac{\lambda_i^2}{\lambda_{k+1}^2(\rho_k+1)^2} > \frac{(\lambda_{k+1}\rho_k)^2}{b^2\lambda_{k+1}^2(\rho_k+1)^2}$$
$$= \frac{\rho_k^2}{b^2(\rho_k+1)^2}.$$

Examining the ρ_k terms:

$$\frac{\rho_k^2}{(\rho_k+1)^2} = \frac{1}{\rho_k^{-2}(\rho_k+1)^2}$$
$$= \frac{1}{(1+\rho_k^{-1})^2}.$$

As $\rho_k \geq b$ we have that $\rho_k^{-1} \leq b \Rightarrow 1 + \rho_k^{-1} \leq 1 + b \Rightarrow (1 + \rho_k^{-1})^{-2} \geq (1 + b)^{-2}$.

Putting it together we get that

$$\begin{aligned}\frac{1}{8c_6 c}\sum_{i=1}^{k}\frac{\lambda_i^2}{\lambda_{k+1}^2(\rho_k+1)^2} &\geq \frac{1}{8c_6 c}\sum_{i=1}^{k}\frac{1}{b^2(1+b)^2}\\ &= \frac{k}{8c_6 c\cdot b^2(1+b)^2}\\ &\geq \frac{1}{8c_6 c\cdot b^2(1+b)^2}.\end{aligned}$$

On $i > k$, it is clear that the $\min$ is always given by the second term, as $\lambda_i \leq \lambda_{k+1}$, so we get

$$\begin{aligned}\underline{V} : \overline{V} &= \frac{1}{8c_6 n}\sum_{i>k}\frac{\tilde{\lambda}_i}{\lambda_i}\min\left(1, \frac{\lambda_i^2}{\lambda_{k+1}^2(\rho_k+1)^2}\right) : c\frac{\tilde{\lambda}_i}{\lambda_i}\frac{\lambda_i^2}{n\lambda_{k+1}^2\rho_k^2}\\ &= \frac{1}{8c_6 c}\sum_{i>k}\frac{\rho_k^2}{(\rho_k+1)^2}\\ &\geq \frac{1}{8c_6 c}\sum_{i>k}\frac{1}{(1+b)^2} = \frac{1}{8c_6 c}\frac{p-k}{(1+b)^2} > \frac{1}{8c_6 c(1+b)^2}.\end{aligned}$$

Finally we note that for $b \geq 1$ it is clear that $\min(b^{-2}(1+b)^{-2}, (1+b)^{-2}) = b^{-2}(1+b)^{-2}$. Therefore,

$$\underline{V} : \overline{V} \geq \frac{1}{8c_6 c}b^{-2}(1+b)^{-2}.$$

By setting c in the upper bound such that $c > 8c_6$, we get

$$\underline{V} : \overline{V} \geq \frac{1}{c^2}b^{-2}(1+b)^{-2}.$$

□

G.2 Bias proof

Proof. We will bound the ratio of the lower and upper bounds by bounding the ratios of the corresponding terms in each sum. Observe that for all i, the ratio of the terms is equal to

$$\frac{(\theta_i^*)^2}{\|\theta^*\|^2}\cdot\frac{1}{\left(1+\frac{\lambda_i}{\lambda_{k+1}\rho_k}\right)}.$$

On $i \leq k$,

$$\begin{aligned}&\frac{(\theta_i^*)^2}{\|\theta^*\|^2}\cdot\frac{1}{\left(1+\frac{\lambda_i}{\lambda_{k+1}\rho_k}\right)}\\ &\geq \min_i\frac{(\theta_i^*)^2}{\|\theta^*\|^2}\cdot\frac{1}{\left(1+\frac{\lambda_1}{\lambda_{k+1}}b^{-1}\right)}.\end{aligned}$$

On $i > k$, we have $\lambda_i/\lambda_{k+1} \leq 1$, so

$$\frac{(\theta_i^*)^2}{\|\theta^*\|^2} \cdot \frac{1}{\left(1 + \frac{\lambda_i}{\lambda_{k+1}\rho_k}\right)}$$
$$\geq \min_i \frac{(\theta_i^*)^2}{\|\theta^*\|^2} \cdot \frac{1}{(1 + b^{-1})}.$$

Unfortunately, the looseness in the top k components coming from the gap λ_1/λ_{k+1} dominates the tighter ratios in the bottom $p - k$ components which only contain a model-dependent gap, $\min_i \theta_i^2/\|\theta\|^2$. Future work would seek to resolve this and provide tight upper and lower bounds for the bias terms.

□

H Proof of beneficial and malignant shifts

H.1 Trace conditions for simple shifts

Let Σ_{s} be any source covariance and define Σ_{t} as $\tilde{\lambda}_i = \alpha\lambda_i$ for $i \leq k$ and $\tilde{\lambda}_i = \beta\lambda_i$ for $i > k$ with $\alpha, \beta \geq 0$.

Then $\mathrm{tr}(\Sigma_{\mathsf{s}}) = \sum_{i=1}^k \lambda_i + \sum_{i>k} \lambda_i$ and $\mathrm{tr}(\Sigma_{\mathsf{t}}) = \alpha(\sum_{i=1}^k \tilde{\lambda}_i) + \beta(\sum_{i>k} \tilde{\lambda}_i)$.

For $\alpha > 1, \beta < 1$, if

$$\frac{\sum_{i>k} \lambda_i}{\sum_{i=1}^k \lambda_i} < \frac{\alpha - 1}{1 - \beta}$$

then we have that $\mathrm{tr}(\Sigma_{\mathsf{s}}) < \mathrm{tr}(\Sigma_{\mathsf{t}})$ and if the inequality is flipped then we obtain $\mathrm{tr}(\Sigma_{\mathsf{s}}) > \mathrm{tr}(\Sigma_{\mathsf{t}})$.

For $\alpha < 1, \beta > 1$, if

$$\frac{\sum_{i=1}^k \lambda_i}{\sum_{i>k} \lambda_i} < \frac{\beta - 1}{1 - \alpha}$$

then we have that $\mathrm{tr}(\Sigma_{\mathsf{s}}) < \mathrm{tr}(\Sigma_{\mathsf{t}})$ and if the inequality is flipped then we obtain $\mathrm{tr}(\Sigma_{\mathsf{s}}) > \mathrm{tr}(\Sigma_{\mathsf{t}})$.

H.2 Proof of beneficial and malignant shifts for simple shifts

We restate the theorem for ease.

Theorem 3.4. *(Beneficial and Malignant Multiplicative Shifts on Variance) Let Σ_{s} be a source covariance that satisfies benign source conditions. That is, $\exists\, k$ such that $\rho_k \geq b$ for a universal constant $b > 1$. Define Σ_{t} as $\tilde{\lambda}_i = \alpha\lambda_i$ for $i \leq k$ and $\tilde{\lambda}_i = \beta\lambda_i$ for $i > k$, with $\alpha, \beta \geq 0$.*

1. *If $\alpha < 1, \beta \leq 1$ or $\alpha \leq 1, \beta < 1$ then we obtain a beneficial shift in variance.*
2. *If $\alpha > 1, \beta \geq 1$ or $\alpha \geq 1, \beta > 1$ then we obtain a malignant shift in variance.*
3. *If we are in the mildly overparameterized regime:*

- $\alpha > 1$ *and* $\beta < 1$ *leads to beneficial shifts;*
- $\alpha < 1$ *and* $\beta > 1$ *leads to malignant shifts.*

4. *If we are in the severely overparameterized regime:*
   - $\alpha > 1$ *and* $\beta < 1$ *leads to malignant shifts;*
   - $\alpha < 1$ *and* $\beta > 1$ *leads to beneficial shifts.*

Proof. From Theorem 3.1 and Theorem 3.3, we have that for a universal constant $b > 1$ if $\rho_k \geq b$ we get the following upper and lower bounds on the out-of-distribution variance for some constants c_1, c_2,

$$V_{ood} \leq c_1 \left(\frac{1}{n} \sum_{i=1}^{k} \frac{\tilde{\lambda}_i}{\lambda_i} + n \frac{\sum_{i>k} \tilde{\lambda}_i \lambda_i}{(\sum_{i>k} \lambda_i)^2} \right)$$
$$V_{ood} \geq c_2 \left(\frac{1}{n} \sum_{i=1}^{k} \frac{\tilde{\lambda}_i}{\lambda_i} + n \frac{\sum_{i>k} \tilde{\lambda}_i \lambda_i}{(\sum_{i>k} \lambda_i)^2} \right).$$

Analogously, the in-distribution variance is upper and lower bounded by,

$$V_{id} \leq c_1 \left(\frac{k}{n} + \frac{n}{R_k} \right)$$
$$V_{id} \geq c_2 \left(\frac{k}{n} + \frac{n}{R_k} \right)$$

where $R_k = (\sum_{i>k} \lambda_i)^2 / \sum_{i>k} \lambda_i^2$.

Let Σ_{s} be any source covariance model that satisfies benign source conditions. Define Σ_{t} by,

$$\tilde{\lambda}_i = \begin{cases} \alpha \lambda_i, & i \leq k \\ \beta \lambda_i, & i > k \end{cases}$$

for $\alpha, \beta \geq 0$.

Beneficial shifts. We use the upper bound to specify requirements for the beneficial shifts.

$$\begin{aligned} V_{ood} &\leq c_1 \left(\alpha \frac{k}{n} + \beta \frac{n}{R_k} \right) \\ &= V_{id} + c_1 \left(\frac{k}{n}(\alpha - 1) + \frac{n}{R_k}(\beta - 1) \right). \end{aligned}$$

Let $\alpha > 1, \beta < 1$. To obtain a beneficial shift in this setting we need,

$$\begin{aligned} \frac{n}{R_k}(1 - \beta) &> \frac{k}{n}(\alpha - 1) \\ \Rightarrow \frac{n}{R_k} &> \frac{k}{n} \left(\frac{\alpha - 1}{1 - \beta} \right). \end{aligned}$$

In the case $\alpha < 1, \beta > 1$, to obtain a beneficial shift we need,

$$\begin{aligned} \frac{n}{R_k}(\beta - 1) &< \frac{k}{n}(1 - \alpha) \\ \Rightarrow \frac{n}{R_k} &< \frac{k}{n}\left(\frac{1-\alpha}{\beta - 1}\right). \end{aligned}$$

In the case where $\alpha = 1$ then any $\beta < 1$ leads to beneficial shifts. Similarly when $\beta = 1$, any $\alpha < 1$ leads to beneficial shifts.

Malignant shifts. We use the lower bound to specify requirements for the malignant shift.

$$\begin{aligned} V_{ood} &\geq c_2 \left(\alpha \frac{k}{n} + \beta \frac{n}{R_k}\right) \\ &= V_{id} + c_2 \left(\frac{k}{n}(\alpha - 1) + \frac{n}{R_k}(\beta - 1)\right). \end{aligned}$$

Let $\alpha < 1$ and $\beta > 1$. To obtain a malignant shift in this setting we need,

$$\frac{n}{R_k} > \frac{k}{n}\left(\frac{1-\alpha}{\beta - 1}\right).$$

In the case of $\alpha > 1, \beta < 1$, to obtain a malignant shift we need,

$$\frac{n}{R_k} < \frac{k}{n}\left(\frac{\alpha - 1}{1 - \beta}\right).$$

In the case where $\alpha = 1$ then any $\beta > 1$ leads to malignant shifts. Similarly when $\beta = 1$, any $\alpha > 1$ leads to malignant shifts.

Mild and severe overparameterization. We see that the four cases separate into settings in which we are mildly overparameterized, meaning

$$\frac{n}{R_k} > \frac{k}{n}\left|\frac{\alpha - 1}{1 - \beta}\right|,$$

and settings in which we are severely overparamterized, meaning

$$\frac{n}{R_k} < \frac{k}{n}\left|\frac{\alpha - 1}{1 - \beta}\right|.$$

In each of these regimes of overparameterization, the above proof has delineated whether we achieve beneficial or malignant shifts in all settings of α, β. □

H.3 Generalized (necessary) conditions for beneficial and malignant shifts

Let $\Sigma_{\mathbf{s}}$ be any source covariance matrix that satisfies benign source conditions and define $\Sigma_{\mathbf{t}}$ as

$$\tilde{\lambda}_i = \begin{cases} \alpha_i \lambda_i & i \leq k, \\ \beta_i \lambda_i & i > k \end{cases}$$

with $\alpha_i, \beta_i \geq 0$ for all i.

Then the OOD variance upper bound is given by,

$$\begin{aligned} V_{ood} &\leq c_1 \left(\frac{1}{n} \sum_{i=1}^{k} \alpha_i + n \frac{\sum_{i>k} \beta_i \lambda_i^2}{(\sum_{i>k} \lambda_i)^2} \right) \\ &= V_{id} + c_1 \left(\frac{(\sum_{i=1}^{k} \alpha_i) - k}{n} + n \frac{\sum_{i>k} \lambda_i^2 (\beta_i - 1)}{(\sum_{i>k} \lambda_i)^2} \right) \\ &= V_{id} + c_1 \left(\frac{k}{n} \left(\frac{\sum_{i=1}^{k} \alpha_i}{k} - 1 \right) + \frac{n}{R_k} \left(\frac{\sum_{i>k} \beta_i \lambda_i^2}{\sum_{i>k} \lambda_i^2} - 1 \right) \right), \end{aligned}$$

and the OOD variance lower bound is given by,

$$\begin{aligned} V_{ood} &\geq c_2 \left(\frac{1}{n} \sum_{i=1}^{k} \alpha_i + n \frac{\sum_{i>k} \beta_i \lambda_i^2}{(\sum_{i>k} \lambda_i)^2} \right) \\ &= V_{id} + c_2 \left(\frac{(\sum_{i=1}^{k} \alpha_i) - k}{n} + n \frac{\sum_{i>k} \lambda_i^2 (\beta_i - 1)}{(\sum_{i>k} \lambda_i)^2} \right) \\ &= V_{id} + c_2 \left(\frac{k}{n} \left(\frac{\sum_{i=1}^{k} \alpha_i}{k} - 1 \right) + \frac{n}{R_k} \left(\frac{\sum_{i>k} \beta_i \lambda_i^2}{\sum_{i>k} \lambda_i^2} - 1 \right) \right), \end{aligned}$$

where V_{id} is the ID variance bound.

Again, we use the upper bounds to prove conditions for beneficial shifts and the lower bounds to prove conditions for malignant shifts.

Beneficial shifts. From the upper bound we consider two separate cases for non-trivial beneficial shifts:

1. $\sum_{i=1}^{k} \alpha_i < k$ and $\sum_{i>k} \beta_i \lambda_i^2 > \sum_{i>k} \lambda_i^2$,
2. $\sum_{i=1}^{k} \alpha_i > k$ and $\sum_{i>k} \beta_i \lambda_i^2 < \sum_{i>k} \lambda_i^2$.

We start with the case of $\sum_{i=1}^{k} \alpha_i < k$ and $\sum_{i>k} \beta_i \lambda_i^2 > \sum_{i>k} \lambda_i^2$. If this is satisfied, the only way to achieve a beneficial shift is if

$$\frac{n}{R_k} \left(\frac{\sum_{i>k} \beta_i \lambda_i^2}{\sum_{i>k} \lambda_i^2} - 1 \right) < \frac{k}{n} \left(1 - \frac{\sum_{i=1}^{k} \alpha_i}{k} \right). \tag{22}$$

We also have in this setting that,

$$0 < 1 - \frac{\sum_{i=1}^{k} \alpha_i}{k} \leq 1.$$

In Equation 22 we see a notion of severe overparameterization that leads to beneficial shifts. For instance as $R_k \to \infty$ we see the left-hand-side (LHS) of Equation 22 $\to 0$. So as $R_k \to \infty$ we have that finite n always leads to a beneficial shift in this setting. We note that equivalently if $\beta_i = 1$ for all i then we also have the LHS $\to 0$, just as in the case of severe overparameterization. We will return to the definitions of mild and severe overparameterization for arbitrary shifts after showing the remaining conditions for beneficial and malignant shifts.

Now consider the case of $\sum_{i=1}^{k} \alpha_i > k$ and $\sum_{i>k} \beta_i \lambda_i^2 < \sum_{i>k} \lambda_i^2$. If this is satisfied, the only way to achieve a beneficial shift is if

$$\frac{n}{R_k}\left(1 - \frac{\sum_{i>k} \beta_i \lambda_i^2}{\sum_{i>k} \lambda_i^2}\right) > \frac{k}{n}\left(\frac{\sum_{i=1}^{k} \alpha_i}{k} - 1\right). \tag{23}$$

In this setting it is clear that

$$0 < 1 - \frac{\sum_{i>k} \beta_i \lambda_i^2}{\sum_{i>k} \lambda_i^2} \leq 1.$$

In Equation 23, it is clear that we have a notion of mild overparameterization that leads to beneficial shifts. As above if $\alpha_i = 1$ for all i then we always obtain a beneficial shift in this setting. Otherwise if R_k does not grow too quickly (as in the case with mild overparameterization) then this is a necessary condition to achieve beneficial shifts when $\sum_{i>k} \beta_i \lambda_i^2 < \sum_{i>k} \lambda_i^2$.

Malignant shifts. From the lower bound we once again consider two separate cases for non-trivial malignant shifts:

1. $\sum_{i=1}^{k} \alpha_i > k$ and $\sum_{i>k} \beta_i \lambda_i^2 < \sum_{i>k} \lambda_i^2$,
2. $\sum_{i=1}^{k} \alpha_i < k$ and $\sum_{i>k} \beta_i \lambda_i^2 > \sum_{i>k} \lambda_i^2$.

We start with the case of $\sum_{i=1}^{k} \alpha_i > k$ and $\sum_{i>k} \beta_i \lambda_i^2 < \sum_{i>k} \lambda_i^2$. If this is satisfied then the only way to achieve a malignant shift is if,

$$\frac{n}{R_k}\left(1 - \frac{\sum_{i>k} \beta_i \lambda_i^2}{\sum_{i>k} \lambda_i^2}\right) < \frac{k}{n}\left(\frac{\sum_{i=1}^{k} \alpha_i}{k} - 1\right). \tag{24}$$

In the case of $\sum_{i=1}^{k} \alpha_i < k$ and $\sum_{i>k} \beta_i \lambda_i^2 > \sum_{i>k} \lambda_i^2$ the only way to achieve a malignant shift is if,

$$\frac{n}{R_k}\left(\frac{\sum_{i>k} \beta_i \lambda_i^2}{\sum_{i>k} \lambda_i^2} - 1\right) > \frac{k}{n}\left(1 - \frac{\sum_{i=1}^{k} \alpha_i}{k}\right). \tag{25}$$

We now are ready to define mild and severe overparameterization for arbitrary multiplicative shifts.

Theorem H.1. *(Mild and severe overparameterization for arbitrary multiplicative shifts) Let Σ_{s} be any source covariance matrix that satisfies benign source conditions, meaning $\exists\, k$ such that $\rho_k \geq b$ for a universal constant $b > 1$. Furthermore, let Σ_{t} be defined by $\tilde{\lambda}_i = \alpha_i \lambda_i$ for $i \leq k$ and $\tilde{\lambda}_i = \beta_i \lambda_i$ for $i > k$.*

We will define

$$C := \left| \left(\frac{\sum_{i=1}^{k} \alpha_i}{k} - 1 \right) \left(1 - \frac{\sum_{i>k} \beta_i \lambda_i^2}{\sum_{i>k} \lambda_i^2} \right)^{-1} \right|.$$

Then we are mildly overparameterized if

$$\frac{n}{R_k} = \omega\left(C\frac{k}{n}\right)$$

and we are severely overparameterized if

$$\frac{n}{R_k} = o\left(C\frac{k}{n}\right).$$

We now state our taxonomy of covariate shifts for arbitrary multiplicative shifts.

Theorem H.2. *(Beneficial and Malignant (Arbitrary) Multiplicative Shifts on Variance) Let Σ_{s} be any source covariance matrix that satisfies benign source conditions, meaning $\exists\ k$ such that $\rho_k \geq b$ for a universal constant $b > 1$. Furthermore, let Σ_{t} be defined by $\tilde{\lambda}_i = \alpha_i \lambda_i$ for $i \leq k$ and $\tilde{\lambda}_i = \beta_i \lambda_i$ for $i > k$.*

1. *If $\sum_{i=1}^{k} \alpha_i \leq k$ and $\sum_{i>k} \beta_i \lambda_i^2 < \sum_{i>k} \lambda_i^2$ then we obtain a beneficial shift.*
2. *If $\sum_{i=1}^{k} \alpha_i < k$ and $\sum_{i>k} \beta_i \lambda_i^2 \leq \sum_{i>k} \lambda_i^2$ then we obtain a beneficial shift.*
3. *If $\sum_{i=1}^{k} \alpha_i \geq k$ and $\sum_{i>k} \beta_i \lambda_i^2 > \sum_{i>k} \lambda_i^2$ then we obtain a malignant shift.*
4. *If $\sum_{i=1}^{k} \alpha_i > k$ and $\sum_{i>k} \beta_i \lambda_i^2 \geq \sum_{i>k} \lambda_i^2$ then we obtain a malignant shift.*
5. *If we are in the mildly overparameterized regime:*
   - *$\sum_{i=1}^{k} \alpha_i > k$ and $\sum_{i>k} \beta_i \lambda_i^2 < \sum_{i>k} \lambda_i^2$ leads to beneficial shifts,*
   - *$\sum_{i=1}^{k} \alpha_i < k$ and $\sum_{i>k} \beta_i \lambda_i^2 > \sum_{i>k} \lambda_i^2$ leads to malignant shifts.*
6. *If we are in the severely overparameterized regime:*
   - *$\sum_{i=1}^{k} \alpha_i < k$ and $\sum_{i>k} \beta_i \lambda_i^2 > \sum_{i>k} \lambda_i^2$ leads to beneficial shifts,*
   - *$\sum_{i=1}^{k} \alpha_i > k$ and $\sum_{i>k} \beta_i \lambda_i^2 < \sum_{i>k} \lambda_i^2$ leads to malignant shifts.*

I Experiment details

I.1 Synthetic data experiments

Our synthetic data experiments use source data generated from random Gaussians with covariance structures that are known to exhibit benign overfitting. These structures include the (k, δ, ϵ) spiked covariance models and eigendecay rates given by Bartlett et al. [Bar+20] such as $\lambda_i = i^{-\alpha} \ln^{-\beta}(i+1)$ for $\alpha = 1, \beta > 1$. Target

data is generated from random Gaussians with covariances that lead to beneficial and malignant shifts based on our theories and modifications of the aforementioned source covariance structures.

All ground truth models are sampled uniformly on the p-dimensional hypersphere, as $\theta_{\mathsf{s}}^* \sim \mathcal{S}^{p-1}$. Label noise is sampled as $\varepsilon \sim \mathcal{N}(0,1)$, unless otherwise specified. For a data matrix $X \in \mathbb{R}^{n \times p}$, training labels are obtained as $y = X\theta_{\mathsf{s}}^* + \varepsilon$. Excess risk is computed for unseen testing data from source and target distributions of interest using clean labels.

In Figure 4 we take the source to be the (k, δ, ϵ) spiked model with parameters given by $k = 70$, $\delta = 1$, and $\epsilon = 0.005$. The beneficial shift scales the first k eigenvalues by $\alpha = 1.125$ and the last $p - k$ eigenvalues by $\beta = 0.65$. For the malignant shift we use $\alpha = 0.875$ and $\beta = 1.35$. The minimum-norm linear interpolator is fit to 500 data points sampled from a centered multivariate Gaussian with unit variance and dimension $p = 4900$. The model vector is sampled from a centered Gaussian and scaled to unit norm. The x-axis represents the amount of additive label noise in training. All evaluation is done on clean data. Each point is the average of 40 runs.

In Figure 5, we take the source to be the (k, δ, ϵ) spiked model with source parameters as $k = 10, \delta = 1.0, \epsilon = 1e^{-6}$ and target parameters $\tilde{k} = 10, \tilde{\delta} = 1.35, \tilde{\epsilon} = 6.5e^{-7}$. We use $n = 50$ training data points, $10k$ held-out testing data points in each OOD test set, and vary p from 75 to 1000 dimensions. We solve OLS using the closed-form MNI solution on the source data. Each experiment is averaged over 100 independent runs.

In Figure 2 we train fully-connected neural networks with ReLU activation functions. Data is sampled as above from the covariance structures given by $\lambda_i = i^{-\alpha} \ln^{-\beta}(i+1)$ with varying β to obtain beneficial and malignant shifts. The network architecture is 3 hidden layers, with hidden widths 512 and 2048. Networks are trained with stochastic gradient descent with momentum 0.9 until the training MSE has reached $< 5e^{-6}$. We start with a learning rate of 0.01 and decay by a stepped cosine schedule for 1,500 epochs. We take batch size of 64 and train without weight decay. Each experiment is averaged over 20 independent runs. We train in PyTorch with a single A100 NVIDIA GPU. In these experiments we take $n = 200$ and compare $p = 20$ with $p = 2000$. Label noise is sampled as $\mathcal{N}(0, \sigma^2)$ and we vary σ^2 to show the behaviors at varying train label noise.

In Figure 8 we train full-connected neural networks with ReLU activation functions. Source data is sampled from a mean-centered Gaussian with diagonal covariance matrix with eigenvalues $\lambda_i = i^{-1} \ln^{-1.5}(i+1)$. Target covariate shifts are implemented in the style of Theorem 3.4 where the top k source eigenvalues are multiplied by α and the bottom $p - k$ source eigenvalues are multiplied by β. In this experiment, we take $k = 10, \alpha = 2, \beta = 0.1$ and experiment with $n = 400$ source data samples for $p = 200$ and $p = 4,000$. The network architecture is 3 hidden layers with hidden width $2,048$. Our training setup is the same as given above for prior MLP experiments.

I.2 CIFAR-10 and CIFAR-10C experiments

In Figures 3 and 9 we use a binary variant of CIFAR-10 and CIFAR-10C. For details on the CIFAR-10C dataset, see Hendrycks and Dietterich [HD19]. The binary problem is constructed by selecting only the dog and truck classes. To stay overparameterized, we subsample $n = 500, 1000, 2000$ points in a class-balanced manner. Images are flattened into $p = 3072$ dimensional vectors. We fit our model using the OLS solution for the MNI against $\{0, 1\}$ class labels. We test on the same two classes from CIFAR-10 and CIFAR-10C

Gaussian blur and Gaussian noise corruptions. Recall that these two corruptions were selected for their eigenspectra's similarity to beneficial and malignant shifts, respectively. Label noise is injected by flipping class labels with a given probability.

In Figure 10, we train ResNet18 models on the entire CIFAR-10 dataset and evaluate on the CIFAR-10C test sets for the Gaussian blur and Gaussian noise corruptions. The setting is not high-dimensional because we train on 50000 images with 3072 dimensions. However, the ResNet18 architecture has around 11.7 million parameters, so the level of overparameterization is very high. The training procedure is similar to that used for our MLP experiments. Networks are trained with stochastic gradient descent with a learning rate of 0.1 and stepped cosine decay schedule for 60 epochs. Each point in the plot is an average over 30 independent runs. As before, we train in PyTorch with a single A100 NVIDIA GPU. Label noise takes the form of random label flips with probabilities 0.1 to 0.9.

J Additional experiments

We present a number of additional supporting experiments that show: (1) more cases of the behavior of the MNI and MLPs under covariate shift on synthetic datasets; (2) underparameterized and overparameterized regimes for linear regression under covariate shift for more realistic eigendecay rates outside of (k, δ, ϵ) spiked covariance models; (3) cases in which the MNI is overfit in a *tempered* or *catastrophic* manner and evaluated on OOD datasets constructed based on our results in Theorem 3.4, indicating that our insights hold up for the MNI even when benign source conditions are not satisfied; (4) the value of overparameterization for the MNI trained on CIFAR-10 and evaluated on CIFAR-10C; (5) experiments training ResNet-18 models to interpolation on the full CIFAR-10 dataset and evaluated on CIFAR-10C blur and noise corruptions.

J.1 MNI on Synthetic Data

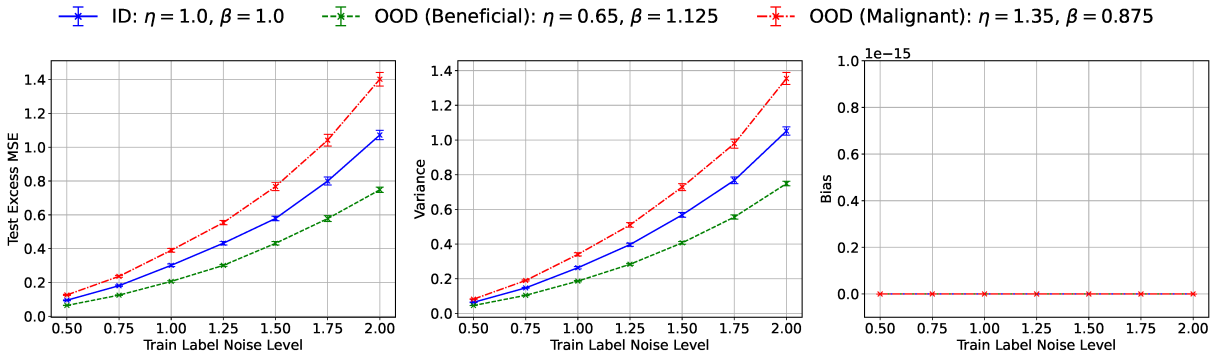


Figure 4: We fit interpolating linear models to random Gaussian data sampled from spiked covariance models with parameters k, δ, ϵ. In this setting, $k = 70, n = 500$, $p = 4900$, $\delta = 1$ and $\epsilon = 0.005$. To illustrate a beneficial shift, we scale the first k eigenvalues by $\alpha = 1.125$ and the last $p - k$ eigenvalues by $\beta = 0.65$. Similarly, for the malignant shift we use $\alpha = 0.875$ and $\beta = 1.35$. All experiments are averaged over 25 independent runs with standard error bars displayed. Note that the bias is consistently below 10^{-16}.

In Figure 4, we experiment with interpolating linear models where Σ_s, Σ_t are given by (k, δ, ϵ)-spike covariances with $k = 70$, $n = 500$, and $p = 4900$. We design problem parameters to show settings in which $\mathrm{tr}(\Sigma_t) > \mathrm{tr}(\Sigma_s)$ and we get a beneficial shift, and $\mathrm{tr}(\Sigma_t) < \mathrm{tr}(\Sigma_s)$ and we get a malignant shift. To do this,

the source covariance matrix is constructed using $\delta = 1$ and $\epsilon = 0.005$. To illustrate a beneficial shift, we scale the first k eigenvalues by $\alpha = 1.125$ and the last $p - k$ eigenvalues by $\beta = 0.65$. Similarly, for the malignant shift we use $\alpha = 0.875$ and $\beta = 1.35$. The resulting plots are significant because they highlight the distinct effects that the first k and last $p - k$ components have on the excess risk.

As illustrated by our main theorems, increasing the energy of an eigenvalue has a negative impact on the risk. Nonetheless, these plots show that where the increase happens plays an important role on how the shift affects generalization. We are able to improve performance by decreasing the energy on the tail and increasing the energy on the head in such a way that the total energy is increased. In short, this setting is a direct connection to our theory and shows clearly that our constructions for beneficial and malignant shifts, when mildly overparameterized, hold up in low and high train label noise regimes, with higher noise exacerbating the effects of the shifts. In addition, Figure 4 demonstrates that the variance generally contributes much more significantly to the overall risk.

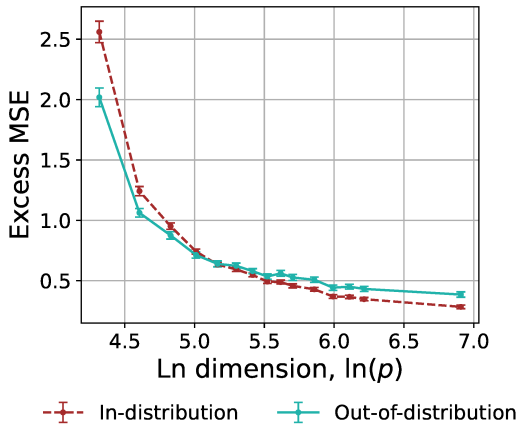
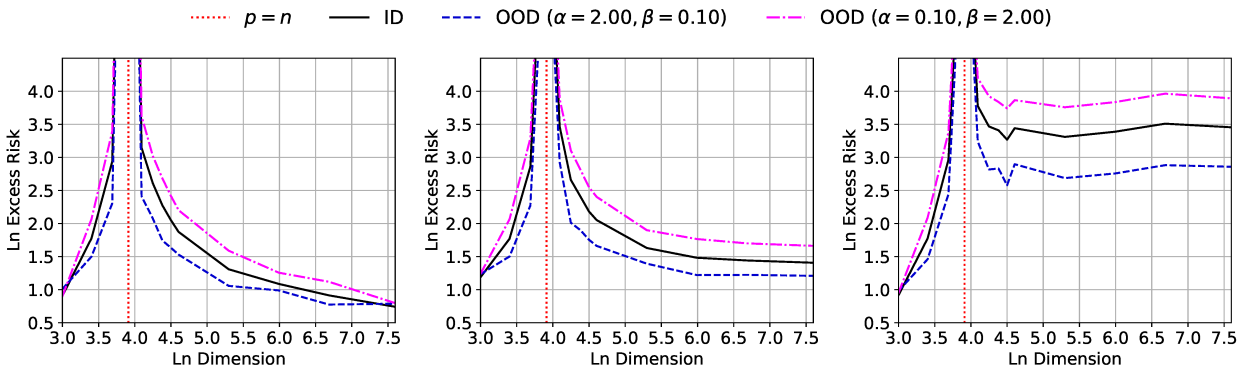


Figure 5: We experiment with the (k, δ, ϵ) spiked covariance models and examine conditions for beneficial and malignant shifts as given in Theorem 3.4. We take $n = 50, k = 10, \delta = 1.0, \epsilon = 1e^{-6}, \tilde{\delta} = 1.5, \tilde{\epsilon} = 5e^{-7}$, and vary p. In all cases, $\mathrm{tr}(\Sigma_t) > \mathrm{tr}(\Sigma_s)$, showing that beneficial shifts of this form can occur. As we increase p while keeping other problem parameters fixed we observe the transition from mild to severe overparameterization and see the cross-over point between the shift going from beneficial to malignant. For both ID and OOD excess risk, we observe that excess risk is a decreasing function of input dimension. Curves are averaged over 100 independent runs.

Figure 5 shows another example of the transition from mild overparameterization to severe overparameterization in the case of (k, δ, ϵ) spiked covariance models. In this example we take $k = 10, n = 50, \delta = 1.0, \epsilon = 1e^{-6}$. Using our shifts defined in Theorem 3.4 we set $\alpha = 1.5$ and $\beta = 0.5$. We plot excess MSE on both ID and OOD test sets vs. the input data dimension, while holding all other problem parameters fixed and clearly observe the transition from beneficial to malignant shifts in keeping with our theorem.

Next, we experimentally show that while our theory is built for benign source covariance structures it holds for non-benign covariances. In particular, we examine eigendecay rates that are known to lead to *tempered* overfitting and *catastrophic* overfitting [Mal+22]. Bartlett et al. [Bar+20] identify the covariance structure given by $\lambda_i = i^{-1}\ln^{-2}(i+1)$ as sufficient for benign overfitting. The rate of $i^{-\alpha}$ for $\alpha > 1$ is akin to a

ridgeless Laplace kernel and corresponds to tempered overfitting. Finally, the rate of $i^{-\ln(i)}$ is akin to a ridgeless Gaussian kernel and corresponds to catastrophic overfitting. This relative ordering is determined by how high-dimensional the tail eigenvalues are, in decreasing order.

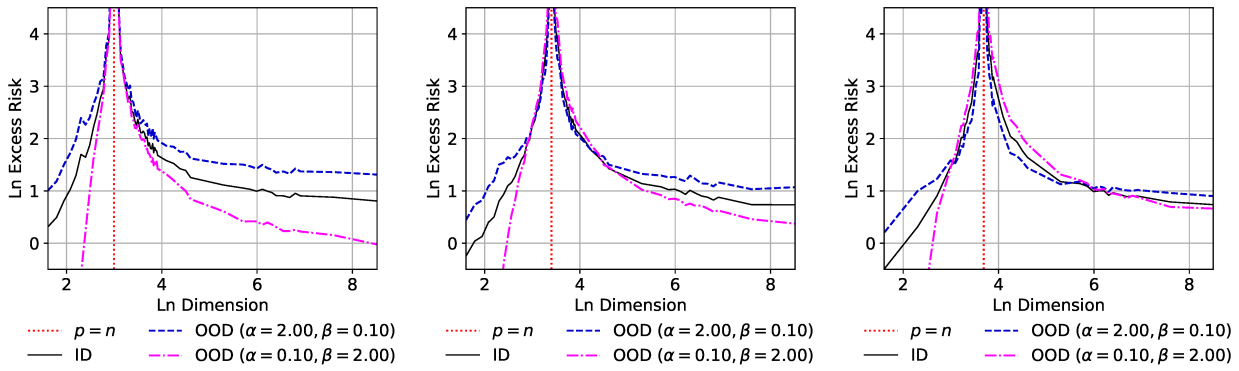


(a) Ridgeless, $\lambda_i = i^{-1}\ln^{-2}(i+1)$ (b) Ridgeless, $\lambda_i = i^{-2}$ (c) Ridgeless, $\lambda_i = i^{-\ln(i)}$

Figure 6: Comparing covariate shift in underparameterized vs. overparameterized linear regression for three different eigendecay rates. In the $p > n$ setting: (a) leads to benign overfitting, (b) leads to tempered overfitting, and (c) leads to catastrophic overfitting. We implement simple multiplicative shifts with α, β as defined in Section 3.1 where we take $n = 50, k = 10$. Ground truth models are sampled uniformly from $\mathcal{S}^{p-1}$ and training label noise is sampled from $\mathcal{N}(0, 2)$. Every curve is averaged over 50 independent runs.

It is clear that even though Theorem 3.4 is for the case in which Σ_s satisfies benign source conditions, the style of beneficial and malignant shift we identify holds for the MNI even when overfit in a non-benign manner. That is, when Σ_s has eigendecay rates that are *tempered* or *catastrophic* we can still obtain non-trivial beneficial and malignant shifts by changing the energy on the signal and noise components in a heterogeneous way.

We also notice in Figure 6 that even when varying the dimension up to $p = 2000$ at $n = 50, k = 10$ we do not quite observe the cross-over from beneficial to malignant shifts in the overparameterized regime. However, we observe that in Figure 6(a) that the two OOD curves begin to cross-over. Given compute budget, we run a variant of Figure 6 where we extend up to $p = 5000$ and take smaller n, e.g. $n = 20, 30, 40$, in order to closer examine the different regimes of overfitting. In addition, we experimentally show results for $p = 5, 10$ which we liken to the classical linear regression regime in which $k = p < n$, meaning all of the signal is captured in the p components. In this setting, α shifts are all that influence the distribution shift behavior. We show these behaviors in Figure 7.

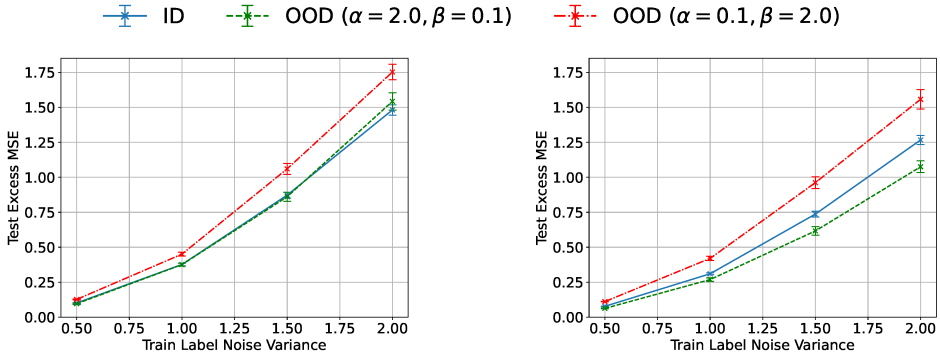


(a) $n = 20$ (b) $n = 30$ (c) $n = 40$

Figure 7: Comparing covariate shift in underparameterized vs. overparameterized linear regression for when $\lambda_i = i^{-1} \ln^{-2}(i+1)$. We implement simple multiplicative shifts with α, β as defined in Section 3.1 where we take $k = 10$ and vary n. Ground truth models are sampled uniformly from $\mathcal{S}^{p-1}$ and training label noise is sampled from $\mathcal{N}(0, 2)$. Every curve is averaged over 100 independent runs.

J.2 MLP on Synthetic Data

We now show additional results for MLPs trained to interpolation on synthetic datasets. This experiment is analogous to that of Figure 2 except that we implement shifts in the style of Theorem 3.4.



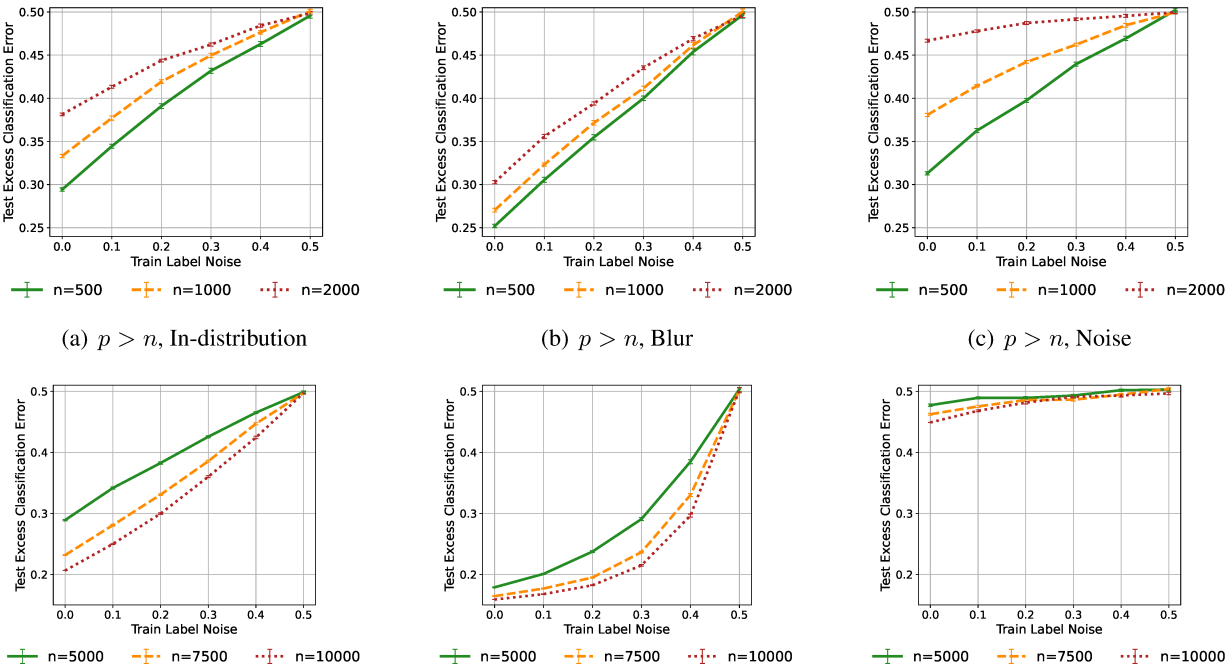
(a) $n = 400, p = 200, h = 2048$ (b) $n = 400, p = 4k, h = 2048$

Figure 8: We implement multiplicative shifts for interpolating 3-layer ReLU MLPs in the style of Theorem 3.4. Source data, X, is sampled from a mean-centered Gaussian with diagonal covariance given by $\lambda_i = i^{-1} \ln^{-1.5}(i+1)$. Ground truth models are sampled as $\theta_s^* \sim \mathcal{S}^{p-1}$ and training label noise is samples as $\varepsilon = \mathcal{N}(0, \sigma_x^2)$. Noisy training labels are obtained as $y = X\theta_s^* + \varepsilon$. The target covariances are obtained by multiplying the top $k = 10$ source eigenvalues by α and the bottom $p - k$ source eigenvalues by β. From our theory, we expect that $\alpha = 2, \beta = 0.1$ leads to beneficial shifts while $\alpha = 0.1, \beta = 2$ leads to malignant shifts. We see this holds up when $p > n$, and that $h > n$ does not change this relationship. All curves are averaged over 20 independent runs and each training run reaches MSE loss $\leq 5e^{-6}$.

In Figure 2 we sampled the ID dataset from a mean-centered Gaussian with diagonal covariance that has

eigenvalues $\lambda_i = i^{-1}\ln^{-3}(i+1)$ and we examined the behavior for OOD datasets under covariate shift where the eigenvalues of the OOD covariance are given by $\lambda_i = i^{-1}\ln^{-2}(i+1)$ and $\lambda_i = i^{-1}\ln^{-4}(i+1)$. In Figure 8 we take the ID data to be sampled from a mean-centered Gaussian with diagonal covariance that has eigenvalues $\lambda_i = i^{-1}\ln^{-1.5}(i+1)$. For the covariance of the OOD datasets, we shift the top $k = 10$ eigenvalues by a factor of α and the bottom $p-k$ eigenvalues by a factor of β, as in the setting of Theorem 3.4. We experiment here with $\alpha = 2, \beta = 0.1$ and $\alpha = 0.1, \beta = 2$. Each model achieves training MSE $\leq 5e^{-6}$. We see the same trends as in Figure 2 with respect to $p > n$ versus $h > n$. In the $p < n$ case, even though $h > n$ we do not clearly observe a beneficial shift as predicted by our high-dimensional linear theory. However, when $p > n$ we do observe beneficial shifts for $\alpha = 2, \beta = 0.1$, as suggested by our theorem for the mildly overparameterized case.

J.3 MNI on CIFAR-10C Experiments



(a) $p > n$, In-distribution (b) $p > n$, Blur (c) $p > n$, Noise

(d) $p < n$, In-distribution (e) $p < n$, Blur (f) $p < n$, Noise

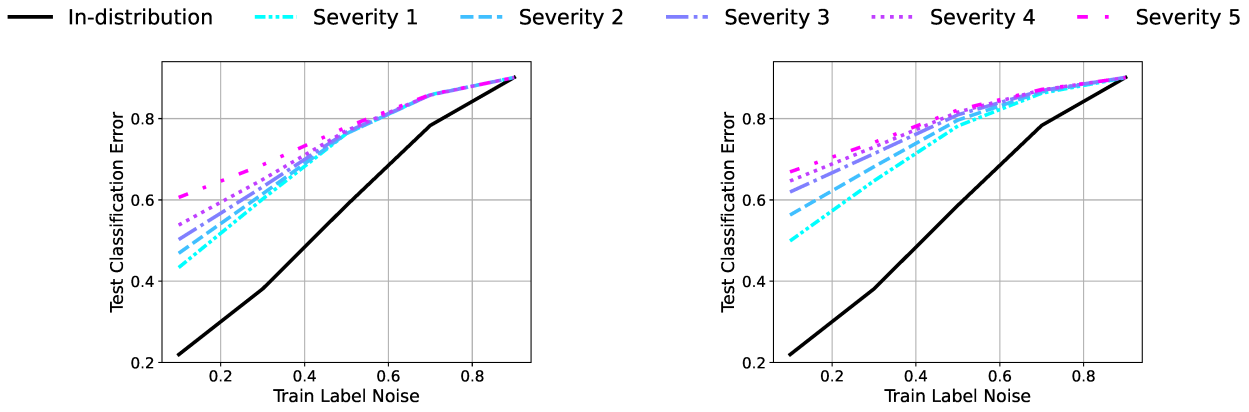
Figure 9: We fit the ridgeless OLS solution to binary CIFAR-10 (dog vs. truck) and test on binary CIFAR-10C under Gaussian blur and noise corruptions. In the top row, we vary the level of overparameterization as $n = 500, 1k, 2k$ and average each curve over 50 independent runs. In this $p > n$ setting the ridgelss OLS solution results in the MNI. In the bottom row we obtain a non-interpolating, ridgeless linear solution. Evaluations are done on severity 3 of CIFAR-10C, however the results hold up across all severities. We plot excess classification error vs training label noise, which is class label flip probability. We see that overparameterization improves robustness of the MNI at all noise levels.

In Figure 9 (a-c) we show that overparameterization improves OOD excess classification error for the MNI fit to binary CIFAR-10 and evaluated on binary CIFAR-10C under Gaussian blur and noise corruptions. The details of these datasets and setups are given in Appendix I. We note that all of the curves in the top

row of this figure are in the overparameterized regime, meaning they are on the right side of the double descent curve. Flattened CIFAR images have $p = 3,072$ and so we vary the number of training subsample sizes over $n = 500, 1000, 2000$ in order to remain in an overparameterized setting. We find that when we are overparameterized, as we reduce n we obtain improved performance. We average over 50 independent runs in each setting and provide standard error bars to show that this observation is not due to specific random samples. We also see that at higher levels of overparameterization, the relative difference in excess classification error between ID, blur, and noise test sets lessens. For example, at 0.0 label noise and $n = 2k$ the average excess error varies from 0.3028 on the blur set to 0.4668 on the noise set for an absolute difference of 0.164, whereas at $n = 500$ the average excess error only varies from 0.252 on the blur set to 0.3131 on the noise set for an absolute difference of 0.0611.

For completeness, in Figure 9 (d - e) we show the above setting in the underparameterized regime where we obtain the linear solution via the ridgeless OLS solution. As these plots are on the left side of the double descent peak, we see that adding more data improves OOD excess classification error. While these models are not interpolating, we observe that noise corruptions lead to nearly *catastrophic* performance, meaning random guessing, on the OOD test sets, whereas blur corruptions lead to more *benign* performance. Finally the ID performance appears to be *tempered*, in showing a nearly linear relationship between train label noise and test excess classification error.

J.4 ResNet on CIFAR-10C Experiments



(a) Interpolating ResNet tested on Blur

(b) Interpolating ResNet tested on Noise

Figure 10: We train ResNet18 on clean CIFAR-10 and evaluate on test sets that has been corrupted by Gaussian blur and Gaussian noise, which correspond to beneficial and malignant shifts, respectively. Labels are flipped with probability 0.1 through 0.9, seen on the x-axis. The setting is not high-dimensional because the training data contains 50000 images, each of which are 3072-dimensional. ResNet18 contains around 11.7 million parameters, so the setting is very overparameterized. We observe that while both shifts negatively affect generalization, the beneficial shift isn't as bad as the malignant shift. This result is similar to those seen in subfigures (a) and (b) in Figure 2, where the data is not high-dimensional but the MLP is overparameterized.

Figure 10 shows the behavior of interpolating ResNets trained on the full CIFAR-10 dataset and evaluated on CIFAR-10C blur and noise corruptions. While these numbers are suboptimal with respect to CNNs on CIFAR-10 we note that they are justified in our setting as our goal is to study interpolating models. At

90% label noise it takes a lot of compute to interpolate the entire CIFAR-10 dataset, especially if using data augmentations, weight decay, or other regularizations. As such, we turn off weight decay and data augmentations for these models to be able to tractably interpolate CIFAR-10 at high noise levels.