

---

# Task Shift: From Classification to Regression in Overparameterized Linear Models

---

Tyler LaBonte\*

Georgia Institute of Technology  
tlabonte@gatech.edu

Kuo-Wei Lai\*

Georgia Institute of Technology  
klai36@gatech.edu

Vidya Muthukumar

Georgia Institute of Technology  
vmuthukumar8@gatech.edu

## Abstract

Modern machine learning methods have recently demonstrated remarkable capability to generalize under *task shift*, where latent knowledge is transferred to a different, often more difficult, task under a similar data distribution. We investigate this phenomenon in an overparameterized linear regression setting where the task shifts from classification during training to regression during evaluation. In the zero-shot case, wherein no regression data is available, we prove that task shift is impossible in both sparse signal and random signal models for any Gaussian covariate distribution. In the few-shot case, wherein limited regression data is available, we propose a simple postprocessing algorithm which asymptotically recovers the ground-truth predictor. Our analysis leverages a fine-grained characterization of individual parameters arising from minimum-norm interpolation which may be of independent interest. Our results show that while minimum-norm interpolators for classification cannot transfer to regression *a priori*, they experience surprisingly structured attenuation which enables successful task shift with limited additional data.

## 1 INTRODUCTION

The fields of modern statistics and machine learning aim to develop models which generalize to a plethora of application-specific tasks. For example, tasks in

computer vision range from classifying images into discrete categories to object detection (Ren et al., 2015), segmentation (Ronneberger et al., 2015), and pose estimation (Cao et al., 2017), while tasks in language modeling could be as basic as next-token prediction (Vaswani et al., 2017), or involve summarization (Liu and Lapata, 2019) or machine translation (Bahdanau et al., 2015). In statistics, basic estimation tasks involve either classification or regression; in the latter we wish to predict real-valued quantities and performance is measured via a continuous error metric. The traditional perspective on task shift establishes a clear hierarchy in difficulty, *e.g.*, a statistical estimator which achieves a certain error rate for a regression task will typically achieve an equal or better rate on the corresponding classification task<sup>1</sup>. Similarly, in empirical machine learning, the most difficult task is considered to be representation learning. Indeed, learned representations are commonly finetuned on simpler downstream tasks and observed to generalize in a *zero-shot* or *few-shot* sense, *i.e.*, when finetuning data is unavailable or limited, respectively (Bengio et al., 2013).

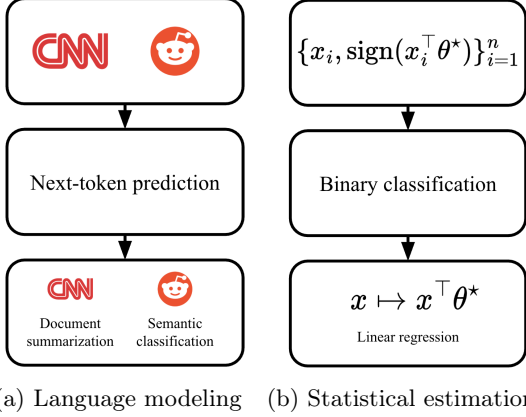
Perhaps more surprising are recent trends in modern machine learning which appear to go in the other direction: using models trained on an “easier” task to successfully solve a “harder” task. Specifically, large language models (LLMs) have shown a remarkable ability to generalize *in-context* — without explicit finetuning — to completing prompt-response pairs despite being trained only on the more basic next-token prediction task (Brown et al., 2020). From a statistical perspective, a particularly intriguing observation is that LLMs trained on next-token prediction can successfully solve linear regression tasks

---

\*Equal contribution; co-first author.

Proceedings of the 28<sup>th</sup> International Conference on Artificial Intelligence and Statistics (AISTATS) 2025, Mai Khao, Thailand. PMLR: Volume 258. Copyright 2025 by the author(s).

<sup>1</sup> This is most directly used in applying logistic regression procedures to classification tasks, but also works for, *e.g.*, least-squares regression (Kline and Berardi, 2005; Rifkin et al., 2003).



**Figure 1: Task shift in language modeling and statistical estimation.** In our task shift setting, latent knowledge is transferred between tasks under a similar conditional distribution or ground-truth signal. Task shift is compelling when the aim is to shift to a fundamentally harder task, with little to no data available from the new task.

by computing the ordinary least-squares estimate in-context (Zhang et al., 2024). With this high-level motivation, we propose the following statistical learning problem formulation:

*Can estimators trained on a classification task generalize, in a zero-shot or few-shot sense, to the regression task on the same data distribution?*

**Our contributions.** We consider linear binary classification on data  $\{(\mathbf{x}_i, \hat{y}_i = \text{sign}(\mathbf{x}_i^\top \boldsymbol{\theta}^*))\}_{i=1}^n$  and investigate whether an estimator trained on the classification task can generalize to the corresponding regression task, *i.e.*, predict the regression label  $\mathbf{x}^\top \boldsymbol{\theta}^*$  of a new datum  $\mathbf{x}$ . We consider the *overparameterized* regime wherein the dimensionality  $d$  of the data greatly exceeds the number of training examples  $n$ , and we study the minimum  $\ell_2$ -norm interpolator (MNI) on the binary labels  $\{\hat{y}_i\}_{i=1}^n$ , which we denote by  $\hat{\boldsymbol{\theta}}$ . We define the *task shift error* as the difference between the regression risk of the classification MNI  $\hat{\boldsymbol{\theta}}$  and the regression risk of the regression MNI, which we denote by  $\tilde{\boldsymbol{\theta}}$ . We show the following results with high probability over the training data:

- Classification data attenuates the signal  $\boldsymbol{\theta}^*$  even in the most favorable possible situation for either task in minimum  $\ell_2$ -norm interpolation, *i.e.*, maximally anisotropic data covariance, sublinearly sparse signal, and the existence of benign overfitting of noise. Therefore, the classification MNI  $\hat{\boldsymbol{\theta}}$  *does not successfully generalize in a*

*zero-shot sense to regression data*, except when effective signal magnitude is equal to a specific, pre-defined constant. See Theorem 8 for a formal statement of this result.

- We also produce an “ansatz” prediction of task shift error for more general signal models under a simplifying assumption on the regression labels (Theorem 12). Corollary 13 then shows a fundamental *tradeoff between regression bias and task shift error* — they cannot be simultaneously statistically consistent. Moreover, we show in Theorem 23 that while a “dense” random signal is known to suffer from poor bias, it achieves vanishing task shift error if the covariance matrix has large effective rank compared to  $n$ , *i.e.*, its eigenvalues decay sufficiently slowly.
- Finally, we consider a  $t$ -sparse ground truth  $\boldsymbol{\theta}^*$  and propose a simple postprocessing algorithm utilizing few-shot regression data. We show the attenuation of the classification MNI is surprisingly structured, culminating in Theorem 14 which proves the support of  $\boldsymbol{\theta}^*$  can be recovered simply by the  $t$  largest elements (in absolute value) of  $\hat{\boldsymbol{\theta}}$ . Our postprocessing algorithm ensures  $\mathcal{O}(\frac{t}{m})$  regression error with  $m$  noisy regression examples or exact recovery from  $t$  noiseless examples — in other words, *successful few-shot generalization from classification to regression*.

Our techniques build on the recent literature on benign overfitting of minimum  $\ell_2$ -norm interpolators in both regression (Bartlett et al., 2020) and classification tasks (Muthukumar et al., 2021; Wang et al., 2023). We are especially inspired by the separation in statistical consistency derived in Muthukumar et al. (2021), which showed that for certain anisotropic ensembles, classification may generalize while regression does not. We substantially develop their tools to provide a fine-grained characterization of the individual magnitudes  $\{|\hat{\theta}_j|\}_{j=1}^d$ . Surprisingly, this characterization can enable the success of our few-shot algorithm even when the minimum  $\ell_2$ -norm interpolator would not generalize for either classification or regression.

## 1.1 Related work

The formulation of task shift in this paper, particularly our focus on the shift from classification to regression tasks, shares both similarities and differences with several prominent areas of research in machine learning. Compared to transfer learning (Pan and Yang, 2009), task shift similarly aims to generalize knowledge from one “job” to another. However,

while transfer learning emphasizes preserving useful features for similar or downstream tasks, task shift focuses on generalizing from a simpler task (*e.g.*, classification) to a more complex one (*e.g.*, regression). Task shift is also related to the concept of distribution shift (Moreno-Torres et al., 2012). In task shift, the source data distribution remains unchanged, but the conditional distributions of the labels differ at test time. Furthermore, our classification-to-regression setup is closely connected to the one-bit compressive sensing problem (Boufounos and Baraniuk, 2008). However, while one-bit compressive sensing focuses on optimal estimators or algorithms based on a known measurement process, our work emphasizes unique properties of the  $\ell_2$ -inductive bias.

The theoretical analysis in this work builds directly on the literature on benign overfitting. Bartlett et al. (2020) Tsigler and Bartlett (2023) characterized benign overfitting for regression estimators, while Muthukumar et al. (2021) provided a survival and contamination analysis for sparse signals in both classification and regression settings. Our analysis substantially develops insights from these works to estimate the support of a sparse signal even when neither classification nor regression generalizes.

We discuss additional related work in Appendix A.

## 2 PRELIMINARIES

We now detail our classification-to-regression task shift setting and introduce our proposed estimators and assumptions on data covariance. We also recap the definitions of *effective dimension* quantities extensively used in analysis of interpolating estimators, *e.g.*, by Bartlett et al. (2020).

**Notation.** We use uppercase bold symbols to denote matrices (*e.g.*,  $\mathbf{X}$ ), lowercase bold symbols to denote vectors (*e.g.*,  $\mathbf{x}$ ), and lowercase italicized symbols to denote scalars (*e.g.*,  $x$ ). We denote random variables using non-italicized symbols, *e.g.*,  $\mathbf{X}$  for a random matrix and  $\mathbf{x}$  for a random vector. Let  $\mathbb{P}_{\mathbf{x}}$  and  $\mathbb{E}_{\mathbf{x}}$  denote a probability and expectation with respect to a random vector  $\mathbf{x}$ , respectively.

Let  $\stackrel{d}{=}$  denote equality in distribution. Let  $\mathcal{S}^c$  denote the complement of a set  $\mathcal{S}$ . Let  $\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$  denote that  $\mathbf{x}$  is sampled from a multivariate Gaussian distribution with mean  $\boldsymbol{\mu}$  and covariance matrix  $\boldsymbol{\Sigma}$ . Let  $\text{diag}(\alpha_1, \dots, \alpha_n)$  denote an  $n \times n$  matrix with  $\alpha_1, \dots, \alpha_n$  on the diagonal and zeroes elsewhere. Let  $\mu_1(\mathbf{A}), \mu_2(\mathbf{A}), \dots$  denote the eigenvalues of a matrix  $\mathbf{A}$  in non-increasing order. Let

$\|\mathbf{A}\| := \mu_1(\mathbf{A})$  and  $\text{Tr}(\mathbf{A}) := \sum_i \mu_i(\mathbf{A})$  denote the operator norm and trace of  $\mathbf{A}$ . For clarity, we use the shorthand  $[t] := \{1, 2, \dots, t\}$  for  $t \in \mathbb{N}$ . Finally, we use  $C, c, c_1, \dots$  to denote constants not depending on  $n, d$  which can change from line to line.

### 2.1 Minimum-norm interpolating estimators

We consider a noiseless linear regression problem where data undergoes centered Gaussian featurization<sup>2</sup> in  $d$  dimensions such that  $\mathbf{x} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$ . We denote the ground-truth regressor, also called *signal*, by  $\boldsymbol{\theta}^* \in \mathbb{R}^d$ , so the regression and classification label for a datum  $\mathbf{x}$  are  $\tilde{y} := \mathbf{x}^\top \boldsymbol{\theta}^* \in \mathbb{R}$  and  $\hat{y} := \text{sign}(\mathbf{x}^\top \boldsymbol{\theta}^*) \in \{\pm 1\}$ , respectively.

We assume noiseless labels in the classification dataset for ease of exposition, but our zero-shot lower bounds and few-shot upper bounds can be easily extended to handle classification label noise<sup>3</sup>.

We assume access to a classification dataset  $\{(\mathbf{x}_i, \hat{y}_i)\}_{i=1}^n$  where  $d \gg n$  and  $\lim_{n \rightarrow \infty} \frac{d}{n} = \infty$ , *i.e.*, the data is heavily overparameterized. We denote the data matrix by  $\mathbf{X} := [\mathbf{x}_1, \dots, \mathbf{x}_n]^\top \in \mathbb{R}^{n \times d}$  and the regression and classification label vectors by  $\tilde{\mathbf{y}} := [\tilde{y}_1, \dots, \tilde{y}_n]^\top \in \mathbb{R}^n$  and  $\hat{\mathbf{y}} := [\hat{y}_1, \dots, \hat{y}_n]^\top \in \{\pm 1\}^n$ . We study the *minimum-norm interpolator* (MNI), which is obtained directly via the implicit bias of gradient descent on the squared loss and also enjoys close links to the implicit bias of popular classification losses (Hsu et al., 2022). The MNIs on regression and classification labels are defined by  $\hat{\boldsymbol{\theta}} := \arg \min \{\|\boldsymbol{\theta}\|_2 : \mathbf{X}\boldsymbol{\theta} = \tilde{\mathbf{y}}\}$  and  $\hat{\boldsymbol{\theta}} := \arg \min \{\|\boldsymbol{\theta}\|_2 : \text{sign}(\mathbf{X}\boldsymbol{\theta}) = \hat{\mathbf{y}}\}$ , respectively. Since  $\mathbf{X}$  is full rank almost surely under Gaussian design, both estimators have simple closed forms:  $\tilde{\boldsymbol{\theta}} = \mathbf{X}^\top (\mathbf{X}\mathbf{X}^\top)^{-1} \tilde{\mathbf{y}}$  and  $\hat{\boldsymbol{\theta}} = \mathbf{X}^\top (\mathbf{X}\mathbf{X}^\top)^{-1} \hat{\mathbf{y}}$ .

The excess risk of any linear estimator  $\boldsymbol{\theta} \in \mathbb{R}^d$  is given by  $L(\boldsymbol{\theta}) := \mathbb{E}_{\mathbf{x}} (\mathbf{x}^\top \boldsymbol{\theta} - \mathbf{x}^\top \boldsymbol{\theta}^*)^2$ . The central goal of this paper is to bound the excess *regression* risk of the MNI *classifier*  $L(\hat{\boldsymbol{\theta}})$  — and, later, postprocessed variants of this classifier. We say  $\hat{\boldsymbol{\theta}}$  *achieves task shift* if it is regression-consistent, *i.e.*,  $\lim_{n, d \rightarrow \infty} L(\hat{\boldsymbol{\theta}}) = 0$ .

### 2.2 Effective rank and covariance structure

As shown in Bartlett et al. (2020); Muthukumar et al. (2021); Tsigler and Bartlett (2023), the performance

<sup>2</sup> We expect many of our results to generalize to the case of independent, sub-Gaussian features by building on Cao et al. (2021); Tsigler and Bartlett (2023).

<sup>3</sup> For example, if  $\mathbb{P}(\hat{y} = \text{sign}(\mathbf{x}^\top \boldsymbol{\theta}^*) \mid \mathbf{x}) = 1 - \nu^*$ , we require  $\nu^* < 0.5$ , *i.e.*, the sign is preserved on average.

of the regression MNI on regression labels or classification MNI on classification labels can be characterized by two notions of *effective rank* of the spectrum of the data covariance matrix  $\Sigma$ . We let  $\lambda_1, \dots, \lambda_d$  denote the eigenvalues of  $\Sigma$  in non-increasing order (i.e.,  $\lambda_j := \mu_j(\Sigma)$  for all  $j \in [d]$ ).

**Definition 1** (Effective rank). *For an index  $k \geq 0$ , two notions of the effective rank of  $\Sigma$  are*

$$r_k(\Sigma) := \frac{\sum_{j=k+1}^d \lambda_j}{\lambda_{k+1}} \text{ and } R_k(\Sigma) := \frac{\left(\sum_{j=k+1}^d \lambda_j\right)^2}{\sum_{j=k+1}^d \lambda_j^2}.$$

These are essentially two notions of effective dimension of a “tail” component of the covariance matrix restricted to eigenvalues  $\lambda_{k+1}, \dots, \lambda_d$ . For a constant  $b > 1$ , denote  $k^* := \min\{k \geq 0 : r_k(\Sigma) \geq bn\}$  where the minimum of the empty set is defined as  $\infty$ . In other words,  $k^*$  is the minimal index after which the spectrum has large (first) effective rank compared to  $n$ . We will make the following assumptions on  $\Sigma$ .

**Assumption 2.** *We assume that  $\Sigma$  is diagonal and positive definite for any  $d, n < \infty$ . We also assume that  $k^* \leq \frac{n}{c}$  for some universal constant  $c > 0$ .*

The diagonal assumption is without loss of generality for our zero-shot lower bounds but required for our few-shot upper bounds. The assumption on effective rank essentially requires a “long tail” on the covariance and ensures that the data is actually high dimensional in nature.

We situate our results in two covariance ensembles considered in Bartlett et al. (2020); Muthukumar et al. (2021) which enable us to state precise rates.

**Definition 3** (Spiked covariance matrix). *A spiked covariance matrix  $\Sigma$  is parameterized by  $p \in (1, \infty)$ ,  $r \in [0, 1)$ , and  $q \in (0, p - r)$ . We set the data dimension to  $d = n^p$ , the length of the “spike” to  $s = n^r$ , and  $a = n^{-q}$  to be a parameter controlling the ratio of the eigenvalues. Then, the ensemble is defined by*

$$\lambda_j := \begin{cases} \frac{ad}{s} & j \in [s] \\ \frac{(1-a)d}{d-s} & \text{otherwise.} \end{cases}$$

*This ensemble satisfies Assumption 2 with  $k^* = s$  when  $q < 1 - r$  and  $k^* = 0$  when  $q > 1 - r$ .*

**Definition 4** (Polynomial decay covariance matrix). *A polynomial decay covariance matrix  $\Sigma$  is parameterized by  $p \in (1, \infty)$  and  $u, v \geq 0$  such that*

$$\lambda_j := j^{-u} \ln^{-v}(j + 1)$$

*and  $d = n^p$ . We consider two versions of the covariance. We set  $u = 1, v = 2$  to study an instance*

*of the regime wherein benign overfitting is achieved, and we set  $u \in [0, 1), v = 0$  to study a case wherein it is not achieved (Bartlett et al., 2020). These parameterizations satisfy Assumption 2, but unlike the spiked covariance matrix, characterizing  $k^*$  is rather delicate. In particular, if  $k^*$  is nonzero we may only be able to characterize its order (see Appendix C.3).*

Formally, we take limits over a sequence of covariance ensembles  $\{\Sigma_n\}_{n=1}^\infty$  (e.g., Theorem 8, Corollaries 9 and 13), but we drop the subscript  $n$  for clarity.

### 3 ZERO-SHOT TASK SHIFT

In this section, we study task shift performance in the zero-shot setting, wherein no regression data is available. We show that task shift is impossible under sparse and random signal models with maximally favorable data covariance. While our negative results are perhaps expected — since the magnitude of the regression labels is irrevocably lost in the classification task — our analysis leads to some unexpected conclusions. Specifically, in Section 3.1, we show that the attenuation of the classification MNI is surprisingly structured (which enables recovery of the true signal up to a magnitude factor), and in Section 3.2, we show that the nature of the failure of zero-shot task shift is closely linked to the index  $k^*$ .

#### 3.1 The case of sparse signal

We first consider a signal with *sublinear sparsity*.

**Assumption 5** ( $t$ -sparse signal model). *Denote the support of the signal  $\theta^*$  by  $\mathcal{S} := \{j \in [d] \mid \theta_j^* \neq 0\}$ . We assume  $\theta^*$  is  $t$ -sparse, that is  $|\mathcal{S}| = t \ll n^{\frac{1}{2}-\epsilon} \ll d$ , where  $\epsilon \in (0, \frac{1}{4})$ . We write  $\theta_j^* := \frac{a_j}{\sqrt{\lambda_j}}$  for  $j \in \mathcal{S}$  with  $a_j \neq 0$ . Moreover, we assume the total signal strength  $\|\Sigma^{\frac{1}{2}}\theta^*\|_2^2 = \sum_{j \in \mathcal{S}} a_j^2$  is constant for all  $n$ .<sup>4</sup>*

We note that the sparsity assumption is justified since in its absence, regression generalization is information-theoretically impossible in the overparameterized regime (Wainwright, 2009; Aeron et al., 2010). For the regression MNI, Tsigler and Bartlett (2023) showed one must have approximate sparsity in the direction of the top eigenvalues to ensure low bias, i.e., sparsity is required for statistical consistency.

Rather than the standard bias-variance decomposition, we directly investigate the relative preservation of the true signal (*survival*) and the pollution

<sup>4</sup> The assumption on total signal strength is only necessary for limits over  $\{\Sigma_n\}_{n=1}^\infty$ , i.e., it is not required for our few-shot results in Section 4.



of false signal (*contamination*), which were shown in Muthukumar et al. (2021) to tightly characterize regression and classification tasks. We define these quantities formally below, using shorthand notation  $\mathbf{A} := \mathbf{X}\mathbf{X}^\top = \sum_{j=1}^d \lambda_j \mathbf{z}_j \mathbf{z}_j^\top$  where  $\mathbf{z}_j := \mathbf{X} \mathbf{e}_j / \sqrt{\lambda_j}$  and  $\mathbf{z}_j \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_n)$  is an independent isotropic Gaussian vector for  $j \in [d]$ .

**Definition 6** (Survival and contamination). *Under Assumptions 2 and 5, we define the survival  $\text{SU}_j$  for all  $j \in \mathcal{S}$  and contamination  $\text{CN}$  by*

$$\text{SU}_j := \frac{\hat{\theta}_j}{\theta_j^*} = \frac{\mathbf{e}_j^\top \mathbf{X}^\top (\mathbf{X}\mathbf{X}^\top)^{-1} \hat{\mathbf{y}}}{\frac{a_j}{\sqrt{\lambda_j}}} = \frac{\lambda_j}{a_j} \mathbf{z}_j^\top \mathbf{A}^{-1} \hat{\mathbf{y}},$$

$$\text{CN} := \sqrt{\sum_{j \in \mathcal{S}^c} \lambda_j \hat{\theta}_j^2}.$$

Intuitively, we desire survival to be close to one and contamination to be small. In the remainder of this section, we substantially generalize the analysis in Muthukumar et al. (2021); Wang et al. (2023), which assumed  $t = 1$  and spiked covariance, to  $t$ -sparse and general covariance models. Using Assumption 5 and Gaussian design, we may write

$$\tilde{\mathbf{y}} := \mathbf{X}\boldsymbol{\theta}^* = \sum_{j \in \mathcal{S}} \sqrt{\lambda_j} \theta_j^* \mathbf{z}_j = \sum_{j \in \mathcal{S}} a_j \mathbf{z}_j,$$

$$\hat{\mathbf{y}} := \text{sign}(\mathbf{X}\boldsymbol{\theta}^*) = \text{sign}\left(\sum_{j \in \mathcal{S}} a_j \mathbf{z}_j\right).$$

This expression enables us to perform careful leave- $t$ -out analyses and decouple the survival and contamination terms in the following key lemma.

**Lemma 7.** *Define  $\bar{\mathcal{S}} := \mathcal{S} \cup [k^*]$  and denote by  $\{\tilde{\lambda}_j\}_{j=1}^{d-|\bar{\mathcal{S}}|}$  the diagonal entries of the matrix  $\boldsymbol{\Sigma}_{-\bar{\mathcal{S}}}$ , i.e.,  $\boldsymbol{\Sigma}$  with rows and columns indexed by  $\bar{\mathcal{S}}$  left out, and  $\mathbf{A}_{-\bar{\mathcal{S}}} = \sum_{j=1}^{d-|\bar{\mathcal{S}}|} \tilde{\lambda}_j \mathbf{z}_j \mathbf{z}_j^\top$ . Under Assumptions 2 and 5 and for large enough  $n$ , we have for all  $j \in \mathcal{S}$ ,*

$$\begin{aligned} \text{SU}_j &\leq \sqrt{\frac{2}{\pi \|\boldsymbol{\Sigma}^{\frac{1}{2}} \boldsymbol{\theta}^*\|_2^2}} \cdot \frac{\left(1 + \frac{c_1 t}{n^{\frac{1}{2}-\epsilon}}\right) \lambda_j \text{Tr}(\mathbf{A}_{-\bar{\mathcal{S}}}^{-1})}{1 + \left(1 - \frac{c_2}{n^{\frac{1}{2}-\epsilon}}\right) \lambda_j \text{Tr}(\mathbf{A}_{-\bar{\mathcal{S}}}^{-1})} \\ &\leq \sqrt{\frac{2}{\pi \|\boldsymbol{\Sigma}^{\frac{1}{2}} \boldsymbol{\theta}^*\|_2^2}} \cdot \frac{\lambda_j \left(\frac{c_3 n}{\tilde{\lambda}_1 r_0(\boldsymbol{\Sigma}_{-\bar{\mathcal{S}}})}\right)}{1 + \lambda_j \left(\frac{n}{c_4 \tilde{\lambda}_1 r_0(\boldsymbol{\Sigma}_{-\bar{\mathcal{S}}})}\right)}, \\ \text{SU}_j &\geq \sqrt{\frac{2}{\pi \|\boldsymbol{\Sigma}^{\frac{1}{2}} \boldsymbol{\theta}^*\|_2^2}} \cdot \frac{\left(1 - \frac{c_5 t}{n^{\frac{1}{2}-\epsilon}}\right) \lambda_j \text{Tr}(\mathbf{A}_{-\bar{\mathcal{S}}}^{-1})}{1 + \left(1 + \frac{c_6}{n^{\frac{1}{2}-\epsilon}}\right) \lambda_j \text{Tr}(\mathbf{A}_{-\bar{\mathcal{S}}}^{-1})} \\ &\geq \sqrt{\frac{2}{\pi \|\boldsymbol{\Sigma}^{\frac{1}{2}} \boldsymbol{\theta}^*\|_2^2}} \cdot \frac{\lambda_j \left(\frac{n}{c_7 \tilde{\lambda}_1 r_0(\boldsymbol{\Sigma}_{-\bar{\mathcal{S}}})}\right)}{1 + \lambda_j \left(\frac{c_8 n}{\tilde{\lambda}_1 r_0(\boldsymbol{\Sigma}_{-\bar{\mathcal{S}}})}\right)}, \end{aligned}$$

with probability at least  $1 - cte^{-n^{2\epsilon}}$ . When  $n, d \rightarrow \infty$ , the limit converges as

$$\lim_{n, d \rightarrow \infty} \text{SU}_j = \sqrt{\frac{2}{\pi \|\boldsymbol{\Sigma}^{\frac{1}{2}} \boldsymbol{\theta}^*\|_2^2}} \cdot \frac{\lambda_j \text{Tr}(\mathbf{A}_{-\bar{\mathcal{S}}}^{-1})}{1 + \lambda_j \text{Tr}(\mathbf{A}_{-\bar{\mathcal{S}}}^{-1})}$$

almost surely. Moreover, we have

$$\text{CN} \leq ct \sqrt{\left(\frac{k^*}{n} + \frac{n}{R_0(\boldsymbol{\Sigma}_{-\bar{\mathcal{S}}})}\right) \ln(n)}$$

with probability at least  $1 - \frac{ct}{n}$ .

Lemma 7 is proved in Appendix B. In addition to being useful for the results of this section, Lemma 7 is also used for our postprocessing algorithm which has access to few-shot data in Section 4. Equipped with bounds of survival and contamination, we can now relate the classification MNI excess risk with  $\{\text{SU}_j\}_{j \in \mathcal{S}}$  and  $\text{CN}$ . Under Assumptions 2 and 5, we can write the excess risk of the classification MNI as

$$\begin{aligned} L(\hat{\boldsymbol{\theta}}) &= \mathbb{E}_{\mathbf{x}} \left[ \left( \mathbf{x}^\top \hat{\boldsymbol{\theta}} - \mathbf{x}^\top \boldsymbol{\theta}^* \right)^2 \right] \\ &= \sum_{j=1}^d \lambda_j \left( \hat{\theta}_j - \theta_j^* \right)^2 \\ &= \sum_{j \in \mathcal{S}} \lambda_j \theta_j^{*2} \left( \frac{\hat{\theta}_j}{\theta_j^*} - 1 \right)^2 + \sum_{j \in \mathcal{S}^c} \lambda_j \hat{\theta}_j^2 \\ &= \sum_{j \in \mathcal{S}} a_j^2 (\text{SU}_j - 1)^2 + \text{CN}^2, \end{aligned} \quad (1)$$

where we substitute the expressions for  $\theta_j^*$  from Assumption 5 and  $\{\text{SU}_j\}_{j \in \mathcal{S}}$  and  $\text{CN}$  from Definition 6. Note that this decomposition is valid for any estimator of  $\boldsymbol{\theta}^*$  (not just the classification MNI  $\hat{\boldsymbol{\theta}}$ ). Substituting the bounds on survival and contamination from Lemma 7 yields our main result of this section.

**Theorem 8.** *Define  $b_j := \lim_{n, d \rightarrow \infty} \lambda_j \text{Tr}(\mathbf{A}_{-\bar{\mathcal{S}}}^{-1})$ . Under Assumptions 2 and 5, for any covariance  $\boldsymbol{\Sigma}$  satisfying  $\lim_{n, d \rightarrow \infty} \frac{t^2 \cdot k^* \cdot \ln(n)}{n} = \lim_{n, d \rightarrow \infty} \frac{t^2 \cdot n \cdot \ln(n)}{R_0(\boldsymbol{\Sigma}_{-\bar{\mathcal{S}}})} = 0$ , we have*

$$\lim_{n, d \rightarrow \infty} L(\hat{\boldsymbol{\theta}}) = \sum_{j \in \mathcal{S}} a_j^2 \left( \sqrt{\frac{2}{\pi \|\boldsymbol{\Sigma}^{\frac{1}{2}} \boldsymbol{\theta}^*\|_2^2}} \frac{b_j}{1 + b_j} - 1 \right)^2$$

almost surely.

Theorem 8 is proved in Appendix C.1. It shows that, even for data covariances that satisfy benign overfitting of noise in linear regression (Bartlett et al., 2020),

perfect survival of signal is required for consistent task shift. Notably, it is not possible for the classification MNI  $\hat{\theta}$  to satisfy regression consistency for all possible magnitudes of the ground truth, *i.e.*, for all possible values of  $\|\Sigma^{\frac{1}{2}}\theta^*\|_2^2$ . This is because the coefficients  $\{b_j\}_{j \in \mathcal{S}}$  in Theorem 8 clearly do not depend on  $\theta^*$ . We can, however, ask the more specialized question (posed in one-bit compressive sensing, *e.g.*, Plan and Vershynin (2012)) of whether it is possible to generalize on all signals of a specific magnitude. Theorem 8 shows this will be the case if and only if  $b_j \rightarrow \infty$  for all  $j \in \mathcal{S}$  and  $\|\Sigma^{\frac{1}{2}}\theta^*\|_2^2 = \frac{2}{\pi}$ : a positive result in the flavor of one-bit compressive sensing.

As a corollary of Theorem 8, we present characterizations of zero-shot task shift for the spiked covariance model (Definition 3). We present similar results for the more delicate polynomial decay covariance model (Definition 4) in Appendix C.3 (Corollary 18).

**Corollary 9.** *Suppose Assumptions 2 and 5 are satisfied with  $t \ll \min\left\{\sqrt{\frac{n}{k^* \cdot \ln(n)}}, \sqrt{\frac{n^{p-1}}{\ln(n)}}\right\}$ . Under the spiked covariance model (Definition 3),*

- For  $q < 1 - r$ , we have:  $\lim_{n, d \rightarrow \infty} L(\hat{\theta}) =$

$$\sum_{j \leq s, j \in \mathcal{S}} a_j^2 \left( \sqrt{\frac{2}{\pi \|\Sigma^{\frac{1}{2}}\theta^*\|_2^2}} - 1 \right)^2 + \sum_{j > s, j \in \mathcal{S}} a_j^2.$$

*This implies regression consistency if and only if the signal magnitude is fixed at  $\|\Sigma^{\frac{1}{2}}\theta^*\|_2^2 = \frac{2}{\pi}$  and  $a_j = 0$  for all  $j \in \mathcal{S} \cap \{s+1, \dots, d\}$ , *i.e.*, the signal is only supported within the “spike”. The latter condition is also required for regression consistency of the regression MNI (Tsigler and Bartlett, 2023).*

- For  $q > 1 - r$ , we have:  $\lim_{n, d \rightarrow \infty} L(\hat{\theta}) = \sum_{j \in \mathcal{S}} a_j^2$ .

*As in the case of the regression MNI (Muthukumar et al., 2021), regression consistency would not be possible unless we had zero signal, *i.e.*,  $\theta^* = 0$ .*

Corollary 9 is proved in Appendix C.2.

### 3.2 The case of random signal

We now provide results for *random signal models* which may be dense. In this section, we study a general random signal model and introduce a simplifying ansatz which enables upper and lower bounding the task shift error. In Appendix E (Theorem 23), we show a more specific “dense” random signal model —

wherein  $\theta^*$  has similar magnitude in all dimensions — which does not require the simplifying ansatz.

The key idea is to interpret classification labels as regression labels under a dependent noise model and explicitly connect to characterizations of the regression MNI (Bartlett et al., 2020; Tsigler and Bartlett, 2023). For regression labels, one usually has additive sub-Gaussian noise, which means that  $\tilde{y} - \mathbf{x}^\top \theta^*$  is sub-Gaussian and conditionally independent given  $\mathbf{x}$ . But for classification labels, we have

$$\hat{y} - \mathbf{x}^\top \theta^* = \frac{\text{sign}(\mathbf{x}^\top \theta^*) - \mathbf{x}^\top \theta^*}{\mathbf{x}^\top \theta^*} \cdot \mathbf{x}^\top \theta^*,$$

which is clearly dependent on  $\mathbf{x}$ . We will write  $\hat{y} - \tilde{y} := \mathbf{D}\mathbf{X}\theta^*$ , where  $\mathbf{D} := \text{diag}(d_1, \dots, d_n)$  and  $d_i := \frac{\text{sign}(\mathbf{x}^\top \theta^*) - \mathbf{x}^\top \theta^*}{\mathbf{x}^\top \theta^*}$ , to make this relationship explicit.

We begin with a decomposition of the regression risk of the classification MNI, proved in Appendix D.1.

**Lemma 10.** *The regression risk of the classification MNI  $\hat{\theta}$  can be decomposed as*

$$L(\hat{\theta}) = L(\tilde{\theta}) + \mathbb{E}_{\mathbf{x}} \left( \mathbf{x}^\top \hat{\theta} - \mathbf{x}^\top \tilde{\theta} \right)^2.$$

Clearly,  $L(\tilde{\theta})$  is the *regression error* of the regression MNI: when there is no noise in regression labels this is equivalent to the *bias*. Likewise, we refer to  $\mathbb{E}_{\mathbf{x}}(\mathbf{x}^\top \hat{\theta} - \mathbf{x}^\top \tilde{\theta})^2$  as the *task shift error*, which can be interpreted as the “variance” under our dependent noise model.

The bias is a standard term, characterized as follows.

**Lemma 11.** *For any  $\Sigma$ , there exists a constant  $c \geq 1$  such that the following hold.*

**Upper bound.** (Bartlett et al., 2020, Lemma 35). *For any  $\theta^*$  (not necessarily random), we have*

$$L(\tilde{\theta}) \leq \sqrt{c} \|\Sigma\| \|\theta^*\|_2^2 \max \left( \sqrt{\frac{r_0(\Sigma)}{n}}, \frac{r_0(\Sigma)}{n} \right)$$

*with probability at least  $1 - e^{-\frac{n}{c}}$ .*

**Lower bound.** (Tsigler and Bartlett, 2023, Lemma 8). *Suppose random signal  $\theta^*$  is generated from the ground truth  $\bar{\theta}$  by  $\theta_j^* = r_j \bar{\theta}_j$  where each  $r_j$  is an independent Rademacher random variable. We have*

$$\mathbb{E}_{\theta^*} L(\tilde{\theta}) \geq \frac{1}{c} \sum_{j=1}^d \frac{\lambda_j \bar{\theta}_j^2}{\left(1 + \frac{n \lambda_j}{\sum_{k=1}^d \lambda_k}\right)^2}$$

*with probability at least  $1 - ce^{-\frac{n}{c}}$ .*

We now provide a characterization of the task shift error under the simplifying ansatz that  $\mathbf{D} = \alpha \mathbf{I}$  for some  $\alpha \neq 0$ . The interpretation for this assumption is that all regression labels have the same magnitude (say, equal to  $R$ ), which would result in  $d_i = \frac{1}{R} - 1 =: \alpha$ ; clearly  $\alpha \neq 0$  except in the special case where  $R = 1$ . Thus, we are considering regression problems that are, in essence, a scaled version of classification. From a technical perspective, it is difficult to obtain closed-form bounds on the task shift error without the simplifying ansatz, as dependencies which arise for general  $\mathbf{D}$  may invalidate certain concentration arguments. Nevertheless, as we expect generic regression problems to be even harder than scaled classification, providing a lower bound even for this simpler setting is meaningful. Our next theorem does precisely this, via an extension of benign overfitting techniques to our dependent noise model.

**Theorem 12.** *For any  $\Sigma$ , there exist constants  $c, c_1 \geq 1$  such that the following hold.*

**Upper bound.** *If  $k^* < \frac{n}{c_1}$ , then for any  $\theta^*$  (not necessarily random), we have*

$$\mathbb{E}_{\mathbf{x}} \left( \mathbf{x}^\top \hat{\theta} - \mathbf{x}^\top \tilde{\theta} \right)^2 \leq cn \left( \frac{k^*}{n} + \frac{n}{R_{k^*}(\Sigma)} \right) \|\Sigma^{\frac{1}{2}} \theta^*\|_2^2$$

*with probability at least  $1 - 18e^{-\frac{n}{c}}$ .*

**Lower bound.** *Suppose  $\mathbf{D} = \alpha \mathbf{I}$  and  $\theta^*$  is any random signal such that  $\mathbb{E} \theta_j^{*2} \geq \sigma^2$  for all  $j \in [d]$ . If  $k^* < \frac{n}{c_1}$ , then*

$$\begin{aligned} \mathbb{E}_{\theta^*, \mathbf{x}} \left( \mathbf{x}^\top \hat{\theta} - \mathbf{x}^\top \tilde{\theta} \right)^2 \\ \geq \frac{\alpha^2 \sigma^2}{c} \left( \sum_{j=1}^{k^*} \lambda_j + \frac{n}{R_{k^*}(\Sigma)} \sum_{j=k^*+1}^d \lambda_j \right) \end{aligned}$$

*with probability at least  $1 - 14e^{-\frac{n}{c}}$ . On the other hand, if  $k^* \geq \frac{n}{c_1}$ , then*

$$\mathbb{E}_{\theta^*, \mathbf{x}} \left( \mathbf{x}^\top \hat{\theta} - \mathbf{x}^\top \tilde{\theta} \right)^2 \geq \frac{\alpha^2 \sigma^2}{c}$$

*with probability at least  $1 - 10e^{-\frac{n}{c}}$ .*

Theorem 12 is proved in Appendix D.2. An interesting consequence of Theorem 12 is that there do exist covariance ensembles for which the task shift error decays to zero — implying that the classification and regression MNIs would generalize equivalently on a regression task! However, these are ensembles for which  $k^* = 0$ , and therefore regression bias would stay constant. Ultimately, our results imply a fundamental tradeoff between bias error and task shift error for random signals, stated below.

**Corollary 13.** *For any sequence  $\{\Sigma_n\}_{n=1}^\infty$ , denote  $\kappa := \lim_{n,d \rightarrow \infty} k^*$ . Suppose random signal  $\theta^*$  is generated from the ground truth  $\tilde{\theta}$  by  $\theta_j^* = r_j \tilde{\theta}_j$ , where each  $r_j$  is an independent Rademacher random variable, such that  $\|\Sigma^{\frac{1}{2}} \tilde{\theta}\|_2^2$  is constant for all  $n$ . Assume that  $\tilde{\theta}_j^2 \geq 1$  for all  $1 \leq j \leq d$ . Then, under the same conditions as the lower bound of Theorem 12, the almost sure limits of bias and task shift error are characterized in two distinct regimes:*

1.  $\kappa = 0$ : the limiting bias is nonzero, i.e.,

$$\lim_{n,d \rightarrow \infty} \mathbb{E}_{\theta^*} L(\tilde{\theta}) \geq \frac{\|\Sigma^{\frac{1}{2}} \tilde{\theta}\|_2^2}{c}.$$

2.  $\kappa > 0$ : the limiting task shift error is nonzero, i.e.,

$$\lim_{n,d \rightarrow \infty} \mathbb{E}_{\theta^*, \mathbf{x}} \left( \mathbf{x}^\top \hat{\theta} - \mathbf{x}^\top \tilde{\theta} \right)^2 \geq \frac{\alpha^2}{c}.$$

The proof of Corollary 13 is in Appendix D.5.

## 4 FEW-SHOT TASK SHIFT

In the previous sections, we demonstrated that task shift from classification to regression without any regression information is generally unachievable. Therefore, in this section, we investigate task shift in the *few-shot* setting, where limited regression information is available. We propose a simple two-step approach to recover a sparse signal  $\theta^*$ . In the first step, we leverage the structured attenuation of the classification MNI to recover the support of  $\theta^*$ . Second, we perform least-squares regression with reduction to the dimensionality of the support to recover the magnitude of  $\theta^*$ . Our results require a diagonal covariance matrix (Assumption 2).

### 4.1 Support recovery via attenuation

The survival and contamination bounds of Lemma 7 show that while the classification MNI  $\hat{\theta}$  attenuates the sparse signal  $\theta^*$ , it does so in a highly structured manner. This suggests it is possible to distinguish the support components of  $\theta^*$  using the relative magnitudes of entries of  $\hat{\theta}$ . If the true signal is supported within the top  $k^*$  indices of the covariance spectrum, the survival is bounded below by a constant; in contrast, contamination decays to zero with  $n$  at a faster rate than the survival terms (Muthukumar et al., 2021). Surprisingly, even when the signal is supported outside the top  $k^*$  indices, its decay rate may still be slower than the non-support components.

---

**Algorithm 1** Support recovery

---

**Require:**  $\mathbf{X}, \hat{\mathbf{y}}, (t \text{ or } \{\lambda_j\}_{j=1}^d)$   
 $\hat{\boldsymbol{\theta}} \leftarrow \mathbf{X}^\top (\mathbf{X}\mathbf{X}^\top)^{-1} \hat{\mathbf{y}}$   
 $\mathcal{S} \leftarrow \emptyset$   
**if**  $t$  is known **then**  
 $\mathcal{S} \leftarrow \arg \text{top } t(|\hat{\boldsymbol{\theta}}|)$   
**else if**  $\boldsymbol{\Sigma}$  is known **then**  
**for**  $j \in [d]$  **do**  
**if**  $|\hat{\theta}_j| \geq \lambda_j^{-\frac{1}{2}}$  **then**  
 $\mathcal{S} \leftarrow \mathcal{S} \cup \{j\}$   
**return**  $\mathcal{S}$

---

With this in mind, we propose Algorithm 1 assuming we either know the sparsity level  $t$  or the covariance spectrum  $\{\lambda_j\}_{j=1}^d$  (equivalently  $\boldsymbol{\Sigma}$  by assumption). Below, we state the general-purpose support recovery guarantee for Algorithm 1.

**Theorem 14.** *Under Assumptions 2 and 5, suppose  $\mathcal{S} \subseteq [k^*]$ , i.e.,  $\boldsymbol{\theta}^*$  is only supported in the top  $k^*$  indices of the covariance spectrum. Denote by  $\{\tilde{\lambda}_j\}_{j=1}^{d-k^*}$  the diagonal entries of the matrix  $\boldsymbol{\Sigma}_{-[k^*]}$ , i.e.,  $\boldsymbol{\Sigma}$  with the first  $k^*$  rows and columns left out. Algorithm 1 recovers the support of the true regressor with probability at least  $1 - ctde^{-n^{2\epsilon}}$  if either (1)  $t$  is known and the additional conditions  $\lambda_j \ll \frac{\lambda_q \cdot n^{1-2\epsilon}}{t^2}$  and  $\lambda_j \ll \frac{(\tilde{\lambda}_1 r_0 (\boldsymbol{\Sigma}_{-[k^*]})^2)}{t^2 \cdot (n^{1+2\epsilon}) \cdot \lambda_\ell}$  hold for all  $j \in \mathcal{S}$ ,  $q \in \mathcal{S}^c \cap [k^*]$ ,  $\ell > k^*$ , or (2)  $\boldsymbol{\Sigma}$  is known.*

The first case of Theorem 14 utilizes Lemma 7 to ensure the lower bound of support components of  $\hat{\boldsymbol{\theta}}$  is larger than the upper bound of non-support components of  $\hat{\boldsymbol{\theta}}$ . The two additional conditions in the first case of Theorem 14 are necessary to avoid scenarios wherein the support components of  $\boldsymbol{\theta}^*$  decay very quickly with  $n$  (recall that  $\theta_j^* := a_j \lambda_j^{-\frac{1}{2}}$ , so a larger  $\lambda_j$  implies faster decay). Roughly, the first condition states that support eigenvalues should not be much larger than non-support eigenvalues in the top  $k^*$  indices, and the second condition states that the top  $k^*$  eigenvalues should not be much larger than the squared sum of the “tail” eigenvalues. We analyze these conditions for specific covariance ensembles in Appendix F. On the other hand, the second case of Theorem 14 does not require any additional assumptions, as we show  $|\hat{\theta}_j|$  is lower bounded by  $\lambda_j^{-\frac{1}{2}}$  if  $j \in \mathcal{S}$  but decays at a faster rate if  $j \notin \mathcal{S}$ .

We provide the proof of Theorem 14 in Appendix F. We include extensions to cases wherein  $\boldsymbol{\theta}^*$  is supported outside the top  $k^*$  indices of the covariance spectrum for spiked and polynomial decay models —

---

**Algorithm 2** Least-squares on recovered support

---

**Require:**  $\mathbf{X}, \hat{\mathbf{y}}, \{\mathbf{x}'_i, y'_i\}_{i=1}^m$   
 $\hat{\boldsymbol{\theta}} \leftarrow \mathbf{X}^\top (\mathbf{X}\mathbf{X}^\top)^{-1} \hat{\mathbf{y}}$   
 $\mathcal{S} \leftarrow \text{from Algorithm 1}$   
 $\hat{\boldsymbol{\theta}}_{\text{post}} \leftarrow \arg \min_{\boldsymbol{\theta} \in \mathbb{R}^d} \sum_{i=1}^m (\mathbf{x}'_i^\top \boldsymbol{\theta} - y'_i)^2$   
s.t.  $\theta_j = 0 \quad \forall j \in \mathcal{S}^c$   
**return**  $\hat{\boldsymbol{\theta}}_{\text{post}}$

---

surprisingly including isotropic covariance, despite isotropy not being conducive to generalization even in classification tasks (Muthukumar et al., 2021).

## 4.2 Least-squares on recovered support

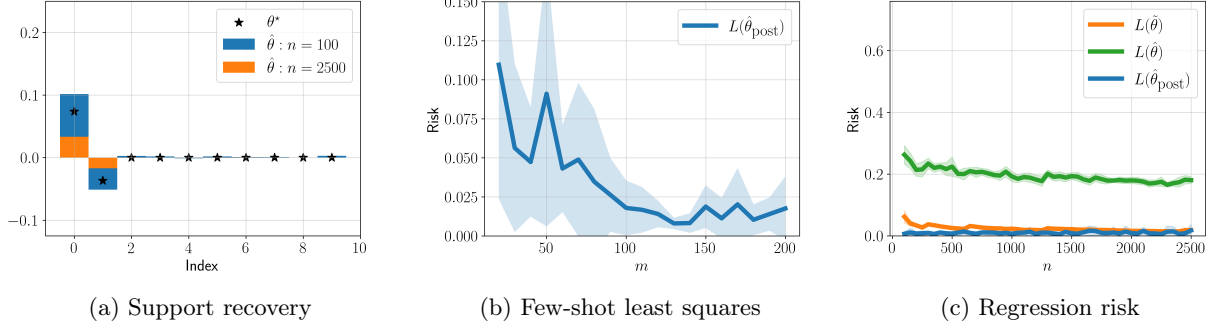
In this section, we leverage few-shot regression data  $\{(\mathbf{x}'_i, y'_i)\}_{i=1}^m$  to recover the magnitude of the  $t$ -sparse signal  $\boldsymbol{\theta}^*$ . Since the support of  $\boldsymbol{\theta}^*$  has already been recovered, we employ a straightforward least-squares estimation technique considering only the  $t$  components of each regression datum which lie in the support. Algorithm 2 describes this method in detail.

We recall that the few-shot regression dataset is allowed to be noisy, i.e., for some  $\sigma^2 \geq 0$ , we may have  $y' = \mathbf{x}'_i^\top \boldsymbol{\theta}^* + \varepsilon$  where  $\varepsilon \sim \mathcal{N}(0, \sigma^2 \mathbf{I})$ . Therefore, provided that  $m \gg t$ , Algorithm 2 enjoys the standard least-squares guarantee of  $\mathcal{O}(\frac{t\sigma^2}{m})$  regression risk, or zero regression error with only  $t$  noiseless samples. Either of these imply the desired regression consistency, i.e., task shift is achieved.

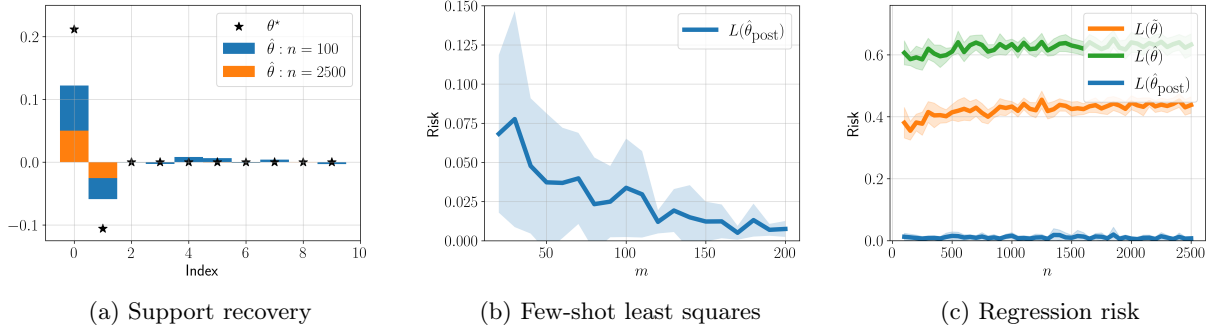
In Figure 2, we demonstrate the performance of our postprocessing procedure, combining Algorithm 1 and Algorithm 2. Notably, our task shift estimator  $\hat{\boldsymbol{\theta}}_{\text{post}}$  generalizes even for covariance ensembles wherein minimum  $\ell_2$ -norm interpolation fails, i.e., the regression MNI is statistically inconsistent with respect to regression labels. In fact, Appendix I demonstrates that this success persists even when the classification MNI is inconsistent for classification tasks — including the case of isotropic covariance.

A remaining question is whether we can recover the magnitudes of the support of  $\hat{\boldsymbol{\theta}}$  without *any* few-shot regression data. Lemma 7 implies that if the effective signal strength  $\|\boldsymbol{\Sigma}^{\frac{1}{2}} \boldsymbol{\theta}^*\|_2^2$  is known, then a simple scaling procedure is sufficient. Specifically, for all  $j \in \mathcal{S}$  we may set  $\hat{\theta}_{\text{post},j} = \hat{\theta}_j \sqrt{\frac{\pi}{2} \|\boldsymbol{\Sigma}^{\frac{1}{2}} \boldsymbol{\theta}^*\|_2^2}$ , and  $\hat{\theta}_{\text{post},j} = 0$  otherwise. Lemma 7 then directly implies that  $\hat{\boldsymbol{\theta}}_{\text{post}} \rightarrow \boldsymbol{\theta}^*$  as  $n, d \rightarrow \infty$  as long as  $b_j \rightarrow 1$  for all  $j \in \mathcal{S}$ . But while this approach is specialized to Gaussian covariates and noiseless classification data, Algorithm 2 is more robust to modeling assumptions,





(i) **Task shift for spiked covariance when regression generalizes.** We set  $p = 1.5$ ,  $q = 0.5$ , and  $r = 0.25$  so that  $q < 1 - r$ . In this regime, both our task shift estimator  $\hat{\theta}_{\text{post}}$  and the regression MNI  $\hat{\theta}$  generalize.



(ii) **Task shift for spiked covariance when regression does not generalize.** We set  $p = 1.5$ ,  $q = 0.5$ , and  $r = 0.55$  so that  $q > 1 - r$ . In this regime, our task shift estimator  $\hat{\theta}_{\text{post}}$  generalizes, but the regression MNI  $\hat{\theta}$  does not. Note that  $k^* = 0$  here, so  $\theta^*$  is necessarily supported outside the “spike”.

**Figure 2: Postprocessing achieves task shift even when minimum  $\ell_2$ -norm interpolation fails for both classification and regression.** The left column demonstrates the survival of  $t$ -sparse signal support components in the classification MNI  $\hat{\theta}$  while non-support components decay quickly. The middle column shows the  $\mathcal{O}(\frac{t}{m})$  regression error of least-squares with reduction to  $t$  dimensions using  $m$  regression samples under standard Gaussian noise. Finally, the right column displays the regression risk of the classification MNI, regression MNI, and our postprocessed predictor. The signal  $\theta^*$  is 2-sparse with  $a_1 = 1$  and  $a_2 = -0.5$  (see Assumption 5). The middle column fixes  $n = 2500$ . We plot the mean and standard deviation over 10 draws of the training dataset  $\mathbf{X}$ . See Appendix I for additional simulations.

and we believe it can handle even unknown label noise and sub-Gaussian covariates by building on Cao et al. (2021); Wang et al. (2023).

## 5 DISCUSSION

Our results paint a pessimistic picture for zero-shot task shift (perhaps as expected), but an optimistic one for the few-shot case. Our key insight is that the attenuation of the classification MNI is surprisingly structured, which suggests one can get more “mileage” out of the MNI than previously known, including for few-shot task shift to regression. A principal open question is whether there exist alternative formulations (e.g., shifting from multiclass classification to regression) more conducive to zero-shot task shift. More close-ended questions include providing a suc-

cessful few-shot procedure in the absence of any sparsity or data covariance assumption and studying task shift for minimum  $\ell_p$ -norm estimators where  $p \neq 2$ .

While we use task shift in large language models (LLMs) only as a motivating example for our theoretical investigation, our work could inform future analyses of LLMs and in-context learning. In particular, our survival and contamination analysis may be extended to the neural tangent kernel regime via recent frameworks for kernel interpolation, e.g., Mallinar et al. (2022); Tsigler and Bartlett (2023); Kaushik et al. (2024). Furthermore, recent work has characterized linear attention as high-dimensional linear regression under a specific data embedding (Lu et al., 2024), which could be analyzed in our framework to explain few-shot task shift in linear Transformers.

**Acknowledgments.** We thank Jacob Abernethy for the compute assistance and anonymous reviewers for helpful feedback. T.L. acknowledges support from the DoD NDSEG Fellowship. K.L. acknowledges support from an ARC-ACO Fellowship provided by Georgia Tech. V.M. acknowledges support from the NSF (awards CCF-223915 and IIS-2212182), Adobe Research, and Amazon Research.

## References

- Shuchin Aeron, Venkatesh Saligrama, and Manqi Zhao. Information theoretic bounds for compressed sensing. *IEEE Transactions on Information Theory*, 56(10):5111–5130, 2010.
- Pranjal Awasthi, Maria-Florina Balcan, Nika Haghtalab, and Hongyang Zhang. Learning and 1-bit compressed sensing under asymmetric noise. In *Conference on Learning Theory (COLT)*, 2016.
- Dzmitry Bahdanau, KyungHyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. In *International Conference on Learning Representations (ICLR)*, 2015.
- Peter L. Bartlett, Philip M. Long, Gábor Lugosi, and Alexander Tsigler. Benign overfitting in linear regression. *Proceedings of the National Academy of Sciences (PNAS)*, 117(48):30063–30070, 2020.
- Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(3):1798–1828, 2013.
- Petros T. Boufounos and Richard G. Baraniuk. 1-bit compressive sensing. In *Conference on Information Sciences and Systems (CISS)*, 2008.
- Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2020.
- Yuan Cao, Quanquan Gu, and Mikhail Belkin. Risk bounds for over-parameterized maximum margin classification on sub-gaussian mixtures. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2021.
- Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2D pose estimation using part affinity fields. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.
- Xiaobin Chang, Yongxin Yang, Tao Xiang, and Timothy M. Hospedales. Disjoint label space transfer learning with common factorised space. In *AAAI Conference on Artificial Intelligence (AAAI)*, 2019.
- Scott Shaobing Chen, David L Donoho, and Michael A Saunders. Atomic decomposition by basis pursuit. *SIAM Review*, 43(1):129–159, 2001.
- Geoffrey Chinot, Felix Kuchelmeister, Matthias Löffler, and Sara van de Geer. AdaBoost and robust one-bit compressed sensing. *Mathematical Statistics and Learning*, 5(1):117–158, 2022.
- Yehuda Dar and Richard G. Baraniuk. Double double descent: On generalization errors in transfer learning between linear regression tasks. *SIAM Journal on Mathematics of Data Science (SIMODS)*, 4(4):1447–1472, 2022.
- Sjoerd Dirksen and Shahar Mendelson. Non-Gaussian hyperplane tessellations and robust one-bit compressed sensing. *Journal of the European Mathematical Society*, 23(9):2913–2947, 2021.
- Saurabh Garg, Sivaraman Balakrishnan, and Zachary C. Lipton. Domain adaptation under open set label shift. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2022.
- Sivakant Gopi, Praneeth Netrapalli, Prateek Jain, and Aditya Nori. One-bit compressed sensing: Provable support and vector recovery. In *International Conference on Machine Learning (ICML)*, 2013.
- Trevor Hastie, Andrea Montanari, Saharon Rosset, and Ryan J. Tibshirani. Surprises in high-dimensional ridgeless least squares interpolation. *Annals of Statistics*, 50(2):949–986, 2022.
- Daniel Hsu, Sham M. Kakade, John Langford, and Tong Zhang. Multi-label prediction via compressed sensing. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2009.
- Daniel Hsu, Vidya Muthukumar, and Ji Xu. On the proliferation of support vectors in high dimensions. *Journal of Statistical Mechanics: Theory and Experiment*, 2022(11):114011, 2022.
- Ziwei Ji and Matus Telgarsky. The implicit bias of gradient descent on nonseparable data. In *Conference on Learning Theory (COLT)*, pages 1772–1798, 2019.

- Chiraag Kaushik, Andrew D. McRae, Mark Davenport, and Vidya Muthukumar. New equivalences between interpolation and SVMs: Kernels and structured features. *SIAM Journal on Mathematics of Data Science (SIMODS)*, 6(3):761–787, 2024.
- Douglas M. Kline and Victor L. Berardi. Revisiting squared-error and cross-entropy functions for training neural network classifiers. *Neural Computing & Applications*, 14:310–318, 2005.
- Karin Knudson, Rayan Saab, and Rachel Ward. One-bit compressive sensing with norm estimation. *IEEE Transactions on Information Theory*, 62(5):2748–2758, 2016.
- Daniel LeJeune, Jiayu Liu, and Reinhard Heckel. Monotonic risk relationships under distribution shifts for regularized risk minimization. *Journal of Machine Learning Research (JMLR)*, 25:1–37, 2024.
- Yang Liu and Mirella Lapata. Text summarization with pretrained encoders. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2019.
- Yue M. Lu, Mary I. Letey, Jacob A. Zavatone-Veth, Anindita Maiti, and Cengiz Pehlevan. Asymptotic theory of in-context learning by linear attention. *arXiv preprint 2405.11751*, 2024.
- Neil Mallinar, James B. 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.
- Neil Mallinar, Austin Zane, Spencer Frei, and Bin Yu. Minimum-norm interpolation under covariate shift. In *International Conference on Machine Learning (ICML)*, 2024.
- Jose G. Moreno-Torres, Troy Raeder, Rocío Alaiz-Rodríguez, Nitesh V Chawla, and Francisco Herrera. A unifying view on dataset shift in classification. *Pattern Recognition*, 45(1):521–530, 2012.
- Vidya Muthukumar, Adhyayan Narang, Vignesh Subramanian, Mikhail Belkin, Daniel Hsu, and Anant Sahai. Classification vs regression in overparameterized regimes: Does the loss function matter? *Journal of Machine Learning Research (JMLR)*, 22(222):1–69, 2021.
- Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359, 2009.
- Yaniv Plan and Roman Vershynin. Robust 1-bit compressed sensing and sparse logistic regression: A convex programming approach. *IEEE Transactions on Information Theory*, 59(1):482–494, 2012.
- Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: towards real-time object detection with region proposal network. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2015.
- Ryan Rifkin, Gene Yeo, Tomaso Poggio, et al. Regularized least-squares classification. *NATO Science Series: Subseries III: Computer and Systems Sciences*, 190:131–154, 2003.
- Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2015.
- Mark Rudelson and Roman Vershynin. Hanson-wright inequality and sub-gaussian concentration. *Electronic Communications in Probability*, 18(1):1–9, 2013.
- Daniel Soudry, Elad Hoffer, Mor Shpigel Nacson, Suriya Gunasekar, and Nathan Srebro. The implicit bias of gradient descent on separable data. *Journal of Machine Learning Research (JMLR)*, 19(1):1–57, 2018.
- Lawrence Stewart, Francis Bach, Quentin Berthet, and Jean-Philippe Vert. Regression as classification: Influence of task formulation on neural network features. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2023.
- Remi Tachet, Han Zhao, Yu-Xiang Wang, and Geoff Gordon. Domain adaptation with conditional distribution matching and generalized label shift. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2020.
- Nilesh Tripuraneni, Michael I. Jordan, and Chi Jin. On the theory of transfer learning: The importance of task diversity. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2020.
- 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.
- Alexander Tsigler and Peter L. Bartlett. Benign overfitting in ridge regression. *Journal of Machine Learning Research (JMLR)*, 24:1–76, 2023.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you

- need. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2017.
- Roman Vershynin. Introduction to the non-asymptotic analysis of random matrices. In Y. Eldar and G. Kutyniok, editors, *Compressed Sensing, Theory and Applications*, chapter 5. Cambridge University Press, 2012.
- Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge University Press, 2018.
- Martin J Wainwright. Information-theoretic limits on sparsity recovery in the high-dimensional and noisy setting. *IEEE Transactions on Information Theory*, 55(12):5728–5741, 2009.
- Ke Wang and Christos Thrampoulidis. Binary classification of Gaussian mixtures: Abundance of support vectors, benign overfitting, and regularization. *SIAM Journal on Mathematics of Data Science (SIMODS)*, 4:260–284, 2022.
- Ke Wang, Vidya Muthukumar, and Christos Thrampoulidis. Benign overfitting in multiclass classification: All roads lead to interpolation. *IEEE Transactions on Information Theory*, 69:7909–7952, 2023.
- David Wu and Anant Sahai. Precise asymptotic generalization for multiclass classification with overparameterized linear models. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2023.
- Kun Zhang, Bernhard Schölkopf, Krikamol Muandet, and Zhikun Wang. Domain adaptation under target and conditional shift. In *International Conference on Machine Learning (ICML)*, 2013.
- Ruiqi Zhang, Spencer Frei, and Peter Bartlett. Trained transformers learn linear models in-context. *Journal of Machine Learning Research (JMLR)*, 25(1):1–55, 2024.

## Checklist

- For all models and algorithms presented, check if you include:
  - A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes/No/Not Applicable]
  - An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes/No/Not Applicable]
  - (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Yes/No/Not Applicable]
- For any theoretical claim, check if you include:
  - Statements of the full set of assumptions of all theoretical results. [Yes/No/Not Applicable]
  - Complete proofs of all theoretical results. [Yes/No/Not Applicable]
  - Clear explanations of any assumptions. [Yes/No/Not Applicable]

We present our proofs in the Appendix and detail our assumptions in Section 2.
- For all figures and tables that present empirical results, check if you include:
  - The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes/No/Not Applicable]
  - All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes/No/Not Applicable]
  - A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes/No/Not Applicable]
  - A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Yes/No/Not Applicable]

We include the training details, statistics, and computing infrastructure in Section 4 and Appendix I. Our code is available at <https://github.com/tmlabonte/taskshift>.
- If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
  - Citations of the creator if your work uses existing assets. [Yes/No/Not Applicable]
  - The license information of the assets, if applicable. [Yes/No/Not Applicable]
  - New assets either in the supplemental material or as a URL, if applicable. [Yes/No/Not Applicable]



- (d) Information about consent from data providers/curators. [Yes/No/**Not Applicable**]
  - (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Yes/No/**Not Applicable**]
5. If you used crowdsourcing or conducted research with human subjects, check if you include:
- (a) The full text of instructions given to participants and screenshots. [Yes/No/**Not Applicable**]
  - (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Yes/No/**Not Applicable**]
  - (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Yes/No/**Not Applicable**]

# Appendix

## Table of Contents

---

<b>A</b>	<b>Expanded related work</b>	
<b>B</b>	<b>Key lemma: general survival and contamination bounds</b>	
<b>C</b>	<b>Zero-shot task shift in the case of sparse signal</b>	
C.1	Characterization of zero-shot task shift for general covariance	...
C.2	Closed-form expressions for spiked covariance	...
C.3	Closed-form expressions for polynomial decay covariance	...
<b>D</b>	<b>Zero-shot task shift in the case of random signal</b>	
D.1	Task shift error decomposition	...
D.2	Upper and lower bounds on task shift error	...
D.3	Deviation of classification and regression labels	...
D.4	Benign overfitting analysis in our dependent noise model	...
D.5	Tradeoff between regression bias and task shift error	...
<b>E</b>	<b>Zero-shot task shift in the case of dense signal</b>	
E.1	Dense random signal model without simplifying ansatz	...
E.2	Analysis of bias and task shift error via benign overfitting	...
E.3	Concentration of task shift error terms	...
<b>F</b>	<b>Support recovery: analysis beyond survival and contamination</b>	
F.1	Characterization of classification MNI parameters	...
F.2	Spiked and polynomial decay covariance: support inside the top $k^*$ indices	...
F.3	Spiked and polynomial decay covariance: support outside the top $k^*$ indices	...
<b>G</b>	<b>Survival and contamination auxiliary lemmas</b>	
<b>H</b>	<b>Background lemmas</b>	
<b>I</b>	<b>Additional simulations</b>	

---

## A Expanded related work

We organize related work under four verticals.

**Task shift *vis-a-vis* transfer learning.** Our problem formulation differs from the popular *transfer learning* paradigm, which utilizes a pretrained representation on downstream tasks, *e.g.*, via finetuning or knowledge transfer. Numerous works have analyzed the generalization of transfer learning, including for high-dimensional linear regression (Dar and Baraniuk, 2022) and general function classes (Tripuraneni et al., 2020). In the transfer learning literature, improved sample complexity guarantees are often provided as compared to learning each of the tasks from scratch; our few-shot results in Section 4 also have this flavor at a high level. In our *task shift* setting, the ground-truth signal does not change between the tasks — only the nature of the task in terms of its metric, *i.e.*, test loss function, changes. Task shift is particularly compelling when the aim is to shift to a fundamentally harder task, as in this work where we transfer a classification estimator to a regression task. Conversely, recent work formulating regression as a multi-class classification problem showed that despite discretization, the classification loss can aid in feature learning (Stewart et al., 2023). Earlier, multi-label classification problems (*i.e.*, prediction of a vector-valued discrete output) were modeled as binary classification utilizing a sparsity assumption on the labels (Hsu et al., 2009).

**Task shift *vis-a-vis* distribution shift.** Though ultimately very different, task shift shares some similarities with *distribution shift*, wherein the training distribution  $p$  and test distribution  $q$  over the feature space  $\mathcal{X}$  and label space  $\mathcal{Y}$  differ. Task shift is not directly related to the common settings of *covariate shift*, wherein  $p(x) \neq q(x)$  but  $p(y | x) = q(y | x)$ , or *label shift*, wherein  $p(y) \neq q(y)$  but  $p(x | y) = q(x | y)$ . In particular, recent results leveraging benign overfitting and random matrix theory to analyze covariate shift (Tripuraneni et al., 2021; Mallinar et al., 2024; LeJeune et al., 2024) are generally inapplicable in our setting. Task shift is more closely related to *concept drift*, where  $p(x) = q(x)$  but  $p(y | x) \neq q(y | x)$ , though concept drift is typically studied in the context of temporal changes (Moreno-Torres et al., 2012). Task shift also has similarities to generalized settings including the “generalized target shift” of Zhang et al. (2013), in which  $p(y) \neq q(y)$  and  $p(x | y)$  changes with constraints, and the “generalized label shift” of Tachet et al. (2020), in which  $p(y) \neq q(y)$  and  $p(g(x) | y) = q(g(x) | y)$  for some representation  $g$ . While these works assume that  $\mathcal{X}$  and  $\mathcal{Y}$  remain constant during the distribution shift, in task shift we study different or even disjoint label spaces  $\mathcal{Y}_{\text{train}}$  and  $\mathcal{Y}_{\text{test}}$ . Similar models have been proposed, such as Disjoint Label Space Transfer Learning (Chang et al., 2019), wherein  $\mathcal{Y}_{\text{train}}$  and  $\mathcal{Y}_{\text{test}}$  are completely disjoint but share a common representation, and Open Set Label Shift (Garg et al., 2022), wherein  $p(y) \neq q(y)$  and new classes may arrive during test-time as long as  $p(x | y)$  is constant for existing classes. Yet, our task shift formulation goes one step further in that the objective function also changes.

**One-bit compressive sensing.** Our problem formulation for zero-shot estimation is deeply connected to the *one-bit compressive sensing* problem (Boufounos and Baraniuk, 2008). In this setting, the data matrix  $\mathbf{X}$  is designed, the ground-truth signal  $\boldsymbol{\theta}^* \in \mathbb{R}^d$  is unknown but  $t$ -sparse, and the objective is signal recovery from  $n \ll d$  quantized measurements of the form  $\hat{y}_i = \text{sign}(\mathbf{x}_i^\top \boldsymbol{\theta}^*)$ . The most prominent difference between our frameworks is that the focus in one-bit compressive sensing is designing the *optimal estimation procedure* with knowledge of the measurement process. This procedure involves either solving a convex program to maximize the average margin on training data subject to a  $\ell_1$ -norm constraint (Plan and Vershynin, 2012; Awasthi et al., 2016; Chinot et al., 2022) or combinatorial optimization routines (Gopi et al., 2013). In contrast, we have no control over the design of the estimator and assume *de facto* the  $\ell_2$ -inductive bias, which due to the implicit bias of gradient descent is one of the most commonly observed in machine learning (Soudry et al., 2018; Ji and Telgarsky, 2019). Accordingly, while one-bit compressive sensing tends to consider isotropic or near-isotropic ensembles, we consider a gamut of anisotropic ensembles which can be more favorable to the  $\ell_2$ -inductive bias<sup>5</sup>. Moreover, one-bit compressive sensing allows the design of biases which leak information about not only the signal direction but also its magnitude. For example, one can design *known* biases  $\boldsymbol{\tau} \in \mathbb{R}^n$  and modify the measurements to  $\hat{y}_i = \text{sign}(\mathbf{x}_i^\top \boldsymbol{\theta}^* + \tau_i)$  (Knudson et al., 2016; Dirksen and Mendelson, 2021).

<sup>5</sup> While, surprisingly, our few-shot guarantees hold even for the isotropic ensemble, the sample complexity is suboptimal compared to one-bit compressive sensing, requiring  $n = d^{\frac{1}{p}}$  for some  $p > 1$ .

We have no such flexibility in our framework — any error terms that may arise would be of the form of unknown regression or classification label noise, and would therefore worsen our zero-shot lower bounds. Nevertheless, our results in Theorem 8 show that even under the weaker  $\ell_2$ -inductive bias, positive results in the flavor of one-bit compressive sensing are possible: specifically, we can estimate the signal correctly *only if* its total magnitude is known and given to be  $\|\Sigma^{\frac{1}{2}}\theta^*\|_2^2 = \frac{2}{\pi}$ .

**The  $\ell_2$ -inductive bias and benign overfitting.** The minimum  $\ell_2$ -norm interpolator has been shown, for certain “effectively high dimensional” data covariances, to overfit noise in a *benign* manner, meaning that the extra error arising from such interpolation can decay to zero as  $n, d \rightarrow \infty$  (Bartlett et al., 2020; Tsigler and Bartlett, 2023). Despite the possible absence of noise in our setting, we find effective dimension useful — indeed, we propose an interpretation of the difference between regression and classification labels as effective noise. The classical shortcoming with minimum  $\ell_2$ -norm solutions is not overfitting noise, but instead their propensity to *attenuate* signal (Chen et al., 2001); even for 1-sparse  $\theta^* = e_1$  and  $d \gg n$  we would have  $\hat{\theta}_1 \rightarrow 0$  for isotropic data covariance (Hastie et al., 2022). Statistically speaking, minimum  $\ell_2$ -norm solutions suffer from a constant bias and therefore inconsistent regression error even when trained on regression labels. Despite this, Muthukumar et al. (2021); Cao et al. (2021); Wang and Thrampoulidis (2022) showed that for sufficiently anisotropic ensembles, one can achieve *classification consistency on classification labels* when *regression would not be consistent on regression labels*. The key insight, developed primarily for 1-sparse signal and spiked covariance models, was that the relative magnitude of the true feature  $\hat{\theta}_1$  (quantified through a metric called *survival*) is preserved with respect to the total magnitude of the “false” features  $\{\hat{\theta}_j\}_{j=2}^d$  (quantified through a metric called *contamination*).

We substantially develop this insight to show that the relative feature magnitudes can be used to estimate the support of a sparse signal even when neither classification nor regression generalizes — including isotropic covariance, the worst-case data model for minimum  $\ell_2$ -norm solutions — which in turn enables few-shot regression consistency. Our support recovery procedure is also generally applicable; it only requires a diagonal covariance and either the sparsity level  $t$  or the covariance spectrum  $\{\lambda_j\}_{j=1}^d$  to be known.

## B Key lemma: general survival and contamination bounds

In this section, we provide the proof of Lemma 7: our extension of survival and contamination bounds to  $t$ -sparse signal  $\theta^*$  and general covariance matrices  $\Sigma$ . In contrast to the 1-sparse  $\theta^*$  result of Muthukumar et al. (2021), we consider a more general  $t$ -sparse  $\theta^*$  setting. To overcome this technical obstacle, we utilize the leave- $k$ -out technique of Wang et al. (2023) to complete the proof. For bounds on the survival term  $SU_{s_\ell}$ , where  $s_\ell$  is the  $\ell^{th}$  element in  $\mathcal{S}$  for  $\ell \in [t]$ , we show the result when  $\ell = 1$  without loss of generality. Before we begin the proof of Lemma 7, we define the following notation for  $Q_{s_\ell}$  and  $\tilde{Q}_{s_\ell}$ :

$$Q_{s_\ell} := \mathbf{z}_{s_1}^\top \mathbf{A}_{-s_1:s_\ell}^{-1} \hat{\mathbf{y}} \quad (2)$$

$$\tilde{Q}_{s_\ell} := \mathbf{z}_{s_1}^\top \mathbf{A}_{-s_1:s_\ell}^{-1} \mathbf{z}_{s_1}, \quad (3)$$

where we denote shorthand notation  $\mathbf{A} := \mathbf{X}\mathbf{X}^\top$  and the *leave- $\ell$ -out* matrices corresponding to  $\mathbf{A}_{-s_1:s_\ell} := \sum_{j=1, j \neq s_1, \dots, s_\ell}^d \lambda_j \mathbf{z}_j \mathbf{z}_j^\top$ . We also need the following auxiliary lemmas which are generalized from lemmas in Wang et al. (2023). Note that lemmas in Wang et al. (2023) only apply to leave-*top*  $k$ -out matrices, while we generalize them into leave-*discrete*  $t$ -out matrices. Proofs of these lemmas can be found in Appendix G.

**Lemma 15** (Generalization of Lemma 15 in Wang et al. (2023)). *Under Assumptions 2 and 5, for large enough  $n$ , we have*

$$\begin{aligned} Q_{s_t} &\leq \text{sign}(a_{s_1}) \sqrt{\frac{2\beta_{s_1}}{\pi}} \text{Tr}(\mathbf{A}_{-s_1:s_t}^{-1}) + 2c_1 \|\mathbf{A}_{-s_1:s_t}^{-1}\|_2 \cdot n^{\frac{1}{2}+\epsilon}, \\ Q_{s_t} &\geq \text{sign}(a_{s_1}) \sqrt{\frac{2\beta_{s_1}}{\pi}} \text{Tr}(\mathbf{A}_{-s_1:s_t}^{-1}) - 2c_1 \|\mathbf{A}_{-s_1:s_t}^{-1}\|_2 \cdot n^{\frac{1}{2}+\epsilon}, \end{aligned}$$

where  $\beta_{s_\ell} := \frac{\lambda_{s_\ell} \theta_{s_\ell}^{*2}}{\sum_{j \in \mathcal{S}} \lambda_j \theta_j^{*2}} = \frac{a_{s_\ell}^2}{\sum_{j \in \mathcal{S}} a_j^2}$  for  $\ell \in [t]$ , with probability at least  $1 - 2e^{-n^{2\epsilon}}$ .



**Lemma 16** (Generalization of Lemma 16 in Wang et al. (2023)). We have that  $Q_{s_1}$  is tight up to an additive factor in  $Q_{s_t}$  as

$$\left(1 - \text{sign}(a_{s_1}) \frac{c_4 t}{c_3 n^{\frac{1}{2}-\epsilon} - 1}\right) Q_{s_t} \leq Q_{s_1} \leq \left(1 + \text{sign}(a_{s_1}) \frac{c_4 t}{c_3 n^{\frac{1}{2}-\epsilon} - 1}\right) Q_{s_t}$$

with probability at least  $1 - cte^{-n^{2\epsilon}}$ .

**Lemma 17** (Generalization of Lemma 21 in Wang et al. (2023)). Define  $\check{\mathbf{y}}_{s_\ell} := \check{\mathbf{y}}_{s_{\ell-1}} - \frac{\lambda_{s_\ell} \mathbf{z}_{s_\ell}^\top \mathbf{A}_{-s_1:s_\ell}^{-1} \check{\mathbf{y}}_{s_{\ell-1}}}{1 + \lambda_{s_\ell} \mathbf{z}_{s_\ell}^\top \mathbf{A}_{-s_1:s_\ell}^{-1} \mathbf{z}_{s_\ell}} \mathbf{z}_{s_\ell}$  and  $\check{\mathbf{y}}_{s_0} = \hat{\mathbf{y}}$  for  $\ell \in [t]$ . Then,

$$\left| \frac{\lambda_{s_\ell} \mathbf{z}_{s_\ell}^\top \mathbf{A}_{-s_1:s_\ell}^{-1} \check{\mathbf{y}}_{s_{\ell-1}}}{1 + \lambda_{s_\ell} \mathbf{z}_{s_\ell}^\top \mathbf{A}_{-s_1:s_\ell}^{-1} \mathbf{z}_{s_\ell}} \right| \leq c$$

with probability at least  $1 - ct^2 e^{-n^{2\epsilon}}$ .

Now we can prove Lemma 7.

*Proof.* (Lemma 7) We start with the survival upper and lower bound. By the survival definition in Definition 6, we can write  $\text{SU}_{s_1}$  as

$$\begin{aligned} \text{SU}_{s_1} &= \frac{\lambda_{s_1}}{a_{s_1}} \mathbf{z}_{s_1}^\top \mathbf{A}^{-1} \hat{\mathbf{y}} \\ &= \frac{\lambda_{s_1}}{a_{s_1}} \mathbf{z}_{s_1}^\top \left( \mathbf{A}_{-s_1:s_1}^{-1} - \frac{\lambda_{s_1} \mathbf{A}_{-s_1:s_1}^{-1} \mathbf{z}_{s_1} \mathbf{z}_{s_1}^\top \mathbf{A}_{-s_1:s_1}^{-1}}{1 + \lambda_{s_1} \mathbf{z}_{s_1}^\top \mathbf{A}_{-s_1:s_1}^{-1} \mathbf{z}_{s_1}} \right) \hat{\mathbf{y}} \\ &= \frac{\lambda_{s_1}}{a_{s_1}} \frac{\mathbf{z}_{s_1}^\top \mathbf{A}_{-s_1:s_1}^{-1} \hat{\mathbf{y}}}{1 + \lambda_{s_1} \mathbf{z}_{s_1}^\top \mathbf{A}_{-s_1:s_1}^{-1} \mathbf{z}_{s_1}} \\ &= \frac{\lambda_{s_1}}{a_{s_1}} \frac{Q_{s_1}}{1 + \lambda_{s_1} \tilde{Q}_{s_1}} \\ &= \frac{\lambda_{s_1}}{|a_{s_1}|} \frac{\text{sign}(a_{s_1}) Q_{s_1}}{1 + \lambda_{s_1} \tilde{Q}_{s_1}}, \end{aligned} \tag{4}$$

where we apply the Sherman-Morrison-Woodbury identity in the second equality. Next, recall that  $\tilde{Q}_{s_\ell} := \mathbf{z}_{s_1}^\top \mathbf{A}_{-s_1:s_\ell}^{-1} \mathbf{z}_{s_1}$  in Equation (3), and by the Hanson-Wright inequality (Lemma 32), we have

$$\begin{aligned} \tilde{Q}_{s_1} &\leq \text{Tr}(\mathbf{A}_{-s_1:s_1}^{-1}) + c_1 \|\mathbf{A}_{-s_1:s_1}^{-1}\|_2 \cdot n^{\frac{1}{2}+\epsilon} \\ &\leq \text{Tr}(\mathbf{A}_{-(s_1:s_t) \cup [k^*]}^{-1}) + c_1 \|\mathbf{A}_{-(s_1:s_t) \cup [k^*]}^{-1}\|_2 \cdot n^{\frac{1}{2}+\epsilon}, \end{aligned}$$

where the second inequality follows because  $\mathbf{A}_{-s_1:s_1} \succeq \mathbf{A}_{-(s_1:s_t) \cup [k^*]}$  and therefore  $\mathbf{A}_{-s_1:s_1}^{-1} \preceq \mathbf{A}_{-(s_1:s_t) \cup [k^*]}^{-1}$ . On the other hand,

$$\begin{aligned} \tilde{Q}_{s_1} &\geq \text{Tr}(\mathbf{A}_{-s_1:s_1}^{-1}) - c_1 \|\mathbf{A}_{-s_1:s_1}^{-1}\|_2 \cdot n^{\frac{1}{2}+\epsilon} \\ &\geq \left(1 - \frac{c}{n}\right)^{k^*+t-1} \text{Tr}(\mathbf{A}_{-(s_1:s_t) \cup [k^*]}^{-1}) - c_1 \|\mathbf{A}_{-(s_1:s_t) \cup [k^*]}^{-1}\|_2 \cdot n^{\frac{1}{2}+\epsilon}, \end{aligned}$$

where the second inequality follows from the trace lower bound in Lemma 36. Next, we need to derive the bounds for  $\text{sign}(a_{s_1}) Q_{s_1}$  term in Equation (4). Note that we need to adjust the bounds of  $Q_{s_1}$  in Lemma 16 according to the sign of  $a_{s_1}$ . Considering the sign of  $a_{s_1}$ , from Lemma 16, we can have

$$\left(1 - \frac{c_4 t}{c_3 n^{\frac{1}{2}-\epsilon} - 1}\right) \text{sign}(a_{s_1}) Q_{s_t} \leq \text{sign}(a_{s_1}) Q_{s_1} \leq \left(1 + \frac{c_4 t}{c_3 n^{\frac{1}{2}-\epsilon} - 1}\right) \text{sign}(a_{s_1}) Q_{s_t}.$$

From Lemma 15, we can further bound the  $\text{sign}(a_{s_1})Q_{s_t}$  term as

$$\begin{aligned}\text{sign}(a_{s_1})Q_{s_t} &\leq \sqrt{\frac{2\beta_{s_1}}{\pi}} \text{Tr}(\mathbf{A}_{-s_1:s_t}^{-1}) + 2c_1 \|\mathbf{A}_{-s_1:s_t}^{-1}\|_2 \cdot n^{\frac{1}{2}+\epsilon}, \\ \text{sign}(a_{s_1})Q_{s_t} &\geq \sqrt{\frac{2\beta_{s_1}}{\pi}} \text{Tr}(\mathbf{A}_{-s_1:s_t}^{-1}) - 2c_1 \|\mathbf{A}_{-s_1:s_t}^{-1}\|_2 \cdot n^{\frac{1}{2}+\epsilon}.\end{aligned}$$

As a result, we can apply the bounds of  $\tilde{Q}_{s_1}$  and the bounds of  $\text{sign}(a_{s_1})Q_{s_t}$  above to get the upper and lower bound of  $\text{SU}_{s_1}$  from Equation (4). We then obtain the  $\text{SU}_{s_1}$  upper bound as

$$\begin{aligned}(4) &\leq \left(\frac{\lambda_{s_1}}{|a_{s_1}|}\right) \frac{\left(1 + \frac{c_4 t}{c_3 n^{\frac{1}{2}-\epsilon}-1}\right) \left(\sqrt{\frac{2\beta_{s_1}}{\pi}} \text{Tr}(\mathbf{A}_{-s_1:s_t}^{-1}) + 2c_1 \|\mathbf{A}_{-s_1:s_t}^{-1}\|_2 \cdot n^{\frac{1}{2}+\epsilon}\right)}{1 + \lambda_{s_1} \left((1 - \frac{c}{n})^{k^*+t-1} \text{Tr}(\mathbf{A}_{-(s_1:s_t)\cup[k^*]}^{-1}) - c_1 \|\mathbf{A}_{-(s_1:s_t)\cup[k^*]}^{-1}\|_2 \cdot n^{\frac{1}{2}+\epsilon}\right)} \\ &= \sqrt{\frac{2}{\pi \|\Sigma^{\frac{1}{2}}\theta^*\|_2^2}} \left( \frac{\left(1 + \frac{c_4 t}{c_3 n^{\frac{1}{2}-\epsilon}-1}\right) \lambda_{s_1} \left(\text{Tr}(\mathbf{A}_{-s_1:s_t}^{-1}) + c_2 \|\mathbf{A}_{-s_1:s_t}^{-1}\|_2 \cdot n^{\frac{1}{2}+\epsilon}\right)}{1 + \lambda_{s_1} \left((1 - \frac{c}{n})^{k^*+t-1} \text{Tr}(\mathbf{A}_{-(s_1:s_t)\cup[k^*]}^{-1}) - c_1 \|\mathbf{A}_{-(s_1:s_t)\cup[k^*]}^{-1}\|_2 \cdot n^{\frac{1}{2}+\epsilon}\right)} \right), \quad (5)\end{aligned}$$

and the lower bound of  $\text{SU}_{s_1}$  as

$$\begin{aligned}(4) &\geq \left(\frac{\lambda_{s_1}}{|a_{s_1}|}\right) \frac{\left(1 - \frac{c_4 t}{c_3 n^{\frac{1}{2}-\epsilon}-1}\right) \left(\sqrt{\frac{2\beta_{s_1}}{\pi}} \text{Tr}(\mathbf{A}_{-s_1:s_t}^{-1}) - 2c_1 \|\mathbf{A}_{-s_1:s_t}^{-1}\|_2 \cdot n^{\frac{1}{2}+\epsilon}\right)}{1 + \lambda_{s_1} \left(\text{Tr}(\mathbf{A}_{-(s_1:s_t)\cup[k^*]}^{-1}) + c_1 \|\mathbf{A}_{-(s_1:s_t)\cup[k^*]}^{-1}\|_2 \cdot n^{\frac{1}{2}+\epsilon}\right)} \\ &= \sqrt{\frac{2}{\pi \|\Sigma^{\frac{1}{2}}\theta^*\|_2^2}} \left( \frac{\left(1 - \frac{c_4 t}{c_3 n^{\frac{1}{2}-\epsilon}-1}\right) \lambda_{s_1} \left(\text{Tr}(\mathbf{A}_{-s_1:s_t}^{-1}) - c_2 \|\mathbf{A}_{-s_1:s_t}^{-1}\|_2 \cdot n^{\frac{1}{2}+\epsilon}\right)}{1 + \lambda_{s_1} \left(\text{Tr}(\mathbf{A}_{-(s_1:s_t)\cup[k^*]}^{-1}) + c_1 \|\mathbf{A}_{-(s_1:s_t)\cup[k^*]}^{-1}\|_2 \cdot n^{\frac{1}{2}+\epsilon}\right)} \right). \quad (6)\end{aligned}$$

In the equalities, we use  $\frac{1}{|a_{s_1}|} \sqrt{\frac{2\beta_{s_1}}{\pi}} = \frac{1}{|a_{s_1}|} \sqrt{\frac{2a_{s_1}^2}{\pi \sum_{j \in S} a_j^2}} = \sqrt{\frac{2}{\pi \|\Sigma^{\frac{1}{2}}\theta^*\|_2^2}}$ . Next, we need to upper bound  $\text{Tr}(\mathbf{A}_{-s_1:s_t}^{-1})$  and  $\|\mathbf{A}_{-s_1:s_t}^{-1}\|_2$  and lower bound  $\text{Tr}(\mathbf{A}_{-(s_1:s_t)\cup[k^*]}^{-1})$  terms; we achieve this by relating  $\mathbf{A}_{-s_1:s_t}$  to  $\mathbf{A}_{-(s_1:s_t)\cup[k^*]}$ . Lemma 36 provides that

$$\text{Tr}(\mathbf{A}_{-(s_1:s_t)\cup[k^*]}^{-1}) \geq \text{Tr}(\mathbf{A}_{-s_1:s_t}^{-1}) \geq \left(1 - \frac{c}{n}\right)^{k^*} \text{Tr}(\mathbf{A}_{-(s_1:s_t)\cup[k^*]}^{-1}),$$

and we also have  $\mathbf{A}_{-s_1:s_t}^{-1} \preceq \mathbf{A}_{-(s_1:s_t)\cup[k^*]}^{-1}$ . As a result, Equation (5) and Equation (6) now become

$$\text{SU}_{s_1} \leq \sqrt{\frac{2}{\pi \|\Sigma^{\frac{1}{2}}\theta^*\|_2^2}} \left( \frac{\left(1 + \frac{c_4 t}{c_3 n^{\frac{1}{2}-\epsilon}-1}\right) \lambda_{s_1} \left(\text{Tr}(\mathbf{A}_{-(s_1:s_t)\cup[k^*]}^{-1}) + c_2 \|\mathbf{A}_{-(s_1:s_t)\cup[k^*]}^{-1}\|_2 \cdot n^{\frac{1}{2}+\epsilon}\right)}{1 + \lambda_{s_1} \left((1 - \frac{c}{n})^{k^*+t-1} \text{Tr}(\mathbf{A}_{-(s_1:s_t)\cup[k^*]}^{-1}) - c_1 \|\mathbf{A}_{-(s_1:s_t)\cup[k^*]}^{-1}\|_2 \cdot n^{\frac{1}{2}+\epsilon}\right)} \right), \quad (7)$$

$$\text{SU}_{s_1} \geq \sqrt{\frac{2}{\pi \|\Sigma^{\frac{1}{2}}\theta^*\|_2^2}} \left( \frac{\left(1 - \frac{c_4 t}{c_3 n^{\frac{1}{2}-\epsilon}-1}\right) \lambda_{s_1} \left((1 - \frac{c}{n})^{k^*} \text{Tr}(\mathbf{A}_{-(s_1:s_t)\cup[k^*]}^{-1}) - c_2 \|\mathbf{A}_{-(s_1:s_t)\cup[k^*]}^{-1}\|_2 \cdot n^{\frac{1}{2}+\epsilon}\right)}{1 + \lambda_{s_1} \left(\text{Tr}(\mathbf{A}_{-(s_1:s_t)\cup[k^*]}^{-1}) + c_1 \|\mathbf{A}_{-(s_1:s_t)\cup[k^*]}^{-1}\|_2 \cdot n^{\frac{1}{2}+\epsilon}\right)} \right). \quad (8)$$

Next, Lemma 35 shows that all eigenvalues of  $\mathbf{A}_{-(s_1:s_t)\cup[k^*]}^{-1}$  are identical up to a constant such that

$$\frac{1}{cn} \leq \frac{\|\mathbf{A}_{-[k^*]\cup(s_1:s_t)}^{-1}\|_2}{\text{Tr}(\mathbf{A}_{-[k^*]\cup(s_1:s_t)}^{-1})} \leq \frac{c}{n}.$$

By dividing  $\text{Tr} \left( \mathbf{A}_{-[k^*] \cup (s_1:s_t)}^{-1} \right)$  in the numerator and denominator of Equation (7), we can have the upper bound of  $\text{SU}_{s_1}$  as

$$\begin{aligned}
\text{SU}_{s_1} &\leq \sqrt{\frac{2}{\pi \|\Sigma^{\frac{1}{2}} \boldsymbol{\theta}^*\|^2_2}} \left( \frac{\left(1 + \frac{c_4 t}{c_3 n^{\frac{1}{2}-\epsilon}-1}\right) \lambda_{s_1} \left( \text{Tr} \left( \mathbf{A}_{-(s_1:s_t) \cup [k^*]}^{-1} \right) + c_2 \left\| \mathbf{A}_{-(s_1:s_t) \cup [k^*]}^{-1} \right\|_2 \cdot n^{\frac{1}{2}+\epsilon} \right)}{1 + \lambda_{s_1} \left( \left(1 - \frac{c}{n}\right)^{k^*+t-1} \text{Tr} \left( \mathbf{A}_{-(s_1:s_t) \cup [k^*]}^{-1} \right) - c_1 \left\| \mathbf{A}_{-(s_1:s_t) \cup [k^*]}^{-1} \right\|_2 \cdot n^{\frac{1}{2}+\epsilon} \right)} \right) \\
&= \sqrt{\frac{2}{\pi \|\Sigma^{\frac{1}{2}} \boldsymbol{\theta}^*\|^2_2}} \left( \frac{\left(1 + \frac{c_4 t}{c_3 n^{\frac{1}{2}-\epsilon}-1}\right) \lambda_{s_1} \text{Tr} \left( \mathbf{A}_{-(s_1:s_t) \cup [k^*]}^{-1} \right) \left(1 + \frac{c_2 \left\| \mathbf{A}_{-(s_1:s_t) \cup [k^*]}^{-1} \right\|_2 \cdot n^{\frac{1}{2}+\epsilon}}{\text{Tr} \left( \mathbf{A}_{-(s_1:s_t) \cup [k^*]}^{-1} \right)} \right)}{1 + \lambda_{s_1} \text{Tr} \left( \mathbf{A}_{-(s_1:s_t) \cup [k^*]}^{-1} \right) \left( \left(1 - \frac{c}{n}\right)^{k^*+t-1} - \frac{c_1 \left\| \mathbf{A}_{-(s_1:s_t) \cup [k^*]}^{-1} \right\|_2 \cdot n^{\frac{1}{2}+\epsilon}}{\text{Tr} \left( \mathbf{A}_{-(s_1:s_t) \cup [k^*]}^{-1} \right)} \right)} \right) \\
&\leq \sqrt{\frac{2}{\pi \|\Sigma^{\frac{1}{2}} \boldsymbol{\theta}^*\|^2_2}} \left( \frac{\left(1 + \frac{c_4 t}{c_3 n^{\frac{1}{2}-\epsilon}-1}\right) \lambda_{s_1} \text{Tr} \left( \mathbf{A}_{-(s_1:s_t) \cup [k^*]}^{-1} \right) \left(1 + \frac{c_3}{n^{\frac{1}{2}-\epsilon}}\right)}{1 + \lambda_{s_1} \text{Tr} \left( \mathbf{A}_{-(s_1:s_t) \cup [k^*]}^{-1} \right) \left( \left(1 - \frac{c}{n}\right)^{k^*+t-1} - \frac{c_4}{n^{\frac{1}{2}-\epsilon}} \right)} \right) \\
&\leq \sqrt{\frac{2}{\pi \|\Sigma^{\frac{1}{2}} \boldsymbol{\theta}^*\|^2_2}} \left( \frac{\left(1 + \frac{c_5 t}{n^{\frac{1}{2}-\epsilon}}\right) \lambda_{s_1} \text{Tr} \left( \mathbf{A}_{-(s_1:s_t) \cup [k^*]}^{-1} \right)}{1 + \left(1 - \frac{c_6}{n^{\frac{1}{2}-\epsilon}}\right) \lambda_{s_1} \text{Tr} \left( \mathbf{A}_{-(s_1:s_t) \cup [k^*]}^{-1} \right)} \right), \tag{9}
\end{aligned}$$

where we apply the bounds in Lemma 35 in the second inequality, and in the last inequality we introduce some new constants since  $t \ll n^{\frac{1}{2}-\epsilon}$  and  $\epsilon \in (0, \frac{1}{4})$ . We repeat the same derivation for the lower bound of  $\text{SU}_{s_1}$ . From Equation (8), we have

$$\begin{aligned}
\text{SU}_{s_1} &\geq \sqrt{\frac{2}{\pi \|\Sigma^{\frac{1}{2}} \boldsymbol{\theta}^*\|^2_2}} \left( \frac{\left(1 - \frac{c_4 t}{c_3 n^{\frac{1}{2}-\epsilon}-1}\right) \lambda_{s_1} \left( \left(1 - \frac{c}{n}\right)^{k^*} \text{Tr} \left( \mathbf{A}_{-(s_1:s_t) \cup [k^*]}^{-1} \right) - c_2 \left\| \mathbf{A}_{-(s_1:s_t) \cup [k^*]}^{-1} \right\|_2 \cdot n^{\frac{1}{2}+\epsilon} \right)}{1 + \lambda_{s_1} \left( \text{Tr} \left( \mathbf{A}_{-(s_1:s_t) \cup [k^*]}^{-1} \right) + c_1 \left\| \mathbf{A}_{-(s_1:s_t) \cup [k^*]}^{-1} \right\|_2 \cdot n^{\frac{1}{2}+\epsilon} \right)} \right) \\
&= \sqrt{\frac{2}{\pi \|\Sigma^{\frac{1}{2}} \boldsymbol{\theta}^*\|^2_2}} \left( \frac{\left(1 - \frac{c_4 t}{c_3 n^{\frac{1}{2}-\epsilon}-1}\right) \lambda_{s_1} \text{Tr} \left( \mathbf{A}_{-(s_1:s_t) \cup [k^*]}^{-1} \right) \left( \left(1 - \frac{c}{n}\right)^{k^*} - \frac{c_2 \left\| \mathbf{A}_{-(s_1:s_t) \cup [k^*]}^{-1} \right\|_2 \cdot n^{\frac{1}{2}+\epsilon}}{\text{Tr} \left( \mathbf{A}_{-(s_1:s_t) \cup [k^*]}^{-1} \right)} \right)}{1 + \lambda_{s_1} \text{Tr} \left( \mathbf{A}_{-(s_1:s_t) \cup [k^*]}^{-1} \right) \left( 1 + \frac{c_1 \left\| \mathbf{A}_{-(s_1:s_t) \cup [k^*]}^{-1} \right\|_2 \cdot n^{\frac{1}{2}+\epsilon}}{\text{Tr} \left( \mathbf{A}_{-(s_1:s_t) \cup [k^*]}^{-1} \right)} \right)} \right) \\
&\geq \sqrt{\frac{2}{\pi \|\Sigma^{\frac{1}{2}} \boldsymbol{\theta}^*\|^2_2}} \left( \frac{\left(1 - \frac{c_4 t}{c_3 n^{\frac{1}{2}-\epsilon}-1}\right) \lambda_{s_1} \text{Tr} \left( \mathbf{A}_{-(s_1:s_t) \cup [k^*]}^{-1} \right) \left( \left(1 - \frac{c}{n}\right)^{k^*} - \frac{c_3}{n^{\frac{1}{2}-\epsilon}} \right)}{1 + \lambda_{s_1} \text{Tr} \left( \mathbf{A}_{-(s_1:s_t) \cup [k^*]}^{-1} \right) \left( 1 + \frac{c_4}{n^{\frac{1}{2}-\epsilon}} \right)} \right) \\
&\geq \sqrt{\frac{2}{\pi \|\Sigma^{\frac{1}{2}} \boldsymbol{\theta}^*\|^2_2}} \left( \frac{\left(1 - \frac{c_5 t}{n^{\frac{1}{2}-\epsilon}}\right) \lambda_{s_1} \text{Tr} \left( \mathbf{A}_{-(s_1:s_t) \cup [k^*]}^{-1} \right)}{1 + \left(1 + \frac{c_6}{n^{\frac{1}{2}-\epsilon}}\right) \lambda_{s_1} \text{Tr} \left( \mathbf{A}_{-(s_1:s_t) \cup [k^*]}^{-1} \right)} \right). \tag{10}
\end{aligned}$$

Finally, Lemma 35 also implies the bounds of  $\text{Tr} \left( \mathbf{A}_{-(s_1:s_t) \cup [k^*]}^{-1} \right)$  such that

$$\begin{aligned}
\text{Tr} \left( \mathbf{A}_{-(s_1:s_t) \cup [k^*]}^{-1} \right) &= \sum_{i=1}^n \frac{1}{\mu_i \left( \mathbf{A}_{-(s_1:s_t) \cup [k^*]} \right)} \leq \frac{cn}{\tilde{\lambda}_1 r_0 \left( \Sigma_{-(s_1:s_t) \cup [k^*]} \right)}, \\
\text{Tr} \left( \mathbf{A}_{-(s_1:s_t) \cup [k^*]}^{-1} \right) &\geq \frac{n}{c \tilde{\lambda}_1 r_0 \left( \Sigma_{-(s_1:s_t) \cup [k^*]} \right)},
\end{aligned}$$

where we denote  $\{\tilde{\lambda}_j\}_{j=1}^{d-|(s_1:s_t) \cup [k^*]|}$  the diagonal entries of the leave- $t$  and  $k^*$ -out covariance operator  $\Sigma_{-(s_1:s_t) \cup [k^*]}$ . By substituting the bounds of  $\text{Tr} \left( \mathbf{A}_{-(s_1:s_t) \cup [k^*]}^{-1} \right)$  into Equation (9) and Equation (10),

the survival proof is done. Next, we show the convergence of  $\text{SU}_{s_1}$ . From Equation (9), since  $t \ll n^{\frac{1}{2}-\epsilon}$  and  $\epsilon \in (0, \frac{1}{4})$ , for  $n, d \rightarrow \infty$ , we have

$$\begin{aligned} \lim_{n, d \rightarrow \infty} \text{SU}_{s_1} &\leq \lim_{n, d \rightarrow \infty} \sqrt{\frac{2}{\pi \|\Sigma^{\frac{1}{2}} \boldsymbol{\theta}^*\|_2^2}} \left( \frac{\left(1 + \frac{c_5 t}{n^{\frac{1}{2}-\epsilon}}\right) \lambda_{s_1} \text{Tr}(\mathbf{A}_{-(s_1:s_t) \cup [k^*]}^{-1})}{1 + \left(1 - \frac{c_6}{n^{\frac{1}{2}-\epsilon}}\right) \lambda_{s_1} \text{Tr}(\mathbf{A}_{-(s_1:s_t) \cup [k^*]}^{-1})} \right) \\ &= \sqrt{\frac{2}{\pi \|\Sigma^{\frac{1}{2}} \boldsymbol{\theta}^*\|_2^2}} \frac{\lambda_{s_1} \text{Tr}(\mathbf{A}_{-(s_1:s_t) \cup [k^*]}^{-1})}{1 + \lambda_{s_1} \text{Tr}(\mathbf{A}_{-(s_1:s_t) \cup [k^*]}^{-1})}. \end{aligned}$$

Similarly, for the lower bound of  $\text{SU}_{s_1}$ , from Equation (10), we have

$$\begin{aligned} \lim_{n, d \rightarrow \infty} \text{SU}_{s_1} &\geq \lim_{n, d \rightarrow \infty} \sqrt{\frac{2}{\pi \|\Sigma^{\frac{1}{2}} \boldsymbol{\theta}^*\|_2^2}} \left( \frac{\left(1 - \frac{c_5 t}{n^{\frac{1}{2}-\epsilon}}\right) \lambda_{s_1} \text{Tr}(\mathbf{A}_{-(s_1:s_t) \cup [k^*]}^{-1})}{1 + \left(1 + \frac{c_6}{n^{\frac{1}{2}-\epsilon}}\right) \lambda_{s_1} \text{Tr}(\mathbf{A}_{-(s_1:s_t) \cup [k^*]}^{-1})} \right) \\ &= \sqrt{\frac{2}{\pi \|\Sigma^{\frac{1}{2}} \boldsymbol{\theta}^*\|_2^2}} \frac{\lambda_{s_1} \text{Tr}(\mathbf{A}_{-(s_1:s_t) \cup [k^*]}^{-1})}{1 + \lambda_{s_1} \text{Tr}(\mathbf{A}_{-(s_1:s_t) \cup [k^*]}^{-1})}. \end{aligned}$$

Therefore, we can conclude the convergence of  $\text{SU}_{s_1}$  by its matching upper and lower bounds.

Next, we prove the contamination upper bound, and the proof follows the proof idea of Lemma 5 in Wang et al. (2023) closely. We start with the classification MNI with the indices not supported in  $\boldsymbol{\theta}^*$ . For  $j \in S^c$ , we have

$$\begin{aligned} \hat{\theta}_j &= \sqrt{\lambda_j} \mathbf{z}_j^\top \mathbf{A}^{-1} \hat{\mathbf{y}} \\ &= \sqrt{\lambda_j} \mathbf{z}_j^\top \left( \mathbf{A}_{-s_1:s_1}^{-1} - \frac{\lambda_{s_1} \mathbf{A}_{-s_1:s_1}^{-1} \mathbf{z}_{s_1} \mathbf{z}_{s_1}^\top \mathbf{A}_{-s_1:s_1}^{-1}}{1 + \lambda_{s_1} \mathbf{z}_{s_1}^\top \mathbf{A}_{-s_1:s_1}^{-1} \mathbf{z}_{s_1}} \right) \hat{\mathbf{y}} \\ &= \sqrt{\lambda_j} \mathbf{z}_j^\top \mathbf{A}_{-s_1:s_1}^{-1} \left( \underbrace{\hat{\mathbf{y}} - \frac{\lambda_{s_1} \mathbf{z}_{s_1}^\top \mathbf{A}_{-s_1:s_1}^{-1} \hat{\mathbf{y}}}{1 + \lambda_{s_1} \mathbf{z}_{s_1}^\top \mathbf{A}_{-s_1:s_1}^{-1} \mathbf{z}_{s_1}}}_{=: \check{\mathbf{y}}_{s_1}} \right) \\ &= \sqrt{\lambda_j} \mathbf{z}_j^\top \mathbf{A}_{-s_1:s_1}^{-1} \check{\mathbf{y}}_{s_1} \\ &= \sqrt{\lambda_j} \mathbf{z}_j^\top \mathbf{A}_{-s_1:s_t}^{-1} \check{\mathbf{y}}_{s_t}, \end{aligned} \tag{11}$$

where we apply the Sherman-Morrison-Woodbury identity recursively and also denote  $\check{\mathbf{y}}_{s_\ell} := \check{\mathbf{y}}_{s_{\ell-1}} - \frac{\lambda_{s_\ell} \mathbf{z}_{s_\ell}^\top \mathbf{A}_{-s_1:s_\ell}^{-1} \check{\mathbf{y}}_{s_{\ell-1}}}{1 + \lambda_{s_\ell} \mathbf{z}_{s_\ell}^\top \mathbf{A}_{-s_1:s_\ell}^{-1} \mathbf{z}_{s_\ell}}$  and  $\check{\mathbf{y}}_{s_0} = \hat{\mathbf{y}}$  for  $\ell \in [t]$ . Next, we take the square of  $\hat{\theta}_j$  and according to the CN definition in Definition 6, we have  $\hat{\theta}_j^2 = \lambda_j \check{\mathbf{y}}_{s_t}^\top \mathbf{A}_{-s_1:s_t}^{-1} \mathbf{z}_j \mathbf{z}_j^\top \mathbf{A}_{-s_1:s_t}^{-1} \check{\mathbf{y}}_{s_t}$  and therefore

$$\begin{aligned} \text{CN}^2 &= \sum_{j \in S^c} \lambda_j \hat{\theta}_j^2 \\ &= \sum_{j \in S^c} \lambda_j^2 \check{\mathbf{y}}_{s_t}^\top \mathbf{A}_{-s_1:s_t}^{-1} \mathbf{z}_j \mathbf{z}_j^\top \mathbf{A}_{-s_1:s_t}^{-1} \check{\mathbf{y}}_{s_t} \\ &= \underbrace{\check{\mathbf{y}}_{s_t}^\top \mathbf{A}_{-s_1:s_t}^{-1} \left( \sum_{j \in S^c} \lambda_j^2 \mathbf{z}_j \mathbf{z}_j^\top \right) \mathbf{A}_{-s_1:s_t}^{-1} \check{\mathbf{y}}_{s_t}}_{=: \tilde{\mathbf{C}}}. \end{aligned}$$

Next, we apply the triangle inequality such that  $\sqrt{(\mathbf{x} - \mathbf{y})^\top \mathbf{M} (\mathbf{x} - \mathbf{y})} \leq \sqrt{\mathbf{x}^\top \mathbf{M} \mathbf{x}} + \sqrt{\mathbf{y}^\top \mathbf{M} \mathbf{y}}$   $t$  times to



decompose  $\tilde{\mathbf{y}}_{s_t}$  and get

$$\text{CN} = \sqrt{\tilde{\mathbf{y}}_{s_t}^\top \tilde{\mathbf{C}} \tilde{\mathbf{y}}_{s_t}} \leq \sqrt{\hat{\mathbf{y}}^\top \tilde{\mathbf{C}} \hat{\mathbf{y}}} + \sum_{\ell=1}^t \sqrt{\left( \frac{\lambda_{s_\ell} \mathbf{z}_{s_\ell}^\top \mathbf{A}_{-s_1:s_\ell}^{-1} \tilde{\mathbf{y}}_{s_{\ell-1}}}{1 + \lambda_{s_\ell} \mathbf{z}_{s_\ell}^\top \mathbf{A}_{-s_1:s_\ell}^{-1} \mathbf{z}_{s_\ell}} \right)^2 \mathbf{z}_{s_\ell}^\top \tilde{\mathbf{C}} \mathbf{z}_{s_\ell}}. \quad (12)$$

Since  $\mathcal{S} = \{s_1, \dots, s_t\}$ ,  $\hat{\mathbf{y}}$  and all  $\mathbf{z}_{s_\ell}$  are independent to  $\tilde{\mathbf{C}}$ , we can apply the Hanson-Wright inequality in Lemma 32 to obtain

$$\begin{aligned} \hat{\mathbf{y}}^\top \tilde{\mathbf{C}} \hat{\mathbf{y}} &\leq \text{Tr}(\tilde{\mathbf{C}}) \left(1 + \frac{1}{c}\right) \ln(n) \quad \text{and} \\ \mathbf{z}_{s_\ell}^\top \tilde{\mathbf{C}} \mathbf{z}_{s_\ell} &\leq \text{Tr}(\tilde{\mathbf{C}}) \left(1 + \frac{1}{c}\right) \ln(n), \end{aligned}$$

with probability at least  $1 - \frac{1}{n}$ . Substitute these inequalities into Equation (12), we get

$$\begin{aligned} \text{CN} &\leq \left(1 + \sum_{\ell=1}^t \left| \frac{\lambda_{s_\ell} \mathbf{z}_{s_\ell}^\top \mathbf{A}_{-s_1:s_\ell}^{-1} \tilde{\mathbf{y}}_{s_{\ell-1}}}{1 + \lambda_{s_\ell} \mathbf{z}_{s_\ell}^\top \mathbf{A}_{-s_1:s_\ell}^{-1} \mathbf{z}_{s_\ell}} \right| \right) \sqrt{\text{Tr}(\tilde{\mathbf{C}}) \left(1 + \frac{1}{c}\right) \ln(n)} \\ &\leq (1 + tc_2) \sqrt{\text{Tr}(\tilde{\mathbf{C}}) \left(1 + \frac{1}{c}\right) \ln(n)}, \end{aligned}$$

where we apply Lemma 17 in the last inequality. It remains to upper bound  $\text{Tr}(\tilde{\mathbf{C}})$  to complete the proof of the upper bound of CN. Then we can use Lemma 11 in Bartlett et al. (2020) to show

$$\begin{aligned} \text{Tr}(\tilde{\mathbf{C}}) &= \text{Tr} \left( \mathbf{A}_{-s_1:s_t}^{-1} \left( \sum_{j \in \mathcal{S}^c} \lambda_j^2 \mathbf{z}_j \mathbf{z}_j^\top \right) \mathbf{A}_{-s_1:s_t}^{-1} \right) \\ &= \sum_{j \in \mathcal{S}^c} \lambda_j^2 \mathbf{z}_j^\top \mathbf{A}_{-s_1:s_t}^{-2} \mathbf{z}_j \\ &= \sum_{j \leq k^*, j \in \mathcal{S}^c} \lambda_j^2 \mathbf{z}_j^\top \mathbf{A}_{-s_1:s_t}^{-2} \mathbf{z}_j + \sum_{j > k^*, j \in \mathcal{S}^c} \lambda_j^2 \mathbf{z}_j^\top \mathbf{A}_{-s_1:s_t}^{-2} \mathbf{z}_j \\ &\leq c \left( \frac{k^*}{n} + n \frac{\sum_{j=k^*+1, j \in \mathcal{S}^c}^d \lambda_j^2}{\left( \sum_{j=k^*+1, j \in \mathcal{S}^c}^d \lambda_j \right)^2} \right). \end{aligned}$$

The proof of the upper bound is completed.  $\square$

## C Zero-shot task shift in the case of sparse signal

In this section, we provide proofs from Section 3.1 concerning our sparse signal model. In Section C.1, we leverage Lemma 7 to prove the convergence of the regression risk of the classification MNI for general covariances. The results of Section C.1 are left in terms of the inverse leave- $t$ -out Gram matrix; in the following sections, we provide more precise derivations for specific covariance ensembles. In Section C.2, we provide closed-form expressions for the limiting regression risk of the classification MNI under the spiked covariance model. Finally, in Section C.3, we provide corresponding expressions for the more delicate case of the polynomial decay covariance model.

### C.1 Characterization of zero-shot task shift for general covariance

In this section, we provide the proof of Theorem 8.

*Proof.* (Theorem 8)

The excess risk expression of Equation (1) in terms of  $\{\text{SU}_j\}_{j \in \mathcal{S}}$  and  $\text{CN}$  gives us

$$L(\hat{\boldsymbol{\theta}}) = \sum_{j \in \mathcal{S}} a_j^2 (\text{SU}_j - 1)^2 + \text{CN}^2.$$

We now characterize the limiting regression risk  $\lim_{n,d \rightarrow \infty} L(\hat{\boldsymbol{\theta}})$ . Since the covariance matrix is benign and satisfies  $\lim_{n,d \rightarrow \infty} \frac{t^2 \cdot k^* \cdot \ln(n)}{n} = \lim_{n,d \rightarrow \infty} \frac{t^2 \cdot n \cdot \ln(n)}{R_0(\boldsymbol{\Sigma}_{-\mathcal{S}})} = 0$ , we can apply Lemma 7. We then have

$$\lim_{n,d \rightarrow \infty} \text{CN}^2 \leq \lim_{n,d \rightarrow \infty} t^2 \left( \frac{k^*}{n} + \frac{n}{R_0(\boldsymbol{\Sigma}_{-\mathcal{S}})} \right) \ln(n) = 0.$$

Since  $\text{CN}^2 \geq 0$  by definition, we have  $\lim_{n,d \rightarrow \infty} \text{CN}^2 = 0$ . Hence, we have  $\lim_{n,d \rightarrow \infty} L(\hat{\boldsymbol{\theta}}) = \sum_{j \in \mathcal{S}} a_j^2 (\text{SU}_j - 1)^2$ . Then, Lemma 7 tells us that

$$\lim_{n,d \rightarrow \infty} \text{SU}_j = \sqrt{\frac{2}{\pi \|\boldsymbol{\Sigma}^{\frac{1}{2}} \boldsymbol{\theta}^*\|_2^2}} \cdot \frac{\lambda_j \text{Tr}(\mathbf{A}_{-\mathcal{S}}^{-1})}{1 + \lambda_j \text{Tr}(\mathbf{A}_{-\mathcal{S}}^{-1})}.$$

Denoting  $b_j := \lim_{n,d \rightarrow \infty} \lambda_j \text{Tr}(\mathbf{A}_{-\mathcal{S}}^{-1})$  as shorthand and putting it all together, we have

$$\lim_{n,d \rightarrow \infty} L(\hat{\boldsymbol{\theta}}) = \lim_{n,d \rightarrow \infty} \sum_{j \in \mathcal{S}} a_j^2 (\text{SU}_j - 1)^2 = \sum_{j \in \mathcal{S}} a_j^2 \left( \lim_{n,d \rightarrow \infty} \text{SU}_j - 1 \right)^2 = \sum_{j \in \mathcal{S}} a_j^2 \left( \sqrt{\frac{2}{\pi \|\boldsymbol{\Sigma}^{\frac{1}{2}} \boldsymbol{\theta}^*\|_2^2}} \frac{b_j}{1 + b_j} - 1 \right)^2.$$

This completes the proof of the theorem.  $\square$

## C.2 Closed-form expressions for spiked covariance

In this section, we provide the proof of Corollary 9.

*Proof.* (Corollary 9)

First of all, we show that the following limit holds

$$\lim_{n,d \rightarrow \infty} \frac{t^2 \cdot k^* \cdot \ln(n)}{n} = \lim_{n,d \rightarrow \infty} \frac{t^2 \cdot n \cdot \ln(n)}{R_0(\boldsymbol{\Sigma}_{-\mathcal{S}})} = 0.$$

This ensures that the assumption in Theorem 8 is satisfied. By the definition of spiked covariance (Definition 3), we observe that  $k^* = s = n^r$  for  $q < 1 - r$  and  $k^* = 0$  for  $q > 1 - r$ . This implies  $\lim_{n,d \rightarrow \infty} \frac{k^* \cdot \ln(n)}{n} = 0$ . On the other hand, we also have  $R_0(\boldsymbol{\Sigma}_{-\mathcal{S}}) = d - |\bar{\mathcal{S}}| = \frac{n^p}{c}$  which leads to  $\lim_{n,d \rightarrow \infty} \frac{n \cdot \ln(n)}{R_0(\boldsymbol{\Sigma}_{-\mathcal{S}})} = 0$ . Additionally, the corollary assumption guarantees that  $t \ll \min\{\sqrt{\frac{n}{k^* \cdot \ln(n)}}, \sqrt{\frac{n^{p-1}}{\ln(n)}}\}$ . Combining these results, we conclude that  $\lim_{n,d \rightarrow \infty} \frac{t^2 \cdot k^* \cdot \ln(n)}{n} = \lim_{n,d \rightarrow \infty} \frac{t^2 \cdot n \cdot \ln(n)}{R_0(\boldsymbol{\Sigma}_{-\mathcal{S}})} = 0$ , and therefore, Theorem 8 holds.

Next, recall that  $b_j := \lim_{n,d \rightarrow \infty} \lambda_j \text{Tr}(\mathbf{A}_{-\mathcal{S}}^{-1})$ , according to Theorem 8, when  $n, d \rightarrow \infty$ , the limiting regression risk is equal to

$$\begin{aligned} \lim_{n,d \rightarrow \infty} L(\hat{\boldsymbol{\theta}}) &= \sum_{j \in \mathcal{S}} a_j^2 \left( \sqrt{\frac{2}{\pi \|\boldsymbol{\Sigma}^{\frac{1}{2}} \boldsymbol{\theta}^*\|_2^2}} \frac{b_j}{1 + b_j} - 1 \right)^2 \\ &= \sum_{j \leq k^*, j \in \mathcal{S}} a_j^2 \left( \sqrt{\frac{2}{\pi \|\boldsymbol{\Sigma}^{\frac{1}{2}} \boldsymbol{\theta}^*\|_2^2}} \frac{b_j}{1 + b_j} - 1 \right)^2 + \sum_{j > k^*, j \in \mathcal{S}} a_j^2 \left( \sqrt{\frac{2}{\pi \|\boldsymbol{\Sigma}^{\frac{1}{2}} \boldsymbol{\theta}^*\|_2^2}} \frac{b_j}{1 + b_j} - 1 \right)^2, \quad (13) \end{aligned}$$

where we separate the summation into two sections. Therefore, it suffices to characterize  $\frac{b_j}{1+b_j}$  for two sections:  $j \leq k^*$  and  $j > k^*$ . In both sections, we have

$$\frac{b_j}{1+b_j} = \frac{1}{\frac{1}{b_j} + 1} = \lim_{n,d \rightarrow \infty} \frac{1}{\left( \frac{1}{\lambda_j \text{Tr}(\mathbf{A}_{-\bar{\mathcal{S}}}^{-1})} \right) + 1}. \quad (14)$$

Next, since  $r_0(\Sigma_{-\bar{\mathcal{S}}}) \geq bn$ , we can bound  $\text{Tr}(\mathbf{A}_{-\bar{\mathcal{S}}}^{-1})$  according to Lemma 10 in [Bartlett et al. \(2020\)](#) such that

$$\begin{aligned} \text{Tr}(\mathbf{A}_{-\bar{\mathcal{S}}}^{-1}) &= \sum_{i=1}^n \frac{1}{\mu_i(\mathbf{A}_{-\bar{\mathcal{S}}})} \leq \frac{n}{\mu_n(\mathbf{A}_{-\bar{\mathcal{S}}})} \leq \frac{cn}{\bar{\lambda}_1 r_0(\Sigma_{-\bar{\mathcal{S}}})} = \frac{cn}{\sum_{j=k^*+1, j \in \mathcal{S}^c}^d \lambda_j} \\ \text{Tr}(\mathbf{A}_{-\bar{\mathcal{S}}}^{-1}) &= \sum_{i=1}^n \frac{1}{\mu_i(\mathbf{A}_{-\bar{\mathcal{S}}})} \geq \frac{n}{\mu_1(\mathbf{A}_{-\bar{\mathcal{S}}})} \geq \frac{n}{c \bar{\lambda}_1 r_0(\Sigma_{-\bar{\mathcal{S}}})} = \frac{n}{c \sum_{j=k^*+1, j \in \mathcal{S}^c}^d \lambda_j}. \end{aligned}$$

Recall that, in the above, we denoted  $\{\tilde{\lambda}_j\}_{j=1}^{d-|\bar{\mathcal{S}}|}$  the eigenvalues of the leave- $t$  and  $k^*$ -out covariance matrix  $\Sigma_{-\bar{\mathcal{S}}}$ . Therefore, we can upper and lower bound Equation (14) as below:

$$\lim_{n,d \rightarrow \infty} \frac{1}{\left( \frac{c \sum_{k=k^*+1, k \in \mathcal{S}^c}^d \lambda_k}{\lambda_j n} \right) + 1} \leq \frac{b_j}{1+b_j} \leq \lim_{n,d \rightarrow \infty} \frac{1}{\left( \frac{\sum_{k=k^*+1, k \in \mathcal{S}^c}^d \lambda_k}{c \lambda_j n} \right) + 1}. \quad (15)$$

As a result, it suffices to characterize the limit of  $\frac{\sum_{k=k^*+1, k \in \mathcal{S}^c}^d \lambda_k}{\lambda_j n}$  as  $n, d \rightarrow \infty$ . We discuss our evaluation of the spiked covariance ensemble in two cases.

- **Spiked covariance with  $q < 1 - r$  (Definition 3):**

Recall that, for this choice of parameters, the regression MNI would generalize ([Muthukumar et al., 2021](#)). Define  $\bar{k} := |\bar{\mathcal{S}}|$  as shorthand. From Definition 3, we have  $k^* = s$ . Therefore, for  $j \in \mathcal{S} \cap [k^*]$ ,

$$\begin{aligned} \lim_{n,d \rightarrow \infty} \frac{\sum_{k=k^*+1, k \in \mathcal{S}^c}^d \lambda_k}{\lambda_j n} &= \lim_{n,d \rightarrow \infty} \frac{(d - \bar{k}) \frac{(1-a)d}{d-s}}{\frac{ad}{s} \cdot n} \\ &= \lim_{n,d \rightarrow \infty} \frac{(n^p - \bar{k}) \frac{(1-n^{-q})n^p}{(n^p - n^r)}}{n^{(p-q-r+1)}} \\ &= \lim_{n,d \rightarrow \infty} \frac{(n^p - \bar{k}) \frac{(1-n^{-q})}{(1-n^{r-p})}}{n^{(p-q-r+1)}} \\ &= 0, \end{aligned}$$

where the last equality followed because  $1 - q - r > 0$ . As a result, from Equation (15), we have  $\frac{b_j}{1+b_j} = 1$  for  $j \in \mathcal{S} \cap [k^*]$ . On the other hand, for  $j \in \mathcal{S} \cap \{k^* + 1, \dots, d\}$ , we have

$$\lim_{n,d \rightarrow \infty} \frac{\sum_{k=k^*+1, k \in \mathcal{S}^c}^d \lambda_k}{\lambda_j n} = \lim_{n,d \rightarrow \infty} \frac{(d - \bar{k}) \frac{(1-a)d}{d-s}}{\frac{(1-a)d}{d-s} \cdot n} = \lim_{n,d \rightarrow \infty} \frac{(n^p - \bar{k})}{n} = \infty.$$

Therefore, from Equation (15), we have  $\frac{b_j}{1+b_j} = 0$  for this case. Substituting these values back into Equation (13) completes the proof of the corollary for this case.

- **Spiked covariance with  $q > 1 - r$  (Definition 3):**

Recall that, for this choice of parameters, the regression MNI would *not* generalize ([Muthukumar et al.,](#)

2021). In this case, we have  $k^* = 0$ . Define the number of support indices that are contained within the spike as  $t_1 := |\mathcal{S} \cap [s]|$ , and note that  $0 \leq t_1 \leq \min\{s, t\}$ . Therefore, for  $j \in \mathcal{S} \cap [s]$ , we have

$$\begin{aligned} \lim_{n,d \rightarrow \infty} \frac{\sum_{k=1, k \in \mathcal{S}^c}^d \lambda_k}{\lambda_j n} &= \lim_{n,d \rightarrow \infty} \frac{(s - t_1) \frac{ad}{s} + (d - s - (t - t_1)) \frac{(1-a)d}{d-s}}{\frac{ad}{s} \cdot n} \\ &= \lim_{n,d \rightarrow \infty} \frac{(n^r - t_1) n^{(p-q-r)} + (n^p - n^r - (t - t_1)) \frac{(1-n^{-q})n^p}{(n^p - n^r)}}{n^{(p-q-r+1)}} \\ &= \infty, \end{aligned}$$

where the last equality follows because  $1 - q - r < 0$ . As a result, from Equation (15),  $\frac{b_j}{1+b_j} = 0$  for  $j \in \mathcal{S} \cap [s]$ . Similarly, for  $j \in \mathcal{S} \cap \{s+1, \dots, d\}$ , we have

$$\lim_{n,d \rightarrow \infty} \frac{\sum_{k=k^*+1, k \in \mathcal{S}^c}^d \lambda_k}{\lambda_j n} \geq \lim_{n,d \rightarrow \infty} \frac{\sum_{k=k^*+1, k \in \mathcal{S}^c}^d \lambda_k}{\lambda_1 n} = \infty,$$

where the last equality follows from the preceding calculation. Substituting these values back into Equation (13) completes the proof for this case.  $\square$

### C.3 Closed-form expressions for polynomial decay covariance

Next, we provide corresponding expressions for the limiting regression risk for the more delicate case of the polynomial decay covariance ensemble. It is worth noting that, unlike the spiked covariance case—where regression consistency can be achieved when the true signal is confined to the top  $k^*$  components and the signal magnitude is fixed at  $\|\Sigma^{\frac{1}{2}} \theta^*\|_2^2 = \frac{2}{\pi}$  (Corollary 9)—the polynomial decay covariance case requires stricter conditions. Specifically, the true signal must be restricted to the top  $\tilde{k}_1$  components, which depend on both the polynomial decay parameters and the sample size  $n$ .

**Corollary 18.** *Under Assumptions 2 and 5 with  $t \ll \min \left\{ \sqrt{\frac{n}{k^* \cdot \ln(n)}}, \sqrt{\frac{R_0(\Sigma - \bar{s})}{n \cdot \ln(n)}} \right\}$  and polynomial decay covariance (Definition 4),*

- For  $u \in [0, 1)$ ,  $v = 0$ , and  $p \cdot (1 - u) > 1$ , we have

$$\lim_{n,d \rightarrow \infty} L(\hat{\theta}) = \sum_{j \in \mathcal{S}} a_j^2.$$

*In this case, we have  $L(\hat{\theta}) > 0$ , and regression consistency is not possible unless we have zero signal, i.e.  $\theta^* = \mathbf{0}$ .*

- For  $u \in (0, 1)$ ,  $v = 0$ , and  $p \cdot (1 - u) < 1$ , we have

$$\begin{aligned} \lim_{n,d \rightarrow \infty} L(\hat{\theta}) &\leq \sum_{j \leq \tilde{k}_1, j \in \mathcal{S}} a_j^2 \left( \sqrt{\frac{2}{\pi \|\Sigma^{\frac{1}{2}} \theta^*\|_2^2}} - 1 \right)^2 + \bar{C} \sum_{\tilde{k}_1 < j < \tilde{k}_2, j \in \mathcal{S}} a_j^2 + \sum_{\tilde{k}_2 \leq j, j \in \mathcal{S}} a_j^2, \\ \lim_{n,d \rightarrow \infty} L(\hat{\theta}) &\geq \sum_{j \leq \tilde{k}_1, j \in \mathcal{S}} a_j^2 \left( \sqrt{\frac{2}{\pi \|\Sigma^{\frac{1}{2}} \theta^*\|_2^2}} - 1 \right)^2 + \underline{C} \sum_{\tilde{k}_1 < j < \tilde{k}_2, j \in \mathcal{S}} a_j^2 + \sum_{\tilde{k}_2 \leq j, j \in \mathcal{S}} a_j^2, \end{aligned}$$

*we define  $\tilde{k}_1 := \max \left\{ k \geq 0 : k = o \left( n^{\frac{1-p \cdot (1-u)}{u}} \right) \right\}$  and  $\tilde{k}_2 := \min \left\{ k \geq 0 : k = \omega \left( n^{\frac{1-p \cdot (1-u)}{u}} \right) \right\}$ , and*

$$\bar{C} := \max_{\tilde{k}_1 < j < \tilde{k}_2, j \in \mathcal{S}} \left( \sqrt{\frac{2}{\pi \|\Sigma^{\frac{1}{2}} \theta^*\|_2^2}} c_j - 1 \right)^2, \quad \underline{C} := \min_{\tilde{k}_1 < j < \tilde{k}_2, j \in \mathcal{S}} \left( \sqrt{\frac{2}{\pi \|\Sigma^{\frac{1}{2}} \theta^*\|_2^2}} c_j - 1 \right)^2.$$

Recalling  $b_j := \lim_{n,d \rightarrow \infty} \lambda_j \text{Tr}(\mathbf{A}_{-\bar{S}}^{-1})$ , we define  $c_j := \frac{b_j}{1+b_j}$ , and  $\{c_j\}_{\tilde{k}_1 < j < \tilde{k}_2, j \in \mathcal{S}}$  is a non-increasing sequence in  $(0, 1)$ . This implies regression consistency if and only if the signal magnitude is fixed at  $\|\Sigma^{\frac{1}{2}} \boldsymbol{\theta}^*\|_2^2 = \frac{2}{\pi}$  and  $a_j = 0$  for all  $j > \tilde{k}_1, j \in \mathcal{S}$ . Note that benign overfitting of the regression MNI is attained for this choice of parameters (Bartlett et al., 2020).

*Proof.* As in the proof of Corollary 9, we first need to show  $\lim_{n,d \rightarrow \infty} \frac{t^2 \cdot k^* \cdot \ln(n)}{n} = \lim_{n,d \rightarrow \infty} \frac{t^2 \cdot n \cdot \ln(n)}{R_0(\Sigma_{-\bar{S}})} = 0$ , and therefore, Theorem 8 holds. Then, we can write the limit of the risk in Equation (13) as

$$\begin{aligned} \lim_{n,d \rightarrow \infty} L(\hat{\boldsymbol{\theta}}) &= \sum_{j \in \mathcal{S}} a_j^2 \left( \sqrt{\frac{2}{\pi \|\Sigma^{\frac{1}{2}} \boldsymbol{\theta}^*\|_2^2}} \frac{b_j}{1+b_j} - 1 \right)^2 \\ &= \sum_{j \leq \tilde{k}_1, j \in \mathcal{S}} a_j^2 \left( \sqrt{\frac{2}{\pi \|\Sigma^{\frac{1}{2}} \boldsymbol{\theta}^*\|_2^2}} \frac{b_j}{1+b_j} - 1 \right)^2 + \sum_{\tilde{k}_1 < j < \tilde{k}_2, j \in \mathcal{S}} a_j^2 \left( \sqrt{\frac{2}{\pi \|\Sigma^{\frac{1}{2}} \boldsymbol{\theta}^*\|_2^2}} \frac{b_j}{1+b_j} - 1 \right)^2 \\ &\quad + \sum_{\tilde{k}_2 \leq j, j \in \mathcal{S}} a_j^2 \left( \sqrt{\frac{2}{\pi \|\Sigma^{\frac{1}{2}} \boldsymbol{\theta}^*\|_2^2}} \frac{b_j}{1+b_j} - 1 \right)^2, \quad (16) \end{aligned}$$

where we separate the summation into three sections. Next, we can bound the value of  $\frac{b_j}{1+b_j}$  for each section by Equation (15) as

$$\lim_{n,d \rightarrow \infty} \frac{1}{\left( \frac{c \sum_{k=k^*+1, k \in \mathcal{S}^c}^d \lambda_k}{\lambda_j n} \right) + 1} \leq \frac{b_j}{1+b_j} \leq \lim_{n,d \rightarrow \infty} \frac{1}{\left( \frac{\sum_{k=k^*+1, k \in \mathcal{S}^c}^d \lambda_k}{c \lambda_j n} \right) + 1}. \quad (17)$$

Hence, for each  $j$ , it suffices to characterize the limit of  $\frac{\sum_{k=k^*+1, k \in \mathcal{S}^c}^d \lambda_k}{\lambda_j n}$  as  $n, d \rightarrow \infty$  in order to characterize the value of  $\frac{b_j}{1+b_j}$ . The proof follows the idea of Theorem 31 in Bartlett et al. (2020) closely.

• **Polynomial decay covariance with  $u \in [0, 1)$ ,  $v = 0$ ,  $p \cdot (1 - u) > 1$  (Definition 4):**

Firstly, we show that this choice of parameters satisfies the assumption in Theorem 8 such that  $\lim_{n,d \rightarrow \infty} \frac{t^2 \cdot k^* \cdot \ln(n)}{n} = \lim_{n,d \rightarrow \infty} \frac{t^2 \cdot n \cdot \ln(n)}{R_0(\Sigma_{-\bar{S}})} = 0$ . From Definition 4, we have  $\lambda_j = \frac{1}{j^u}$  and we can show  $\sum_{j=1}^d \lambda_j \asymp d^{1-u} = n^{p \cdot (1-u)}$  by showing

$$\sum_{j=1}^d \lambda_j = 1 + \sum_{j=2}^d \frac{1}{j^u} \leq 1 + \int_1^d \frac{1}{x^u} dx = 1 + \frac{x^{1-u}}{1-u} \Big|_1^d = 1 + \frac{d^{1-u} - 1}{1-u} \quad (18a)$$

$$\sum_{j=1}^d \lambda_j = \sum_{j=1}^d \frac{1}{j^u} \geq \sum_{j=1}^d \frac{1}{d^u} = d^{1-u}. \quad (18b)$$

Therefore, we can derive  $r_0(\Sigma) = \frac{\sum_{j=1}^d \lambda_j}{\lambda_1} = \sum_{j=1}^d \lambda_j \geq d^{1-u} = n^{p \cdot (1-u)} > n$ , so  $k^* = 0$ . Next, we need to verify  $\lim_{n,d \rightarrow \infty} \frac{t^2 \cdot n \cdot \ln(n)}{R_0(\Sigma_{-\bar{S}})} = 0$ , and we have

$$\frac{1}{R_0(\Sigma_{-\bar{S}})} = \frac{\sum_{j=1, j \in \mathcal{S}^c}^d \lambda_j^2}{\left( \sum_{j=1, j \in \mathcal{S}^c}^d \lambda_j \right)^2} = \frac{\sum_{j=1, j \in \mathcal{S}^c}^d \frac{1}{j^{2u}}}{\left( \sum_{j=1, j \in \mathcal{S}^c}^d \frac{1}{j^u} \right)^2} \leq \frac{\left( 1 + \int_1^d \frac{1}{x^{2u}} dx \right)}{\left( \sum_{j=1, j \in \mathcal{S}^c}^d \frac{1}{j^u} \right)^2} = \frac{\left( 1 + \int_1^d \frac{1}{x^{2u}} dx \right)}{(d-t)^2 \cdot d^{-2u}}.$$



The integral  $\int_1^d \frac{1}{x^{2u}} dx$  varies for different values of  $u$ , and we have

$$\int_1^d \frac{1}{x^{2u}} dx = \begin{cases} \left. \frac{x^{1-2u}}{1-2u} \right|_1^d = \frac{d^{1-2u}-1}{1-2u}, & \text{for } u \in [0, 0.5), \\ \ln(x) \Big|_1^d = \ln(d), & \text{for } u = 0.5, \\ \left. \frac{x^{1-2u}}{1-2u} \right|_1^d = \frac{1-d^{1-2u}}{2u-1}, & \text{for } u \in (0.5, 1). \end{cases}$$

Therefore, we can upper bound  $\frac{n \ln(n)}{R_0(\Sigma_{-\mathcal{S}})}$  as

$$\frac{n \ln(n)}{R_0(\Sigma_{-\mathcal{S}})} = \begin{cases} \frac{n \ln(n) \left(1 + \frac{d^{1-2u}-1}{1-2u}\right)}{(d-t)^2 \cdot d^{-2u}} \leq \frac{c_1 n \ln(n)}{d} = \frac{c_1 n \ln(n)}{n^p}, & \text{for } u \in [0, 0.5), \\ \frac{n \ln(n)(1+\ln(d))}{(d-t)^2 \cdot d^{-1}} \leq \frac{c_2 n \ln(n) \ln(d)}{d} = \frac{c_2 n \cdot p \ln^2(n)}{n^p}, & \text{for } u = 0.5, \\ \frac{n \ln(n) \left(1 + \frac{1-d^{1-2u}}{2u-1}\right)}{(d-t)^2 \cdot d^{-2u}} \leq \frac{c_3 n \ln(n)}{d^{2-2u}} = \frac{c_3 n \ln(n)}{n^{p(2-2u)}}, & \text{for } u \in (0.5, 1). \end{cases}$$

Note that since  $p \cdot (1-u) > 1$ , the limit of all three cases goes to zero by L'Hôpital's Rule, and we have  $\lim_{n,d \rightarrow \infty} \frac{n \ln(n)}{R_0(\Sigma_{-\mathcal{S}})} = 0$  for  $u \in [0, 1)$ . Combining these results with the corollary assumption

$t \ll \min \left\{ \sqrt{\frac{n}{k^* \cdot \ln(n)}}, \sqrt{\frac{R_0(\Sigma_{-\mathcal{S}})}{n \cdot \ln(n)}} \right\}$ , we have  $\lim_{n,d \rightarrow \infty} \frac{t^2 \cdot k^* \cdot \ln(n)}{n} = \lim_{n,d \rightarrow \infty} \frac{t^2 \cdot n \cdot \ln(n)}{R_0(\Sigma_{-\mathcal{S}})} = 0$  and Theorem 8

holds. Next, we characterize the limit  $\frac{\sum_{k=1, k \in \mathcal{S}^c}^d \lambda_k}{\lambda_j n}$  as  $n, d \rightarrow \infty$  to evaluate  $\frac{b_j}{1+b_j}$ . For  $j \in [d]$ , since  $p \cdot (1-u) > 1$ , we have

$$\lim_{n,d \rightarrow \infty} \frac{\sum_{k=1, k \in \mathcal{S}^c}^d \lambda_k}{\lambda_j n} \geq \lim_{n,d \rightarrow \infty} \frac{d^{(1-u)}}{c \lambda_1 n} = \lim_{n,d \rightarrow \infty} \frac{n^{p(1-u)}}{cn} = \infty.$$

By substituting the value of  $\lim_{n,d \rightarrow \infty} \frac{\sum_{k=1, k \in \mathcal{S}^c}^d \lambda_k}{\lambda_j n}$  into Equation (17), we get  $\frac{b_j}{1+b_j} = 0$  for  $j \in [d]$ . Finally, we substitute  $\frac{b_j}{1+b_j} = 0$  into Equation (16) and complete the proof for this case.

• **Polynomial decay covariance with  $u \in (0, 1)$ ,  $v = 0$ ,  $p \cdot (1-u) < 1$  (Definition 4):**

Recall that, for this choice of parameters, we attain benign overfitting in regression (Bartlett et al., 2020, Theorem 31). This result implies that  $\lim_{n,d \rightarrow \infty} \frac{k^*}{n} = \lim_{n,d \rightarrow \infty} \frac{n}{R_0(\Sigma_{-\mathcal{S}})} = 0$ . However, to ensure that

Theorem 8 holds, we require a stronger condition  $\lim_{n,d \rightarrow \infty} \frac{t^2 \cdot k^* \cdot \ln(n)}{n} = \lim_{n,d \rightarrow \infty} \frac{t^2 \cdot n \cdot \ln(n)}{R_0(\Sigma_{-\mathcal{S}})} = 0$ . We first check the rate of  $k^*$ . For any  $k \in [d-2]$  and  $\lambda_j = j^{-u}$ , we have

$$F(d) - F(k+1) = \int_{k+1}^d x^{-u} dx \leq \sum_{j=k+1}^d \lambda_j \leq \int_k^d x^{-u} dx = F(d) - F(k),$$

where  $F(x) = \frac{x^{1-u}}{1-u}$ . Therefore, we have  $\sum_{j=k+1}^d \lambda_j = \mathcal{O}(d^{1-u})$ . Next, we can calculate the rate of the effective rank  $r_k = \frac{\sum_{j=k+1}^d \lambda_j}{\lambda_{k+1}} = \mathcal{O}(k^u \cdot d^{1-u})$ . According to the definition of  $k^*$  such that  $k^* := \min\{k \geq 0 : r_k(\Sigma) \geq bn\}$ , we have  $k^* = \mathcal{O}(n^{\frac{1}{u}} \cdot d^{\frac{u-1}{u}}) = \mathcal{O}(n^{\frac{1-p \cdot (1-u)}{u}})$ , where  $\frac{1-p \cdot (1-u)}{u} < 1$  since  $p > 1$  and  $p \cdot (1-u) < 1$ . This result implies  $\lim_{n,d \rightarrow \infty} \frac{k^* \cdot \ln(n)}{n} = 0$ . Next, we show that  $\lim_{n,d \rightarrow \infty} \frac{n \cdot \ln(n)}{R_0(\Sigma_{-\mathcal{S}})} = 0$ . Similar to the rate of  $\sum_{j=k+1}^d \lambda_j$ , we have the rate of  $\sum_{j=k+1}^d \lambda_j^2$  as

$$\sum_{j=k+1}^d \lambda_j^2 = \begin{cases} \mathcal{O}(d^{1-2u}), & \text{for } u \in (0, 0.5), \\ \mathcal{O}(\ln(\frac{d}{k})), & \text{for } u = 0.5, \\ \mathcal{O}(k^{1-2u}), & \text{for } u \in (0.5, 1). \end{cases}$$

Therefore, recalling that  $\sum_{j=k+1}^d \lambda_j = \mathcal{O}(d^{1-u})$ , we can calculate  $R_0(\Sigma_{-\mathcal{S}})$  as

$$R_0(\Sigma_{-\mathcal{S}}) = \frac{\left(\sum_{j=k^*+1, j \in \mathcal{S}^c}^d \lambda_j\right)^2}{\sum_{j=k^*+1, j \in \mathcal{S}^c}^d \lambda_j^2} = \begin{cases} \mathcal{O}(d) = \mathcal{O}(n^p), & \text{for } u \in (0, 0.5), \\ \mathcal{O}\left(\frac{d}{\ln\left(\frac{d}{k^*}\right)}\right) = \mathcal{O}\left(\frac{n^p}{\ln(n)}\right), & \text{for } u = 0.5, \\ \mathcal{O}\left(d^{2-2u} \cdot k^{*2u-1}\right) = \mathcal{O}\left(n^{2-\frac{1-p(1-u)}{u}}\right), & \text{for } u \in (0.5, 1). \end{cases}$$

As a result, we can conclude that

$$\frac{n \ln(n)}{R_0(\Sigma_{-\mathcal{S}})} = \begin{cases} \mathcal{O}\left(\frac{\ln(n)}{n^{p-1}}\right), & \text{for } u \in [0, 0.5), \\ \mathcal{O}\left(\frac{\ln^2(n)}{n^{p-1}}\right), & \text{for } u = 0.5, \\ \mathcal{O}\left(\frac{\ln(n)}{n^{1-\frac{1-p(1-u)}{u}}}\right), & \text{for } u \in (0.5, 1). \end{cases}$$

Since  $\frac{1-p(1-u)}{u} < 1$ , the limit of all three cases goes to zero by L'Hôpital's Rule, and we have  $\lim_{n,d \rightarrow \infty} \frac{n \ln(n)}{R_0(\Sigma_{-\mathcal{S}})} = 0$  for  $u \in (0, 1)$ . Combining these results with the corollary assumption  $t \ll \min\left\{\sqrt{\frac{n}{k^* \cdot \ln(n)}}, \sqrt{\frac{R_0(\Sigma_{-\mathcal{S}})}{n \cdot \ln(n)}}\right\}$ , we have  $\lim_{n,d \rightarrow \infty} \frac{t^2 \cdot k^* \cdot \ln(n)}{n} = \lim_{n,d \rightarrow \infty} \frac{t^2 \cdot n \cdot \ln(n)}{R_0(\Sigma_{-\mathcal{S}})} = 0$  and Theorem 8 holds. Next, from Equation (16), we will present our calculation in three sections:  $j \leq \tilde{k}_1$ ,  $\tilde{k}_1 < j < \tilde{k}_2$  and  $\tilde{k}_2 \leq j$ , where we recall  $\tilde{k}_1 := \max\left\{k \geq 0 : k = o\left(n^{\frac{1-p \cdot (1-u)}{u}}\right)\right\}$  and  $\tilde{k}_2 := \min\left\{k \geq 0 : k = \omega\left(n^{\frac{1-p \cdot (1-u)}{u}}\right)\right\}$ . We first consider the case  $j \leq \tilde{k}_1$ , and we have

$$\lim_{n,d \rightarrow \infty} \frac{\sum_{k=k^*+1, k \in \mathcal{S}^c}^d \lambda_k}{\lambda_j n} \leq \lim_{n,d \rightarrow \infty} \frac{\sum_{j=1}^d \lambda_j}{\lambda_{\tilde{k}_1} n} \leq \lim_{n,d \rightarrow \infty} \frac{cd^{1-u}}{\lambda_{\tilde{k}_1} n} = \lim_{n,d \rightarrow \infty} \frac{cn^{p \cdot (1-u)}}{\lambda_{\tilde{k}_1} n} = \lim_{n,d \rightarrow \infty} \frac{c \cdot \tilde{k}_1^u}{n^{1-p \cdot (1-u)}} = 0,$$

where we apply Equation (18a) in the second inequality. Therefore, by Equation (17), we have  $\frac{b_j}{1+b_j} = 1$  for  $j \leq \tilde{k}_1$ . Next, for  $\tilde{k}_1 < j < \tilde{k}_2$ , we have

$$\begin{aligned} \lim_{n,d \rightarrow \infty} \frac{\sum_{k=k^*+1, k \in \mathcal{S}^c}^d \lambda_k}{\lambda_j n} &\leq \lim_{n,d \rightarrow \infty} \frac{\sum_{j=1}^d \lambda_j}{\lambda_{\tilde{k}_2-1} n} \leq \lim_{n,d \rightarrow \infty} \frac{cd^{1-u}}{\lambda_{\tilde{k}_2-1} n} = \lim_{n,d \rightarrow \infty} \frac{c \cdot (\tilde{k}_2 - 1)^u}{n^{1-p \cdot (1-u)}} < \infty, \\ \lim_{n,d \rightarrow \infty} \frac{\sum_{k=k^*+1, k \in \mathcal{S}^c}^d \lambda_k}{\lambda_j n} &\geq \lim_{n,d \rightarrow \infty} \frac{d^{1-u}}{c \lambda_{\tilde{k}_1+1} n} = \lim_{n,d \rightarrow \infty} \frac{(\tilde{k}_1 + 1)^u}{cn^{1-p \cdot (1-u)}} > 0, \end{aligned}$$

by applying  $\sum_{k=k^*+1, k \in \mathcal{S}^c}^d \lambda_k \asymp d^{1-u}$  and the definition of  $\tilde{k}_1$  and  $\tilde{k}_2$ . As a result, according to Equation (17), we can derive  $0 < \frac{b_j}{1+b_j} < 1$  for  $\tilde{k}_1 < j < \tilde{k}_2$ . Finally, for  $\tilde{k}_2 \leq j$ , we have

$$\lim_{n,d \rightarrow \infty} \frac{\sum_{k=k^*+1, k \in \mathcal{S}^c}^d \lambda_k}{\lambda_j n} \geq \lim_{n,d \rightarrow \infty} \frac{d^{1-u}}{c \lambda_{\tilde{k}_2} n} = \lim_{n,d \rightarrow \infty} \frac{\tilde{k}_2^u}{cn^{1-p \cdot (1-u)}} = \infty.$$

Therefore, according to Equation (17), we have  $\frac{b_j}{1+b_j} = 0$  for  $\tilde{k}_2 \leq j$ . Substituting the value of  $\frac{b_j}{1+b_j}$  into Equation (16) for all three cases completes the proof.

□

## D Zero-shot task shift in the case of random signal

In this section, we provide proofs from Section 3.2 concerning our random signal model. In Section D.1 we detail the decomposition of the regression risk of the classification MNI into regression bias and task shift error. In Section D.2, we provide the main proofs of the upper and lower bounds on task shift error in our random signal model. In Section D.3, we bound the deviation of the classification and regression labels. Then, in Section D.4, we adapt the benign overfitting analysis of Bartlett et al. (2020) to our dependent noise setting. Finally, in Section D.5, we prove a fundamental tradeoff between the statistical consistency of regression bias and task shift error.

### D.1 Task shift error decomposition

In this section, we provide the proof of Lemma 10.

*Proof.* (Lemma 10)

Recall that the MNI on regression labels and the MNI on classification labels are defined as

$$\begin{aligned}\tilde{\boldsymbol{\theta}} &:= \arg \min \{ \|\boldsymbol{\theta}\|_2 : \mathbf{X}\boldsymbol{\theta} = \tilde{\mathbf{y}} \}, \\ \hat{\boldsymbol{\theta}} &:= \arg \min \{ \|\boldsymbol{\theta}\|_2 : \text{sign}(\mathbf{X}\boldsymbol{\theta}) = \hat{\mathbf{y}} \},\end{aligned}$$

respectively, and they have closed forms

$$\begin{aligned}\tilde{\boldsymbol{\theta}} &= \mathbf{X}^\top (\mathbf{X}\mathbf{X}^\top)^{-1} \tilde{\mathbf{y}}, \\ \hat{\boldsymbol{\theta}} &= \mathbf{X}^\top (\mathbf{X}\mathbf{X}^\top)^{-1} \hat{\mathbf{y}}.\end{aligned}$$

Now, we have

$$\begin{aligned}L(\hat{\boldsymbol{\theta}}) &= \mathbb{E}_{\mathbf{x}} \left( \mathbf{x}^\top \hat{\boldsymbol{\theta}} - \mathbf{x}^\top \boldsymbol{\theta}^* \right)^2 \\ &= \mathbb{E}_{\mathbf{x}} \left( \left( \mathbf{x}^\top \tilde{\boldsymbol{\theta}} - \mathbf{x}^\top \boldsymbol{\theta}^* \right) + \left( \mathbf{x}^\top \hat{\boldsymbol{\theta}} - \mathbf{x}^\top \tilde{\boldsymbol{\theta}} \right) \right)^2 \\ &= L(\tilde{\boldsymbol{\theta}}) + \mathbb{E}_{\mathbf{x}} \left( \mathbf{x}^\top \hat{\boldsymbol{\theta}} - \mathbf{x}^\top \tilde{\boldsymbol{\theta}} \right)^2 + 2\mathbb{E}_{\mathbf{x}} \left[ \left( \mathbf{x}^\top \tilde{\boldsymbol{\theta}} - \mathbf{x}^\top \boldsymbol{\theta}^* \right) \left( \mathbf{x}^\top \hat{\boldsymbol{\theta}} - \mathbf{x}^\top \tilde{\boldsymbol{\theta}} \right) \right].\end{aligned}$$

We will show that the third term is precisely zero. Write

$$\mathbf{E} := \left( \mathbf{X}^\top (\mathbf{X}\mathbf{X}^\top)^{-1} \mathbf{X} - \mathbf{I} \right) \boldsymbol{\Sigma} \mathbf{X}^\top (\mathbf{X}\mathbf{X}^\top)^{-1}.$$

Substituting  $\tilde{\mathbf{y}} := \mathbf{X}\boldsymbol{\theta}^*$  and the closed-form expressions for the minimum-norm interpolators,

$$\begin{aligned}\mathbb{E}_{\mathbf{x}} \left[ \left( \mathbf{x}^\top \tilde{\boldsymbol{\theta}} - \mathbf{x}^\top \boldsymbol{\theta}^* \right) \left( \mathbf{x}^\top \hat{\boldsymbol{\theta}} - \mathbf{x}^\top \tilde{\boldsymbol{\theta}} \right) \right] \\ &= \mathbb{E}_{\mathbf{x}} \left[ \left( \mathbf{x}^\top \mathbf{X}^\top (\mathbf{X}\mathbf{X}^\top)^{-1} \mathbf{X}\boldsymbol{\theta}^* - \mathbf{x}^\top \boldsymbol{\theta}^* \right) \left( \mathbf{x}^\top \mathbf{X}^\top (\mathbf{X}\mathbf{X}^\top)^{-1} \hat{\mathbf{y}} - \mathbf{x}^\top \mathbf{X}^\top (\mathbf{X}\mathbf{X}^\top)^{-1} \tilde{\mathbf{y}} \right) \right] \\ &= \mathbb{E}_{\mathbf{x}} \left[ \left( \mathbf{x}^\top \left( \mathbf{X}^\top (\mathbf{X}\mathbf{X}^\top)^{-1} \mathbf{X} - \mathbf{I} \right) \boldsymbol{\theta}^* \right) \left( \mathbf{x}^\top \mathbf{X}^\top (\mathbf{X}\mathbf{X}^\top)^{-1} (\hat{\mathbf{y}} - \tilde{\mathbf{y}}) \right) \right] \\ &= \boldsymbol{\theta}^{*\top} \mathbf{E} (\hat{\mathbf{y}} - \tilde{\mathbf{y}}).\end{aligned}$$

Recall that  $\hat{\mathbf{y}} - \tilde{\mathbf{y}} := \mathbf{D}\mathbf{X}\boldsymbol{\theta}^*$  where  $\mathbf{D} := \text{diag}(d_1, \dots, d_n)$  and  $d_i := \frac{\text{sign}(\mathbf{x}^\top \boldsymbol{\theta}^*) - \mathbf{x}^\top \boldsymbol{\theta}^*}{\mathbf{x}^\top \boldsymbol{\theta}^*}$ . Note  $\mathbf{E}\mathbf{D}\mathbf{X}$  is nilpotent: we have  $\mathbf{X}\mathbf{E} = \mathbf{0}$ , so  $(\mathbf{E}\mathbf{D}\mathbf{X})^2 = \mathbf{0}$ , and hence all eigenvalues of  $\mathbf{E}\mathbf{D}\mathbf{X}$  are zero. Therefore,

$$\boldsymbol{\theta}^{*\top} \mathbf{E} (\hat{\mathbf{y}} - \tilde{\mathbf{y}}) = \boldsymbol{\theta}^{*\top} \mathbf{E}\mathbf{D}\mathbf{X}\boldsymbol{\theta}^* = 0,$$

which completes the proof.  $\square$

## D.2 Upper and lower bounds on task shift error

In this section, we provide the proof of Theorem 12. First, let us define

$$\mathbf{C} := (\mathbf{X}\mathbf{X}^\top)^{-1} \mathbf{X}\Sigma\mathbf{X}^\top (\mathbf{X}\mathbf{X}^\top)^{-1}.$$

Recall also that  $\hat{\mathbf{y}} - \tilde{\mathbf{y}} := \mathbf{D}\mathbf{X}\boldsymbol{\theta}^*$  where  $\mathbf{D} := \text{diag}(d_1, \dots, d_n)$  and  $d_i := \frac{\text{sign}(\mathbf{x}^\top \boldsymbol{\theta}^*) - \mathbf{x}^\top \boldsymbol{\theta}^*}{\mathbf{x}^\top \boldsymbol{\theta}^*}$ .

We now introduce several lemmas. The first lemma upper bounds the deviation of the classification and regression labels. The proof is in Appendix D.3.

**Lemma 19.** *For any  $\Sigma$  and  $\boldsymbol{\theta}^*$ , there exists a constant  $c > 1$  such that*

$$\|\hat{\mathbf{y}} - \tilde{\mathbf{y}}\|_2^2 \leq cn \|\Sigma^{\frac{1}{2}} \boldsymbol{\theta}^*\|_2^2$$

with probability at least  $1 - e^{-\frac{n}{c}}$ .

The remaining lemmas bound traces involving  $\mathbf{C}$  and  $\mathbf{X}$  in high probability. The first of these is a restatement of the main technical result of Bartlett et al. (2020).

**Lemma 20.** *For any  $\Sigma$  and  $\boldsymbol{\theta}^*$ , there exist constants  $c, c_1 \geq 1$  such that the following hold. If  $k^* < \frac{n}{c_1}$ , we have*

$$\frac{1}{c} \left( \frac{k^*}{n} + \frac{n}{R_{k^*}(\Sigma)} \right) \leq \text{Tr}(\mathbf{C}) \leq c \left( \frac{k^*}{n} + \frac{n}{R_{k^*}(\Sigma)} \right)$$

with probability at least  $1 - 17e^{-\frac{n}{c}}$ . On the other hand, if  $k^* \geq \frac{n}{c_1}$ , we have

$$\text{Tr}(\mathbf{C}) \geq \frac{1}{c}$$

with probability at least  $1 - 10e^{-\frac{n}{c}}$ .

The final lemma characterizes the key additional term arising in our dependent noise model. The proof is in Appendix D.4.

**Lemma 21.** *For any  $\Sigma$  and  $\boldsymbol{\theta}^*$ , there exist constants  $c, c_1 \geq 1$  such that the following hold. If  $k^* < \frac{n}{c_1}$ , we have*

$$\frac{1}{c} \left( \sum_{j=1}^{k^*} \lambda_j + \frac{n}{R_{k^*}(\Sigma)} \sum_{j=k^*+1}^d \lambda_j \right) \leq \text{Tr}(\mathbf{C}\mathbf{X}\mathbf{X}^\top) \leq c \left( \sum_{j=1}^{k^*} \lambda_j + \frac{n}{R_{k^*}(\Sigma)} \sum_{j=k^*+1}^d \lambda_j \right)$$

with probability at least  $1 - 14e^{-\frac{n}{c}}$ . On the other hand, if  $k^* \geq \frac{n}{c_1}$ , we have

$$\text{Tr}(\mathbf{C}\mathbf{X}\mathbf{X}^\top) \geq \frac{1}{c}$$

with probability at least  $1 - 10e^{-\frac{n}{c}}$ .

We are now ready to prove Theorem 12.

*Proof.* (Theorem 12)

We begin with the upper bound. Substituting the closed-form expressions for the minimum-norm interpolators,

$$\begin{aligned} \mathbb{E}_{\mathbf{x}} \left( \mathbf{x}^\top \hat{\boldsymbol{\theta}} - \mathbf{x}^\top \tilde{\boldsymbol{\theta}} \right)^2 &= \mathbb{E}_{\mathbf{x}} \left( \mathbf{x}^\top \mathbf{X}^\top (\mathbf{X}\mathbf{X}^\top)^{-1} \hat{\mathbf{y}} - \mathbf{x}^\top \mathbf{X}^\top (\mathbf{X}\mathbf{X}^\top)^{-1} \tilde{\mathbf{y}} \right)^2 \\ &= \mathbb{E}_{\mathbf{x}} \left( \mathbf{x}^\top \mathbf{X}^\top (\mathbf{X}\mathbf{X}^\top)^{-1} (\hat{\mathbf{y}} - \tilde{\mathbf{y}}) \right)^2 \\ &= (\hat{\mathbf{y}} - \tilde{\mathbf{y}})^\top \mathbf{C} (\hat{\mathbf{y}} - \tilde{\mathbf{y}}). \end{aligned} \tag{19}$$

By definition of operator norm and trace,

$$(\hat{\mathbf{y}} - \tilde{\mathbf{y}})^\top \mathbf{C}(\hat{\mathbf{y}} - \tilde{\mathbf{y}}) \leq \|\mathbf{C}\| \|\hat{\mathbf{y}} - \tilde{\mathbf{y}}\|_2^2 \leq \text{Tr}(\mathbf{C}) \|\hat{\mathbf{y}} - \tilde{\mathbf{y}}\|_2^2.$$

Since  $k^* < \frac{n}{c_1}$ , by Lemma 20, there exists a constant  $c_2 \geq 1$  such that

$$\text{Tr}(\mathbf{C}) \leq c_2 \left( \frac{k^*}{n} + \frac{n}{R_k^*(\boldsymbol{\Sigma})} \right) \quad (20)$$

with probability at least  $1 - 17e^{-\frac{n}{c_2}}$ . Moreover, by Lemma 19, there exists a constant  $c_3 \geq 1$  such that

$$\|\hat{\mathbf{y}} - \tilde{\mathbf{y}}\|_2^2 \leq c_3 n \|\boldsymbol{\Sigma}^{\frac{1}{2}} \boldsymbol{\theta}^*\|_2^2 \quad (21)$$

with probability at least  $1 - e^{-\frac{n}{c_3}}$ . Combining Equations 20 and 21 with a union bound completes the upper bound.

Now we will prove the lower bound. Starting from Equation (19) and using the assumption that  $\mathbf{D} = \alpha \mathbf{I}$ ,

$$\begin{aligned} \mathbb{E}_{\boldsymbol{\theta}^*, \mathbf{x}} \left( \mathbf{x}^\top \hat{\boldsymbol{\theta}} - \mathbf{x}^\top \tilde{\boldsymbol{\theta}} \right)^2 &= \mathbb{E}_{\boldsymbol{\theta}^*} (\hat{\mathbf{y}} - \tilde{\mathbf{y}})^\top \mathbf{C}(\hat{\mathbf{y}} - \tilde{\mathbf{y}}) \\ &= \mathbb{E}_{\boldsymbol{\theta}^*} \boldsymbol{\theta}^{*\top} \mathbf{X}^\top \mathbf{D} \mathbf{C} \mathbf{D} \mathbf{X} \boldsymbol{\theta}^* \\ &= \alpha \mathbb{E}_{\boldsymbol{\theta}^*} \boldsymbol{\theta}^{*\top} \mathbf{X}^\top \mathbf{C} \mathbf{X} \boldsymbol{\theta}^*. \end{aligned}$$

By assumption,  $\mathbb{E}_{\boldsymbol{\theta}^*} \boldsymbol{\theta}^* \boldsymbol{\theta}^{*\top} \geq \sigma^2 \mathbf{I}$ . Using the cyclic and linear properties of trace,

$$\alpha \mathbb{E}_{\boldsymbol{\theta}^*} \boldsymbol{\theta}^{*\top} \mathbf{X}^\top \mathbf{C} \mathbf{X} \boldsymbol{\theta}^* = \alpha \text{Tr}(\mathbf{X}^\top \mathbf{C} \mathbf{X} \mathbb{E}_{\boldsymbol{\theta}^*} \boldsymbol{\theta}^* \boldsymbol{\theta}^{*\top}) \geq \alpha \sigma^2 \text{Tr}(\mathbf{X}^\top \mathbf{C} \mathbf{X}). \quad (22)$$

If  $k^* < \frac{n}{c_1}$ , then by Lemma 21 there exists a constant  $c_4 \geq 1$  such that

$$\text{Tr}(\mathbf{X}^\top \mathbf{C} \mathbf{X}) \geq \frac{1}{c_4} \left( \sum_{j=1}^{k^*} \lambda_j + \frac{n}{R_{k^*}^*(\boldsymbol{\Sigma})} \sum_{j=k^*+1}^d \lambda_j \right) \quad (23)$$

with probability at least  $1 - 14e^{-\frac{n}{c_4}}$ . On the other hand, if  $k^* > \frac{n}{c_1}$ , then by Lemma 21,

$$\text{Tr}(\mathbf{X}^\top \mathbf{C} \mathbf{X}) \geq \frac{1}{c_4} \quad (24)$$

with probability at least  $1 - 10e^{-\frac{n}{c_4}}$ . Substituting Equations 23 and 24 into Equation (22) and choosing  $c = \max(c_2 c_3, c_4)$  completes the proof.  $\square$

### D.3 Deviation of classification and regression labels

In this section, we provide the proof of Lemma 19.

*Proof.* (Lemma 19)

First, note that

$$\begin{aligned} \|\hat{\mathbf{y}} - \tilde{\mathbf{y}}\|_2^2 &= \|\hat{\mathbf{y}}\|_2^2 + \|\tilde{\mathbf{y}}\|_2^2 - 2 \sum_{j=1}^n \tilde{y}_j \hat{y}_j \\ &= n + \|\tilde{\mathbf{y}}\|_2^2 - 2 \|\tilde{\mathbf{y}}\|_1. \end{aligned}$$

Write  $\psi^2 := \|\boldsymbol{\Sigma}^{\frac{1}{2}} \boldsymbol{\theta}^*\|_2^2$  so that  $\tilde{y}_i \sim \mathcal{N}(0, \psi^2)$  for all  $1 \leq i \leq n$ . In particular,

$$\mathbb{E}_{\mathbf{x}} \|\hat{\mathbf{y}} - \tilde{\mathbf{y}}\|_2^2 = n + n\psi^2 - \sqrt{\frac{8}{\pi}} n\psi. \quad (25)$$



For  $\gamma > 1$ , define the sub-Gaussian and sub-exponential norms of a random variable  $X$  by

$$\begin{aligned}\|X\|_{\psi_2} &:= \inf \{t > 0 : \mathbb{E}_X \exp(X^2/t^2) < \gamma\} \quad \text{and} \\ \|X\|_{\psi_1} &:= \inf \{t > 0 : \mathbb{E}_X \exp(|X|/t) < \gamma\},\end{aligned}$$

respectively.

By definition,  $\tilde{y}_i$  is sub-Gaussian for any  $i$  with  $\|\tilde{y}_i\|_{\psi_2} = \xi\psi$  where  $\xi > 0$ . Moreover, because  $\tilde{y}_i$  is symmetric,  $\hat{y}_i$  is Bernoulli and therefore sub-Gaussian with  $\|\hat{y}_i\|_{\psi_2} = \xi$ . Combining terms,  $\hat{y}_i - \tilde{y}_i$  is sub-Gaussian with  $\|\hat{y}_i - \tilde{y}_i\|_{\psi_2} = \xi + \xi\psi$ , and  $(\hat{y}_i - \tilde{y}_i)^2$  is sub-exponential with  $\|(\hat{y}_i - \tilde{y}_i)^2\|_{\psi_1} \leq (\xi + \xi\psi)^2$ .

By Bernstein's inequality, there exists a constant  $c_1 > 0$  such that for any  $t \geq 0$ ,

$$\mathbb{P}_{\mathbf{X}} \left( \|\hat{\mathbf{y}} - \tilde{\mathbf{y}}\|_2^2 - \mathbb{E}_{\mathbf{X}} \|\hat{\mathbf{y}} - \tilde{\mathbf{y}}\|_2^2 \geq t \right) \leq \exp \left( -c_1 \min \left( \frac{t^2}{n(\xi + \xi\psi)^4}, \frac{t}{(\xi + \xi\psi)^2} \right) \right).$$

Hence,

$$\|\hat{\mathbf{y}} - \tilde{\mathbf{y}}\|_2^2 \leq \mathbb{E}_{\mathbf{X}} \|\hat{\mathbf{y}} - \tilde{\mathbf{y}}\|_2^2 + \sqrt{\frac{n}{c_1}} \max \left( \sqrt{\frac{n}{c_1}}, \sqrt{n} \right) (\xi + \xi\psi)^2 \quad (26)$$

with probability at least  $1 - e^{-n}$ . Defining  $c_2 := \max(c_1^{-1}, c_1^{-1/2})$  and using  $(a + b)^2 \leq 2(a^2 + b^2)$  for any  $a, b \in \mathbb{R}$ , we have

$$\begin{aligned}\sqrt{\frac{n}{c_1}} \max \left( \sqrt{\frac{n}{c_1}}, \sqrt{n} \right) (\xi + \xi\psi)^2 &\leq c_2 n (\xi + \xi\psi)^2 \\ &\leq 2c_2 n \xi^2 (1 + \psi^2).\end{aligned} \quad (27)$$

Substituting Equations 25 and 27 into Equation (26) and dropping the negative term, we have

$$\begin{aligned}\|\hat{\mathbf{y}} - \tilde{\mathbf{y}}\|_2^2 &\leq n + n\psi^2 + 2c_2 n \xi^2 (1 + \psi^2) \\ &\leq c_2 (1 + 2\xi^2) (1 + \psi^2) n.\end{aligned}$$

Treating  $\psi^2$  as a constant, there exists a constant  $c_3 \geq 1$  such that  $1 + \psi^2 \leq c_3 \psi^2$ . Hence,

$$\|\hat{\mathbf{y}} - \tilde{\mathbf{y}}\|_2^2 \leq c_2 c_3 (1 + 2\xi^2) n \psi^2.$$

Choosing  $c = c_2 c_3$  completes the proof.  $\square$

#### D.4 Benign overfitting analysis in our dependent noise model

In this section, we provide the proof of Lemma 21. First, we write a convenient representation of the data matrix  $\mathbf{X}$ . Let  $\mathbf{Z} \in \mathbb{R}^{n \times d}$  have *i.i.d.* standard Gaussian elements, then  $\mathbf{X} \stackrel{d}{=} \mathbf{Z} \Sigma^{\frac{1}{2}}$ . The matrix  $\mathbf{Z}$  is known as the *whitened* data matrix. We write  $\mathbf{z}_j \in \mathbb{R}^n$  to denote the  $j^{\text{th}}$  column of  $\mathbf{Z}$ .

Using this notation, write

$$\mathbf{A} := \mathbf{Z} \Sigma \mathbf{Z}^\top = \sum_{j=1}^d \lambda_j \mathbf{z}_j \mathbf{z}_j^\top,$$

and similarly,

$$\mathbf{Z} \Sigma^2 \mathbf{Z}^\top = \sum_{j=1}^d \lambda_j^2 \mathbf{z}_j \mathbf{z}_j^\top.$$

Note that  $\mathbf{X} \mathbf{X}^\top \stackrel{d}{=} \mathbf{A}$ . Finally, let  $\mathbf{A}_{-k} := \sum_{j \neq k} \lambda_j \mathbf{z}_j \mathbf{z}_j^\top$  denote the *leave-one-out* Gram matrix for some  $1 \leq k \leq d$ .

We will use the following lemma in the proof. It is a short consequence of the Sherman-Woodbury-Morrison identity applied to a matrix-vector product.

**Lemma 22.** Suppose  $\mathbf{A}$  and  $\mathbf{z}_j$  are defined as above for some  $1 \leq j \leq d$ . Then,

$$\mathbf{A}^{-1} \mathbf{z}_j = \frac{\mathbf{A}_{-j}^{-1} \mathbf{z}_j}{1 + \lambda_j \mathbf{z}_j^\top \mathbf{A}_{-j}^{-1} \mathbf{z}_j}.$$

*Proof.* By the Sherman-Woodbury-Morrison identity,

$$\mathbf{A}^{-1} = \mathbf{A}_{-j}^{-1} - \frac{\lambda_j \mathbf{A}_{-j}^{-1} \mathbf{z}_j \mathbf{z}_j^\top \mathbf{A}_{-j}^{-1}}{1 + \lambda_j \mathbf{z}_j^\top \mathbf{A}_{-j}^{-1} \mathbf{z}_j}.$$

Multiplying by  $\mathbf{z}_j$  on the right we obtain

$$\begin{aligned} \mathbf{A}^{-1} \mathbf{z}_j &= \mathbf{A}_{-j}^{-1} \mathbf{z}_j - \frac{\lambda_j \mathbf{A}_{-j}^{-1} \mathbf{z}_j \mathbf{z}_j^\top \mathbf{A}_{-j}^{-1} \mathbf{z}_j}{1 + \lambda_j \mathbf{z}_j^\top \mathbf{A}_{-j}^{-1} \mathbf{z}_j} \\ &= \frac{\mathbf{A}_{-j}^{-1} \mathbf{z}_j (1 + \lambda_j \mathbf{z}_j^\top \mathbf{A}_{-j}^{-1} \mathbf{z}_j) - \lambda_j \mathbf{A}_{-j}^{-1} \mathbf{z}_j \mathbf{z}_j^\top \mathbf{A}_{-j}^{-1} \mathbf{z}_j}{1 + \lambda_j \mathbf{z}_j^\top \mathbf{A}_{-j}^{-1} \mathbf{z}_j} \\ &= \frac{\mathbf{A}_{-j}^{-1} \mathbf{z}_j}{1 + \lambda_j \mathbf{z}_j^\top \mathbf{A}_{-j}^{-1} \mathbf{z}_j}, \end{aligned}$$

as desired.  $\square$

Now, we begin the main proof of this section.

*Proof.* (Lemma 21)

Substituting the whitened data matrix,

$$\begin{aligned} \text{Tr}(\mathbf{X}^\top \mathbf{C} \mathbf{X}) &= \text{Tr}(\mathbf{X} \mathbf{\Sigma} \mathbf{X}^\top (\mathbf{X} \mathbf{X}^\top)^{-1}) \\ &\stackrel{\text{d}}{=} \text{Tr}(\mathbf{A}^{-1} \mathbf{Z} \mathbf{\Sigma}^2 \mathbf{Z}^\top) \\ &= \sum_{j=1}^d \lambda_j^2 \mathbf{z}_j^\top \mathbf{A}^{-1} \mathbf{z}_j. \end{aligned} \tag{28}$$

It is instructive to compare this term with Lemma 8 of Bartlett et al. (2020). In contrast to their formulation, we have one fewer  $\mathbf{A}^{-1}$  in the center of the expression due to our dependent noise model. The remainder of the proof utilizes a leave-one-out technique similarly to Bartlett et al. (2020), except the  $j^{\text{th}}$  summand carries an additional  $n\lambda_j$ . Applying Lemma 22,

$$\sum_{j=1}^d \lambda_j^2 \mathbf{z}_j^\top \mathbf{A}^{-1} \mathbf{z}_j = \sum_{j=1}^{k^*} \frac{\lambda_j^2 \mathbf{z}_j^\top \mathbf{A}_{-j}^{-1} \mathbf{z}_j}{1 + \lambda_j \mathbf{z}_j^\top \mathbf{A}_{-j}^{-1} \mathbf{z}_j} + \sum_{j=k^*+1}^d \lambda_j^2 \mathbf{z}_j^\top \mathbf{A}^{-1} \mathbf{z}_j.$$

Clearly, the first term is at most  $\sum_{j=1}^{k^*} \lambda_j$ . For the second term, we have

$$\sum_{j=k^*+1}^d \lambda_j^2 \mathbf{z}_j^\top \mathbf{A}^{-1} \mathbf{z}_j \leq \frac{\sum_{j=k^*+1}^d \lambda_j^2 \|\mathbf{z}_j\|_2^2}{\mu_n(\mathbf{A})}.$$

By Lemma 10 of Bartlett et al. (2020), since  $k^* < \frac{n}{c_1}$ , there exists a constant  $c_2 \geq 1$  such that  $\mu_n(\mathbf{A}) \geq \frac{\lambda_{k^*+1} r_{k^*}(\mathbf{\Sigma})}{c_2}$  with probability at least  $1 - 2e^{-\frac{n}{c_2}}$ . Moreover, by Lemma 12 of Bartlett et al. (2020), there exists a constant  $c_3 \geq 1$  such that

$$\sum_{j=k^*+1}^d \lambda_j^2 \|\mathbf{z}_j\|_2^2 \leq c_3 n \sum_{j=k^*+1}^d \lambda_j^2$$

with probability at least  $1 - 2e^{-\frac{n}{c_3}}$ . By a union bound,

$$\begin{aligned} \text{Tr}(\mathbf{X}^\top \mathbf{C} \mathbf{X}) &\leq \sum_{j=1}^{k^*} \lambda_j + \frac{c_2 c_3 n \sum_{j=k^*+1}^d \lambda_j^2}{\lambda_{k^*+1} r_{k^*}(\boldsymbol{\Sigma})} \\ &\leq c_2 c_3 \left( \sum_{j=1}^{k^*} \lambda_j + \frac{n}{R_{k^*}(\boldsymbol{\Sigma})} \sum_{j=k^*+1}^d \lambda_j \right) \end{aligned}$$

with probability at least  $1 - 4e^{-\frac{n}{c_2 c_3}}$ , which completes the upper bound.

We now prove the lower bound. Beginning from Equation (28) and applying Lemma 22,

$$\sum_{j=1}^d \lambda_j^2 \mathbf{z}_j^\top \mathbf{A}^{-1} \mathbf{z}_j = \sum_{j=1}^d \frac{\lambda_j^2 \mathbf{z}_j^\top \mathbf{A}_{-j}^{-1} \mathbf{z}_j}{1 + \lambda_j \mathbf{z}_j^\top \mathbf{A}_{-j}^{-1} \mathbf{z}_j}.$$

By Lemma 14 of Bartlett et al. (2020), for any  $k < \frac{n}{c_1}$ , there exists a constant  $c_4 \geq 1$  such that for any  $1 \leq j \leq d$ ,

$$1 + \lambda_j \mathbf{z}_j^\top \mathbf{A}_{-j}^{-1} \mathbf{z}_j \leq c_4 \lambda_j \mathbf{z}_j^\top \mathbf{A}_{-j}^{-1} \mathbf{z}_j \left( \frac{n \lambda_j + \lambda_{k+1} r_k(\boldsymbol{\Sigma}) + n \lambda_{k+1}}{n \lambda_j} \right)$$

with probability at least  $1 - 5e^{-\frac{n}{c_4}}$ . By Lemma 15 of Bartlett et al. (2020),

$$\begin{aligned} \sum_{j=1}^d \frac{\lambda_j^2 \mathbf{z}_j^\top \mathbf{A}_{-j}^{-1} \mathbf{z}_j}{1 + \lambda_j \mathbf{z}_j^\top \mathbf{A}_{-j}^{-1} \mathbf{z}_j} &\geq \sum_{j=1}^d \frac{\lambda_j^2 \mathbf{z}_j^\top \mathbf{A}_{-j}^{-1} \mathbf{z}_j}{c_4 \lambda_j \mathbf{z}_j^\top \mathbf{A}_{-j}^{-1} \mathbf{z}_j} \left( \frac{n \lambda_j}{n \lambda_j + \lambda_{k+1} r_k(\boldsymbol{\Sigma}) + n \lambda_{k+1}} \right) \\ &= \frac{1}{c_4} \sum_{j=1}^d \frac{n \lambda_j^2}{n \lambda_j + \lambda_{k+1} r_k(\boldsymbol{\Sigma}) + n \lambda_{k+1}} \end{aligned}$$

with probability at least  $1 - 10e^{-\frac{n}{c_4}}$ . By the median inequality, there exist constants  $b, c_5 \geq 1$  such that

$$\frac{1}{c_4} \sum_{j=1}^d \frac{n \lambda_j^2}{\lambda_{k+1} r_k(\boldsymbol{\Sigma}) + n \lambda_{k+1} + n \lambda_j} \geq \frac{1}{b c_5} \sum_{j=1}^d \min \left( \lambda_j, \frac{b n \lambda_j^2}{\lambda_{k+1} r_k(\boldsymbol{\Sigma})}, \frac{\lambda_j^2}{\lambda_{k+1}} \right).$$

If  $r_k(\boldsymbol{\Sigma}) < b n$ , the second term in the minimum is larger than the third term. In this case,

$$\begin{aligned} \frac{1}{b c_5} \sum_{j=1}^d \min \left( \lambda_j, \frac{b n \lambda_j^2}{\lambda_{k+1} r_k(\boldsymbol{\Sigma})}, \frac{\lambda_j^2}{\lambda_{k+1}} \right) &\geq \frac{1}{b c_5} \sum_{j=1}^d \min \left( \lambda_j, \frac{\lambda_j^2}{\lambda_{k+1}} \right) \\ &= \frac{1}{b c_5} \left( \sum_{j=1}^k \lambda_j + \sum_{j=k+1}^d \frac{\lambda_j^2}{\lambda_{k+1}} \right). \end{aligned} \tag{29}$$

On the other hand, if  $r_k(\boldsymbol{\Sigma}) \geq b n$ , the second term in the minimum is smaller than the third term. In this case,

$$\begin{aligned} \frac{1}{b c_5} \sum_{j=1}^d \min \left( \lambda_j, \frac{b n \lambda_j^2}{\lambda_{k+1} r_k(\boldsymbol{\Sigma})}, \frac{\lambda_j^2}{\lambda_{k+1}} \right) &\geq \frac{1}{b c_5} \sum_{j=1}^d \min \left( \lambda_j, \frac{b n \lambda_j^2}{\lambda_{k+1} r_k(\boldsymbol{\Sigma})} \right) \\ &= \frac{1}{b c_5} \min_{1 \leq \ell \leq k} \left( \sum_{j=1}^{\ell} \lambda_j + \frac{b n \sum_{j=\ell+1}^d \lambda_j^2}{\lambda_{k+1} r_k(\boldsymbol{\Sigma})} \right). \end{aligned}$$

Recall that  $k^* := \min\{k \geq 0 : r_k(\mathbf{\Sigma}) \geq bn\}$ . By Lemma 17 of [Bartlett et al. \(2020\)](#), there exists a constant  $c_6 \geq 1$  such that

$$\frac{1}{bc_5} \min_{1 \leq \ell \leq k} \left( \sum_{j=1}^{\ell} \lambda_j + \frac{bn \sum_{j=\ell+1}^d \lambda_j^2}{\lambda_{k+1} r_k(\mathbf{\Sigma})} \right) = \frac{1}{c_4} \left( \sum_{j=1}^{k^*} \lambda_j + \frac{n \sum_{j=k^*+1}^d \lambda_j^2}{\lambda_{k+1} r_k(\mathbf{\Sigma})} \right).$$

Therefore, if  $k^* < \frac{n}{c_1}$ ,

$$\text{Tr}(\mathbf{X}^\top \mathbf{C} \mathbf{X}) \geq \frac{1}{c_6} \left( \sum_{j=1}^{k^*} \lambda_j + \frac{n}{R_{k^*}(\mathbf{\Sigma})} \sum_{j=k^*+1}^d \lambda_j \right).$$

On the other hand, if  $k^* > \frac{n}{c_1}$ , then  $r_k(\mathbf{\Sigma}) < bn$  for all  $k \leq \frac{n}{c_1}$ . Equation (29) then implies

$$\text{Tr}(\mathbf{X}^\top \mathbf{C} \mathbf{X}) \geq \frac{1}{bc_5} \left( \sum_{j=1}^{\frac{n}{c_1}} \lambda_j + \sum_{j=\frac{n}{c_1}+1}^d \frac{\lambda_j^2}{\lambda_{\frac{n}{c_1}+1}} \right) \geq \frac{1}{bc_5}.$$

Choosing  $c = \max(c_2 c_3, c_4, bc_5)$  and taking a union bound over the upper and lower bounds completes the proof.  $\square$

## D.5 Tradeoff between regression bias and task shift error

In this section, we provide the proof of Theorem 13.

*Proof.* (Theorem 13)

Note that for the given signal model we have  $\mathbb{E}_{\theta^*} \theta_j^{*2} = \bar{\theta}_j^2$  and  $\bar{\theta}_j^2 \geq 1$  for all  $1 \leq j \leq d$ . Hence, the assumption of the Theorem 12 lower bound is satisfied with  $\sigma^2 = 1$ .

We first prove the case where  $\kappa = 0$ . In this case, there is some finite  $n$  after which  $r_0(\mathbf{\Sigma}) \geq bn$  for a constant  $b > 1$ . By the lower bound of Lemma 11, there exists a constant  $c_1 \geq 1$  such that

$$\mathbb{E}_{\theta^*} L(\tilde{\theta}) \geq \frac{1}{c_1} \sum_{j=1}^d \frac{\lambda_j \bar{\theta}_j^2}{\left(1 + \frac{n \lambda_j}{\sum_{j=1}^d \lambda_j}\right)^2}$$

with probability at least  $1 - c_1 e^{-\frac{n}{c_1}}$ . Since  $r_0(\mathbf{\Sigma}) \geq bn$ , for any  $1 \leq j \leq d$  we have

$$\frac{n \lambda_j}{\sum_{j=1}^d \lambda_j} \leq \frac{n \lambda_1}{\sum_{j=1}^d \lambda_j} \leq \frac{1}{b}.$$

Therefore,

$$\mathbb{E}_{\theta^*} L(\tilde{\theta}) \geq \frac{1}{c_1} \sum_{j=1}^d \frac{\lambda_j \bar{\theta}_j^2}{\left(1 + \frac{1}{b}\right)^2} = \frac{\|\mathbf{\Sigma}^{\frac{1}{2}} \bar{\theta}\|_2^2}{c_1 \left(1 + \frac{1}{b}\right)^2}.$$

In the limit as  $n \rightarrow \infty$ , the term  $c_1 e^{-\frac{n}{c_1}}$  is zero. Hence,

$$\lim_{n, d \rightarrow \infty} \mathbb{E}_{\theta^*} L(\tilde{\theta}) \geq \frac{\|\mathbf{\Sigma}^{\frac{1}{2}} \bar{\theta}\|_2^2}{c_1 \left(1 + \frac{1}{b}\right)^2}$$

almost surely.

We now prove the case where  $0 < \kappa < n$ . By assumption,  $\|\mathbf{\Sigma}^{\frac{1}{2}} \bar{\theta}\|_2^2$  is constant for all  $n$  and  $\bar{\theta}_j^2 \geq 1$  for all  $1 \leq j \leq d$ . Hence, there exists a constant  $c_2 \geq 1$  such that either (i)  $\lim_{n, d \rightarrow \infty} \sum_{j=1}^{\kappa} \lambda_j \geq \frac{1}{c_2}$  or (ii)  $\lim_{n, d \rightarrow \infty} \sum_{j=\kappa+1}^d \lambda_j \geq \frac{1}{c_2}$ . (In other words, we cannot have both terms go to zero).

In case (i), by the lower bound of Theorem 12, there exist constants  $c_3, c_4 \geq 1$  such that if  $k^* < \frac{n}{c_3}$ , then

$$\mathbb{E}_{\theta^*, \mathbf{x}} \left( \mathbf{x}^\top \hat{\boldsymbol{\theta}} - \mathbf{x}^\top \tilde{\boldsymbol{\theta}} \right)^2 \geq \frac{\alpha^2}{c_4} \left( \sum_{j=1}^{k^*} \lambda_j + \frac{n}{R_{k^*}(\boldsymbol{\Sigma})} \sum_{j=k^*+1}^d \lambda_j \right) \geq \frac{\alpha^2}{c_4} \sum_{j=1}^{k^*} \lambda_j$$

with probability at least  $1 - 14e^{-\frac{n}{c_4}}$ . In the limit as  $n \rightarrow \infty$ , the term  $14e^{-\frac{n}{c_4}}$  is zero, and any  $\kappa < n$  satisfies the condition  $\kappa < \frac{n}{c_3}$ . Using the assumption of case (i),

$$\lim_{n, d \rightarrow \infty} \mathbb{E}_{\theta^*, \mathbf{x}} \left( \mathbf{x}^\top \hat{\boldsymbol{\theta}} - \mathbf{x}^\top \tilde{\boldsymbol{\theta}} \right)^2 \geq \lim_{n, d \rightarrow \infty} \frac{\alpha^2}{c_4} \sum_{j=1}^{\kappa} \lambda_j \geq \frac{\alpha^2}{c_2 c_4}$$

almost surely.

In case (ii), since case (i) is not satisfied, we have  $\lambda_1 \bar{\theta}_1^2, \dots, \lambda_\kappa \bar{\theta}_\kappa^2 \rightarrow 0$ . But by assumption,  $\bar{\theta}_j^2 \geq 1$  for all  $1 \leq j \leq \kappa$ , so then  $\lambda_1, \dots, \lambda_\kappa \rightarrow 0$ . This implies  $\boldsymbol{\Sigma} \rightarrow \mathbf{0}$ , a contradiction with positive-definiteness.

Finally, we prove the case where  $\kappa > n$ . By the lower bound of Theorem 12, we have

$$\mathbb{E}_{\theta^*, \mathbf{x}} \left( \mathbf{x}^\top \hat{\boldsymbol{\theta}} - \mathbf{x}^\top \tilde{\boldsymbol{\theta}} \right)^2 \geq \frac{\alpha^2}{c_4}$$

with probability at least  $1 - 10e^{-\frac{n}{c_4}}$ . In the limit as  $n \rightarrow \infty$ , the term  $10e^{-\frac{n}{c_4}}$  is zero. Hence,

$$\lim_{n, d \rightarrow \infty} \mathbb{E}_{\theta^*, \mathbf{x}} \left( \mathbf{x}^\top \hat{\boldsymbol{\theta}} - \mathbf{x}^\top \tilde{\boldsymbol{\theta}} \right)^2 \geq \frac{\alpha^2}{c_4}$$

almost surely. Choosing  $c = \max \left( c_1 \left( 1 + \frac{1}{b} \right)^2, c_2 c_4 \right)$  completes the proof.  $\square$

## E Zero-shot task shift in the case of dense signal

We conclude the random signal section with a model which does not require the ansatz  $\mathbf{D} \approx \alpha \mathbf{I}$  introduced in Section 3.2. In this section, we first introduce the settings and characterize the task shift of the dense signal model in Section E.1. In Section E.2, we prove that dense signal implies poor bias, and we reduce the task shift error to a benign overfitting term via high-dimensional probability arguments. In Section E.3, we show concentration of task shift error terms via an adaptation of standard sub-Gaussian random matrix analysis.

### E.1 Dense random signal model without simplifying ansatz

We study a “dense” signal, *i.e.*, one which has similar magnitude in all dimensions. Specifically, we let  $\theta_j^* \sim \mathcal{N} \left( 0, \frac{1}{d\lambda_j} \right)$  for all  $j \in [d]$  so that  $\|\boldsymbol{\Sigma}^{\frac{1}{2}} \boldsymbol{\theta}^*\|_2^2 = 1$ . Writing  $\mathbf{X} \stackrel{d}{=} \mathbf{Z} \boldsymbol{\Sigma}^{1/2}$  where  $\mathbf{Z} \in \mathbb{R}^{n \times d}$  has independent standard Gaussian entries, we can see that

$$\tilde{\mathbf{y}} = \mathbf{X} \boldsymbol{\theta}^* \stackrel{d}{=} \left( \mathbf{Z} \boldsymbol{\Sigma}^{1/2} \right) \left( \frac{1}{\sqrt{d}} \boldsymbol{\Sigma}^{-1/2} \mathbf{z} \right) = \frac{1}{\sqrt{d}} \mathbf{Z} \mathbf{z},$$

where  $\mathbf{z} \in \mathbb{R}^d$  is a standard Gaussian vector with independent entries. Therefore, this setting of  $\boldsymbol{\theta}^*$  is equivalent in distribution to scaled Gaussian random signal under isotropic covariance, clearly a “dense” problem instance.

In this regime, we show that while the limiting bias is nonzero, the limiting task shift error is zero as long as the covariance matrix has large effective rank compared to  $n$ . Note that this condition is necessary, but not sufficient, for  $\boldsymbol{\Sigma}$  to exhibit benign overfitting.

**Theorem 23.** *For any  $\boldsymbol{\Sigma}$  there exists a constant  $c \geq 1$  such that the following holds. Suppose  $\boldsymbol{\theta}^*$  is such that  $\theta_j^* \sim \mathcal{N} \left( 0, \frac{1}{d\lambda_j} \right)$  for all  $j \in [d]$ . Then we have  $\lim_{n, d \rightarrow \infty} \mathbb{E}_{\theta^*} L(\tilde{\boldsymbol{\theta}}) \geq \frac{1}{c}$  but  $\lim_{n, d \rightarrow \infty} \mathbb{E}_{\theta^*, \mathbf{x}} \left( \mathbf{x}^\top \hat{\boldsymbol{\theta}} - \mathbf{x}^\top \tilde{\boldsymbol{\theta}} \right)^2 \leq$*

$c \lim_{n,d \rightarrow \infty} \left( \frac{k^*}{n} + \frac{n}{R_{k^*}(\Sigma)} \right)$  almost surely. In particular, if  $\lim_{n,d \rightarrow \infty} \frac{k^*}{n} = \lim_{n,d \rightarrow \infty} \frac{n}{R_{k^*}(\Sigma)} = 0$ , then  $\lim_{n,d \rightarrow \infty} \mathbb{E}_{\theta^*, \mathbf{x}} \left( \mathbf{x}^\top \hat{\boldsymbol{\theta}} - \mathbf{x}^\top \tilde{\boldsymbol{\theta}} \right)^2 = 0$  almost surely.

The proof of Theorem 23 is in the following Appendix E.2.

## E.2 Analysis of bias and task shift error via benign overfitting

In this section, we provide the proof of Theorem 23, which lower bounds the bias and upper bounds the task shift error of a “dense” random signal.

*Proof.* (Theorem 23)

We begin by characterizing the bias term  $\mathbb{E}_{\theta^*} L(\tilde{\boldsymbol{\theta}})$ . Since the Gaussian distribution is symmetric, we may write  $\theta_j^* = r_j \bar{\theta}_j$  where each  $r_j$  is an independent Rademacher random variable and  $\bar{\theta}_j$  is drawn according to a Gaussian distribution. Applying the lower bound of Lemma 11, for a constant  $c_1 \geq 1$  we have

$$\mathbb{E}_{\theta^*} L(\tilde{\boldsymbol{\theta}}) \geq \frac{1}{c_1} \sum_{j=1}^d \frac{\lambda_j \mathbb{E} \bar{\theta}_j^2}{\left( 1 + \frac{n \lambda_j}{\sum_{k=1}^d \lambda_k} \right)^2}$$

with probability at least  $1 - c_1 e^{-\frac{n}{c_1}}$ <sup>6</sup>. As previously mentioned, setting  $\theta_j^* \sim \mathcal{N}\left(0, \frac{1}{d \lambda_j}\right)$  for all  $1 \leq j \leq d$  under general covariance is equivalent in distribution to setting  $\theta_j^* \sim \mathcal{N}\left(0, \frac{1}{d}\right)$  for all  $1 \leq j \leq d$  under isotropic covariance. In this case,  $\mathbb{E} \bar{\theta}_j^2 = \frac{1}{d}$  and  $\lambda_1 = \dots = \lambda_d = 1$ . Therefore, we have

$$\begin{aligned} \mathbb{E}_{\theta^*} L(\tilde{\boldsymbol{\theta}}) &\geq \frac{1}{dc_1} \sum_{j=1}^d \frac{1}{\left( 1 + \frac{n \lambda_j}{\sum_{k=1}^d \lambda_k} \right)^2} \\ &= \frac{1}{dc_1} \sum_{j=1}^d \frac{1}{\left( 1 + \frac{n}{d} \right)^2} \\ &= \frac{1}{c_1} \frac{1}{\left( 1 + \frac{n}{d} \right)^2}. \end{aligned}$$

In the limit as  $n \rightarrow \infty$ , the term  $c_1 e^{-\frac{n}{c_1}}$  is zero. Hence,

$$\lim_{n,d \rightarrow \infty} \mathbb{E}_{\theta^*} L(\tilde{\boldsymbol{\theta}}) \geq \frac{1}{c_1}$$

almost surely.

For the task shift error  $\mathbb{E}_{\theta^*, \mathbf{x}} \left( \mathbf{x}^\top \hat{\boldsymbol{\theta}} - \mathbf{x}^\top \tilde{\boldsymbol{\theta}} \right)^2$ , we begin with a lemma.

**Lemma 24.** Let  $\mathbf{u}_1, \dots, \mathbf{u}_n, \boldsymbol{\varepsilon} \in \mathbb{R}^d$  be independent standard Gaussian random vectors. We have for any  $1 \leq i, k \leq n$ ,

$$\mathbb{E}_{\boldsymbol{\varepsilon}} \left[ \text{sign}(\mathbf{u}_i^\top \boldsymbol{\varepsilon}) \mathbf{u}_k^\top \boldsymbol{\varepsilon} \right] = \sqrt{\frac{2}{\pi}} \frac{\mathbf{u}_i^\top \mathbf{u}_k}{\|\mathbf{u}_i\|_2}.$$

The proof is in Appendix E.3. Now, have by Equation (19) that

$$\begin{aligned} \mathbb{E}_{\theta^*, \mathbf{x}} \left( \mathbf{x}^\top \hat{\boldsymbol{\theta}} - \mathbf{x}^\top \tilde{\boldsymbol{\theta}} \right)^2 &= \mathbb{E}_{\theta^*} \left[ (\hat{\mathbf{y}} - \tilde{\mathbf{y}})^\top \mathbf{C} (\hat{\mathbf{y}} - \tilde{\mathbf{y}}) \right] \\ &= \mathbb{E}_{\theta^*} \left[ \tilde{\mathbf{y}}^\top \mathbf{C} \tilde{\mathbf{y}} \right] - \mathbb{E}_{\theta^*} \left[ \tilde{\mathbf{y}}^\top \mathbf{C} \hat{\mathbf{y}} \right] - \mathbb{E}_{\theta^*} \left[ \hat{\mathbf{y}}^\top \mathbf{C} \tilde{\mathbf{y}} \right] + \mathbb{E}_{\theta^*} \left[ \hat{\mathbf{y}}^\top \mathbf{C} \hat{\mathbf{y}} \right]. \end{aligned}$$

<sup>6</sup> The statement of Lemma 11 is for deterministic  $\tilde{\boldsymbol{\theta}}$ , but the same result holds for  $\tilde{\boldsymbol{\theta}}$  with random coordinates independent of each other and  $\mathbf{r}$  by taking the expectation over  $\tilde{\boldsymbol{\theta}}$ .



By the cyclic property of trace,

$$\begin{aligned}\mathbb{E}_{\boldsymbol{\theta}^*} [\tilde{\mathbf{y}}^\top \mathbf{C} \tilde{\mathbf{y}}] &= \text{Tr} (\mathbf{C} \mathbb{E}_{\boldsymbol{\theta}^*} [\tilde{\mathbf{y}} \tilde{\mathbf{y}}^\top]) \\ &= \text{Tr} (\mathbf{C} \mathbb{E}_{\boldsymbol{\theta}^*} [\mathbf{X} \boldsymbol{\theta}^* \boldsymbol{\theta}^{*\top} \mathbf{X}^\top]) \\ &= \text{Tr} \left( \mathbf{C} \left( \frac{1}{d} \mathbf{u} \mathbf{u}^\top \right) \right),\end{aligned}$$

where  $\mathbf{u} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ . Next,

$$\mathbb{E}_{\boldsymbol{\theta}^*} [\hat{\mathbf{y}}^\top \mathbf{C} \hat{\mathbf{y}}] = \text{Tr} (\mathbf{C} \mathbb{E}_{\boldsymbol{\theta}^*} [\hat{\mathbf{y}} \hat{\mathbf{y}}^\top]) =: \text{Tr} (\mathbf{C} \mathbf{F}),$$

where we define

$$\begin{aligned}F_{ik} &:= \mathbb{E}_{\boldsymbol{\theta}^*} [\text{sign}(\mathbf{x}_i^\top \boldsymbol{\theta}^*) \mathbf{x}_k^\top \boldsymbol{\theta}^*] \\ &= \begin{cases} \frac{1}{\sqrt{d}} \mathbb{E}_{\boldsymbol{\epsilon}} [|\mathbf{u}_i^\top \boldsymbol{\epsilon}|] & i = k \\ \frac{1}{\sqrt{d}} \mathbb{E}_{\boldsymbol{\epsilon}} [\text{sign}(\mathbf{u}_i^\top \boldsymbol{\epsilon}) \mathbf{u}_k^\top \boldsymbol{\epsilon}] & i \neq k \end{cases} \\ &= \begin{cases} \sqrt{\frac{2}{\pi d}} \|\mathbf{u}_i\|_2 & i = k \\ \sqrt{\frac{2}{\pi d}} \frac{\mathbf{u}_i^\top \mathbf{u}_k}{\|\mathbf{u}_i\|_2} & i \neq k \end{cases}\end{aligned}$$

by Lemma 24. Similarly,

$$\mathbb{E}_{\boldsymbol{\theta}^*} [\hat{\mathbf{y}}^\top \mathbf{C} \hat{\mathbf{y}}] = \text{Tr} (\mathbf{C} \mathbb{E}_{\boldsymbol{\theta}^*} [\hat{\mathbf{y}} \hat{\mathbf{y}}^\top]) =: \text{Tr} (\mathbf{C} \mathbf{G}),$$

where we define

$$\begin{aligned}G_{ik} &:= \mathbb{E}_{\boldsymbol{\theta}^*} [\text{sign}(\mathbf{x}_i^\top \boldsymbol{\theta}^*) \text{sign}(\mathbf{x}_k^\top \boldsymbol{\theta}^*)] \\ &= \begin{cases} 1 & i = k \\ \mathbb{E}_{\boldsymbol{\epsilon}} [\text{sign}(\mathbf{u}_i^\top \boldsymbol{\epsilon}) \text{sign}(\mathbf{u}_k^\top \boldsymbol{\epsilon})] & i \neq k \end{cases} \\ &= \begin{cases} 1 & i = k \\ \frac{2}{\pi} \sin^{-1} \left( \frac{\mathbf{u}_i^\top \mathbf{u}_k}{\|\mathbf{u}_i\|_2 \|\mathbf{u}_k\|_2} \right) & i \neq k \end{cases}\end{aligned}$$

by Grothendieck's identity (Lemma 3.6.6 in Vershynin (2018)). Putting everything together, we have

$$\mathbb{E}_{\boldsymbol{\theta}^*, \mathbf{x}} (\mathbf{x}^\top \hat{\boldsymbol{\theta}} - \mathbf{x}^\top \tilde{\boldsymbol{\theta}})^2 = \text{Tr} (\mathbf{C} \mathbf{S})$$

where

$$S_{ik} = \begin{cases} \frac{1}{d} \|\mathbf{u}_i\|_2^2 - 2\sqrt{\frac{2}{\pi d}} \|\mathbf{u}_i\|_2 + 1 & i = k \\ \frac{1}{d} \mathbf{u}_i^\top \mathbf{u}_k - \sqrt{\frac{2}{\pi d}} \frac{\mathbf{u}_i^\top \mathbf{u}_k}{\|\mathbf{u}_k\|_2} - \sqrt{\frac{2}{\pi d}} \frac{\mathbf{u}_i^\top \mathbf{u}_k}{\|\mathbf{u}_k\|_2} + \frac{2}{\pi} \sin^{-1} \left( \frac{\mathbf{u}_i^\top \mathbf{u}_k}{\|\mathbf{u}_i\|_2 \|\mathbf{u}_k\|_2} \right) & i \neq k. \end{cases}$$

By concentration of a standard Gaussian random vector and a union bound, for all  $1 \leq i \leq n$  we have

$$\begin{aligned}d - \sqrt{d} &\leq \|\mathbf{u}_i\|_2^2 \leq d + \sqrt{d}, \\ \sqrt{d - \sqrt{d}} &\leq \|\mathbf{u}_i\|_2 \leq \sqrt{d + \sqrt{d}}\end{aligned}$$

with probability at least  $1 - 2ne^{-d}$ . By Bernstein's inequality and a union bound, the above holds simultaneously with

$$|\mathbf{u}_i^\top \mathbf{u}_k| \leq \sqrt{d + \sqrt{d}}$$

for all  $1 \leq i \neq k \leq n$ , with probability at least  $1 - 2n^3 e^{-d}$ . Therefore, since  $\frac{2}{\pi} |\sin^{-1}(x)| \leq |x|$  for  $-1 \leq x \leq 1$ , there exists a constant  $c_2 > 1$  such that

$$\left| \sin^{-1} \left( \frac{\mathbf{u}_i^\top \mathbf{u}_k}{\|\mathbf{u}_i\|_2 \|\mathbf{u}_k\|_2} \right) \right| \leq \left| \frac{\mathbf{u}_i^\top \mathbf{u}_k}{\|\mathbf{u}_i\|_2 \|\mathbf{u}_k\|_2} \right| \leq \frac{c_2}{\sqrt{d}}$$

over the same randomness as above. Thus, there exists a constant  $c_3 > 1$  such that for all  $1 \leq i, k \leq n$ ,

$$\frac{1}{c_3} \leq S_{ii} \leq c_3 \quad (30)$$

$$|S_{ik}| \leq \frac{c_3}{\sqrt{d}} \quad i \neq k \quad (31)$$

with probability at least  $1 - c_3 n^3 e^{-d}$ . These high-probability bounds will be used to prove the following lemma, detailed in Appendix E.3.

**Lemma 25.** *There exists a constant  $c > 0$  such that  $\mu_1(\mathbf{S}) \leq c$  with probability at least  $1 - cn^3 e^{-d}$ .*

By Lemma 25, since  $\mathbf{C}$  is positive semi-definite, there exists a constant  $c_4 > 0$  such that

$$\text{Tr}(\mathbf{CS}) \leq \text{Tr}(\mathbf{C})\mu_1(\mathbf{S}) \leq c_4 \text{Tr}(\mathbf{C})$$

with probability at least  $1 - c_4 n^3 e^{-d}$ . By Lemma 20, there exists a constant  $c_5 \geq 1$  such that

$$\text{Tr}(\mathbf{C}) \leq c_5 \left( \frac{k^*}{n} + \frac{n}{R_{k^*}(\mathbf{\Sigma})} \right)$$

with probability at least  $1 - 17e^{-\frac{n}{c_5}}$ . In the limit as  $n, d \rightarrow \infty$ , the terms  $c_4 n^3 e^{-d}$  and  $17e^{-\frac{n}{c_5}}$  are zero. Therefore, by a union bound,

$$\lim_{n, d \rightarrow \infty} \mathbb{E}_{\boldsymbol{\theta}^*, \mathbf{x}} \left( \mathbf{x}^\top \hat{\boldsymbol{\theta}} - \mathbf{x}^\top \tilde{\boldsymbol{\theta}} \right)^2 \leq c_4 c_5 \lim_{n, d \rightarrow \infty} \left( \frac{k^*}{n} + \frac{n}{R_{k^*}(\mathbf{\Sigma})} \right)$$

almost surely. Choosing  $c = \max(c_1, c_4 c_5)$  completes the proof.  $\square$

### E.3 Concentration of task shift error terms

In this section, we provide the proofs of the technical Lemmas 24 and 25 used in the proof of Theorem 23.

*Proof.* (Lemma 24)

Note that  $\mathbf{u}_i^\top \boldsymbol{\varepsilon} \sim \mathcal{N}(0, \|\mathbf{u}_i\|_2^2)$  and  $\mathbf{u}_k^\top \boldsymbol{\varepsilon} \sim \mathcal{N}(0, \|\mathbf{u}_k\|_2^2)$ . Their correlation coefficient is

$$\rho_{ik} = \frac{\mathbf{u}_i^\top \mathbf{u}_k}{\|\mathbf{u}_i\|_2 \|\mathbf{u}_k\|_2}. \quad (32)$$

Let  $Z_1 \sim \mathcal{N}(0, 1)$  and  $Z_2 \sim \mathcal{N}(0, 1)$  be Gaussian variables independent of each other and  $\mathbf{u}_i, \mathbf{u}_k$ . We may write

$$\begin{aligned} \mathbf{u}_i^\top \boldsymbol{\varepsilon} &\stackrel{d}{=} \|\mathbf{u}_i\|_2 Z_1 \\ \mathbf{u}_k^\top \boldsymbol{\varepsilon} &\stackrel{d}{=} \rho_{ik} \|\mathbf{u}_k\|_2 Z_1 + \sqrt{1 - \rho_{ik}^2} \|\mathbf{u}_k\|_2 Z_2. \end{aligned}$$

Since  $Z_1$  and  $Z_2$  are independent and centered,

$$\begin{aligned} \mathbb{E}_{\boldsymbol{\varepsilon}} [\text{sign}(\mathbf{u}_i^\top \boldsymbol{\varepsilon}) \mathbf{u}_k^\top \boldsymbol{\varepsilon}] &= \mathbb{E}_{Z_1} [\text{sign}(Z_1) \rho_{ik} \|\mathbf{u}_k\|_2 Z_1] + \mathbb{E}_{Z_1, Z_2} [\text{sign}(Z_1) \sqrt{1 - \rho_{ik}^2} \|\mathbf{u}_k\|_2 Z_2] \\ &= \mathbb{E}_{Z_1} [\text{sign}(Z_1) \rho_{ik} \|\mathbf{u}_k\|_2 Z_1] \\ &= \rho_{ik} \|\mathbf{u}_k\|_2 \mathbb{E}_{Z_1} [\text{sign}(Z_1) Z_1]. \end{aligned}$$

Since  $Z_1$  is standard Gaussian, we have

$$\mathbb{E}_{Z_1} [\text{sign}(Z_1) Z_1] = \mathbb{E}_{Z_1} |Z_1| = \sqrt{\frac{2}{\pi}}.$$

Using the value of  $\rho_{ik}$  derived in Equation (32),

$$\mathbb{E}_{\boldsymbol{\varepsilon}} [\text{sign}(\mathbf{u}_i^\top \boldsymbol{\varepsilon}) \mathbf{u}_k^\top \boldsymbol{\varepsilon}] = \sqrt{\frac{2}{\pi}} \frac{\mathbf{u}_i^\top \mathbf{u}_k}{\|\mathbf{u}_i\|_2}.$$

□

*Proof.* (Lemma 25)

We will adapt the proof of Theorem 5.39 of Vershynin (2012). By the approximate isometry lemma (Lemma 5.36 in Vershynin (2012)), if for some  $c > 0$  we have

$$\left\| \frac{1}{n} \mathbf{S}^\top \mathbf{S} - \mathbf{I} \right\| \leq \max(c, c^2)$$

then  $\mu_1(\mathbf{S}) \leq c$  as desired. Let  $\mathcal{N}$  be a  $\frac{1}{4}$ -net of the unit sphere  $S^{n-1}$  with respect to the Euclidean metric, and by Lemma 5.2 of Vershynin (2012) put  $|\mathcal{N}| \leq 9^n$ . Then, by Lemma 5.4 of Vershynin (2012),

$$\left\| \frac{1}{n} \mathbf{S}^\top \mathbf{S} - \mathbf{I} \right\| \leq 2 \max_{\mathbf{x} \in \mathcal{N}} \left| \left\langle \left( \frac{1}{n} \mathbf{S}^\top \mathbf{S} - \mathbf{I} \right) \mathbf{x}, \mathbf{x} \right\rangle \right| = 2 \max_{\mathbf{x} \in \mathcal{N}} \left| \frac{1}{n} \|\mathbf{S}\mathbf{x}\|_2^2 - 1 \right|.$$

Now consider a fixed vector  $\mathbf{x} \in S^{n-1}$ . We have

$$\begin{aligned} \|\mathbf{S}\mathbf{x}\|_2^2 &= \sum_{i=1}^n \langle \mathbf{S}_i, \mathbf{x} \rangle^2 \\ &= \sum_{i=1}^n \left( \sum_{k=1}^n S_{ik} x_k \right)^2 \\ &\leq \sum_{i=1}^n \left( \sum_{k=1}^n |S_{ik}| |x_k| \right)^2. \end{aligned}$$

By the concentration results in Equations 30 and 31, there exists a constant  $c_1 > 0$  such that

$$\begin{aligned} |S_{ik}| &\leq c_1 \\ |S_{ik}| &\leq \frac{c_1}{\sqrt{d}} \quad i \neq k \end{aligned}$$

with probability at least  $1 - c_1 n^3 e^{-d}$ . In particular,

$$\|\mathbf{S}\mathbf{x}\|_2^2 \leq \sum_{i=1}^n \left( c_1 |x_i| + \sum_{i \neq k} \frac{c_1}{\sqrt{d}} |x_k| \right)^2$$

with probability at least  $1 - c_1 n^3 e^{-d}$ . Using  $(a + b)^2 \leq 2(a^2 + b^2)$  for  $a, b \in \mathbb{R}$  and the fact that  $\mathbf{x} \in S^{n-1}$ ,

$$\begin{aligned} \|\mathbf{S}\mathbf{x}\|_2^2 &\leq 2c_1^2 \left( \sum_{i=1}^n x_i^2 + \frac{1}{d} \left( \sum_{i=1}^n |x_i| \right)^2 \right) \\ &\leq 2c_1^2 \left( 1 + \frac{n}{d} \right) \end{aligned}$$

with probability at least  $1 - c_1 n^3 e^{-d}$ . Therefore

$$\frac{1}{n} \|\mathbf{S}\mathbf{x}\|_2^2 \leq 2c_1^2 \left( \frac{1}{n} + \frac{1}{d} \right),$$

so for  $c_3 > \max(1, 4c_1^2 - 1)$ , we have

$$\left| \frac{1}{n} \|\mathbf{S}\mathbf{x}\|_2^2 - 1 \right| \leq c_3$$

with probability at least  $1 - c_1 n^3 e^{-d}$ .

Taking a union bound over  $\mathcal{N}$ , there exists a constant  $c_4 \geq 1$  such that

$$\mathbb{P} \left( \max_{\mathbf{x} \in \mathcal{N}} \left| \frac{1}{n} \|\mathbf{S}\mathbf{x}\|_2^2 - 1 \right| \geq c \right) \leq 9^n \cdot c_1 n^3 e^{-d} \leq c_4 n^3 e^{n-d}.$$

Therefore, for  $c = \max(c_3^2, c_4^2)$ , we have  $\mu_1(\mathbf{S}) \leq c$  with probability at least  $1 - cn^3 e^{n-d}$ .

□

## F Support recovery: analysis beyond survival and contamination

In this section, we provide all of the proofs of our support recovery results from Section 4 under the diagonal covariance assumption. In Section F.1, we introduce our key lemma of this section, then use it to prove Theorem 14, which characterizes the magnitudes of individual parameters arising from minimum-norm interpolation. In Section F.2, we provide proofs for the support recovery of spiked and polynomial decay  $\Sigma$  when  $\theta^*$  is supported entirely in the top  $k^*$  indices of the spectrum of  $\Sigma$ . Finally, in Section F.3, we show that support recovery works even when  $\theta^*$  is supported outside these top  $k^*$  indices — thereby handling cases wherein  $k^* = 0$ , such as isotropic covariance — under some additional conditions.

We remark that repeated application of the Sherman-Morrison-Woodbury identity induces a linear dependence on the sparsity parameter  $t$  in Lemma 26. This necessitates  $t \ll n^{\frac{1}{2}}$  when the support lies entirely within the top  $k^*$  indices of the covariance spectrum (Section F.2) and  $t \ll n^{\frac{1}{4}}$  otherwise (Section F.3). It is conceivable that these bounds could be improved with a finer analysis. A relevant work is Wu and Sahai (2023), who study a multiclass classification setting where  $\hat{\mathbf{y}}$  is a one-hot encoded vector and develop an improved Hanson-Wright inequality utilizing the sparsity in  $\hat{\mathbf{y}}$  in the multiclass settings. Unfortunately, we cannot directly apply their bound: even though our  $\theta^*$  is sparse, our  $\hat{\mathbf{y}}_i$  have Rademacher distribution and are generally not sparse.

### F.1 Characterization of classification MNI parameters

Before we prove Theorem 14, we prove the following lemma which lower bounds support indices of  $\hat{\theta}$  and upper bounds non-support indices of  $\hat{\theta}$ .

**Lemma 26.** Define  $\bar{\mathcal{S}} := \mathcal{S} \cup [k^*]$  and denote by  $\{\tilde{\lambda}_j\}_{j=1}^{d-|\bar{\mathcal{S}}|}$  the diagonal entries of the matrix  $\Sigma_{-\bar{\mathcal{S}}}$ . Under Assumptions 2 and 5, for large enough  $n$ , we have

$$\begin{aligned} |\hat{\theta}_j| &\geq \sqrt{\frac{2\beta_j}{\lambda_j \pi}} \frac{\lambda_j \left( \frac{n}{c_1 \tilde{\lambda}_1 r_0 (\Sigma_{-\bar{\mathcal{S}}})} \right)}{1 + \lambda_j \left( \frac{c_2 n}{\tilde{\lambda}_1 r_0 (\Sigma_{-\bar{\mathcal{S}}})} \right)} && \text{for } j \in \mathcal{S}, \\ |\hat{\theta}_j| &\leq t \sqrt{\frac{c_3}{\lambda_j n^{1-2\epsilon}}} && \text{for } j \in [k^*] \cap \mathcal{S}^c, \\ |\hat{\theta}_j| &\leq t \sqrt{\frac{c_4 n^{1+2\epsilon} \lambda_j}{\left( \tilde{\lambda}_1 r_0 (\Sigma_{-\bar{\mathcal{S}}}) \right)^2}} && \text{for } k^* < j \leq d, j \in \mathcal{S}^c, \end{aligned}$$

where  $\beta_j := \frac{\lambda_j \theta_j^{*2}}{\sum_{k \in \mathcal{S}} \lambda_k \theta_k^{*2}} = \frac{a_j^2}{\sum_{k \in \mathcal{S}} a_k^2}$  for  $j \in \mathcal{S}$ , with probability at least  $1 - ctde^{-n^{2\epsilon}}$ .

*Proof.* We start with the lower bound for the support  $|\hat{\theta}_j|$  for  $j \in \mathcal{S}$ ; this term is related to the survival term

$\text{SU}_j$ . According to Lemma 7, for any  $j \in \mathcal{S}$ , we have

$$\left| \hat{\theta}_j \right| = \text{SU}_j \cdot |\theta_j^*| = \text{SU}_j \cdot \frac{|a_j|}{\sqrt{\lambda_j}} \geq \sqrt{\frac{2}{\pi \|\Sigma^{\frac{1}{2}} \theta^*\|_2^2}} \cdot \frac{\lambda_j \left( \frac{n}{c_1 \bar{\lambda}_1 r_0 (\Sigma_{-\mathcal{S}})} \right)}{1 + \lambda_j \left( \frac{c_2 n}{\bar{\lambda}_1 r_0 (\Sigma_{-\mathcal{S}})} \right)} \cdot \frac{|a_j|}{\sqrt{\lambda_j}} = \sqrt{\frac{2\beta_j}{\lambda_j \pi}} \frac{\lambda_j \left( \frac{n}{c_1 \bar{\lambda}_1 r_0 (\Sigma_{-\mathcal{S}})} \right)}{1 + \lambda_j \left( \frac{c_2 n}{\bar{\lambda}_1 r_0 (\Sigma_{-\mathcal{S}})} \right)}. \quad (33)$$

Next, we derive the upper bound of  $\left| \hat{\theta}_j \right|$  for  $j \in \mathcal{S}^c$ . This term is related to CN, but note that CN is the summation of all non-support dimensions. Here, we only need bounds for individual  $\hat{\theta}_j$  for  $j \in \mathcal{S}^c$ . Since the covariance operator satisfies  $r_{k^*}(\Sigma) \geq bn$ , we have two cases such that for  $j \in [k^*] \cap \mathcal{S}^c$  and  $k^* < j \leq d, j \in \mathcal{S}^c$  where they may have different covariance eigenvalue magnitude range. Based on the  $\hat{\theta}_j$  definition in Equation (11), we follow the proof steps in Lemma 7 and get

$$\begin{aligned} \left| \hat{\theta}_j \right| &= \left| \sqrt{\lambda_j} \mathbf{z}_j^\top \mathbf{A}_{-s_1:s_t}^{-1} \check{\mathbf{y}}_{s_t} \right| \\ &= \frac{1}{\sqrt{\lambda_j}} \left| \lambda_j \mathbf{z}_j^\top \mathbf{A}_{-s_1:s_t}^{-1} \check{\mathbf{y}}_{s_t} \right| \\ &= \frac{1}{\sqrt{\lambda_j}} \sqrt{\check{\mathbf{y}}_{s_t}^\top \underbrace{\mathbf{A}_{-s_1:s_t}^{-1} (\lambda_j^2 \mathbf{z}_j \mathbf{z}_j^\top) \mathbf{A}_{-s_1:s_t}^{-1}}_{:= \tilde{\mathbf{C}}_j} \check{\mathbf{y}}_{s_t}} \\ &\leq \frac{1}{\sqrt{\lambda_j}} \left( \sqrt{\hat{\mathbf{y}}^\top \tilde{\mathbf{C}}_j \hat{\mathbf{y}}} + \sum_{\ell=1}^t \sqrt{\left( \frac{\lambda_{s_\ell} \mathbf{z}_{s_\ell}^\top \mathbf{A}_{-s_1:s_\ell}^{-1} \check{\mathbf{y}}_{s_{\ell-1}}}{1 + \lambda_{s_\ell} \mathbf{z}_{s_\ell}^\top \mathbf{A}_{-s_1:s_\ell}^{-1} \mathbf{z}_{s_\ell}} \right)^2 \mathbf{z}_{s_\ell}^\top \tilde{\mathbf{C}}_j \mathbf{z}_{s_\ell}} \right), \end{aligned} \quad (34)$$

where we apply the definition of  $\check{\mathbf{y}}_{s_t}$  such that  $\check{\mathbf{y}}_{s_\ell} := \check{\mathbf{y}}_{s_{\ell-1}} - \frac{\lambda_{s_\ell} \mathbf{z}_{s_\ell}^\top \mathbf{A}_{-s_1:s_\ell}^{-1} \check{\mathbf{y}}_{s_{\ell-1}}}{1 + \lambda_{s_\ell} \mathbf{z}_{s_\ell}^\top \mathbf{A}_{-s_1:s_\ell}^{-1} \mathbf{z}_{s_\ell}} \mathbf{z}_{s_\ell}$  and  $\check{\mathbf{y}}_{s_0} = \hat{\mathbf{y}}$  for  $\ell \in [t]$ .

For the inequality, we apply the triangle inequality such that  $\sqrt{(\mathbf{x} - \mathbf{y})^\top \mathbf{M} (\mathbf{x} - \mathbf{y})} \leq \sqrt{\mathbf{x}^\top \mathbf{M} \mathbf{x}} + \sqrt{\mathbf{y}^\top \mathbf{M} \mathbf{y}}$  a total of  $t$  times. Next, we discuss the upper bounds for  $j \in [k^*] \cap \mathcal{S}^c$  and  $k^* < j \leq d, j \in \mathcal{S}^c$  respectively. For  $j \in [k^*] \cap \mathcal{S}^c$ , by the Sherman-Morrison-Woodbury identity, we have  $\hat{\mathbf{y}}^\top \tilde{\mathbf{C}}_j \hat{\mathbf{y}}$  as

$$\hat{\mathbf{y}}^\top \tilde{\mathbf{C}}_j \hat{\mathbf{y}} = \hat{\mathbf{y}}^\top \mathbf{A}_{-s_1:s_t}^{-1} (\lambda_j^2 \mathbf{z}_j \mathbf{z}_j^\top) \mathbf{A}_{-s_1:s_t}^{-1} \hat{\mathbf{y}} = \frac{\lambda_j^2 \hat{\mathbf{y}}^\top \mathbf{A}_{-(s_1:s_t) \cup (j)}^{-1} \mathbf{z}_j \mathbf{z}_j^\top \mathbf{A}_{-(s_1:s_t) \cup (j)}^{-1} \hat{\mathbf{y}}}{\left( 1 + \lambda_j \mathbf{z}_j^\top \mathbf{A}_{-(s_1:s_t) \cup (j)}^{-1} \mathbf{z}_j \right)^2} \leq \frac{\left( \mathbf{z}_j^\top \mathbf{A}_{-(s_1:s_t) \cup (j)}^{-1} \hat{\mathbf{y}} \right)^2}{\left( \mathbf{z}_j^\top \mathbf{A}_{-(s_1:s_t) \cup (j)}^{-1} \mathbf{z}_j \right)^2}.$$

Next, for  $j \in [k^*] \cap \mathcal{S}^c$ , by the Hanson-Wright inequality (Lemma 32), we have the upper bound for  $\sqrt{\hat{\mathbf{y}}^\top \tilde{\mathbf{C}}_j \hat{\mathbf{y}}}$  as

$$\begin{aligned} \sqrt{\hat{\mathbf{y}}^\top \tilde{\mathbf{C}}_j \hat{\mathbf{y}}} &\leq \sqrt{\frac{c_1^2 \left\| \mathbf{A}_{-(s_1:s_t) \cup (j)}^{-1} \right\|_2^2 \cdot n^{1+2\epsilon}}{\left( \text{Tr} \left( \mathbf{A}_{-(s_1:s_t) \cup (j)}^{-1} \right) - c_1 \left\| \mathbf{A}_{-(s_1:s_t) \cup (j)}^{-1} \right\|_2 \cdot n^{\frac{1}{2}+\epsilon} \right)^2}} \\ &\leq \sqrt{\frac{c_1^2 \left\| \mathbf{A}_{-(s_1:s_t) \cup [k^*]}^{-1} \right\|_2^2 \cdot n^{1+2\epsilon}}{\left( \text{Tr} \left( \mathbf{A}_{-(s_1:s_t) \cup (j)}^{-1} \right) - c_1 \left\| \mathbf{A}_{-(s_1:s_t) \cup [k^*]}^{-1} \right\|_2 \cdot n^{\frac{1}{2}+\epsilon} \right)^2}}, \end{aligned} \quad (35)$$

where in the second inequality, we use the fact that  $\mathbf{A}_{-(s_1:s_t) \cup [k^*]}^{-1} \succeq \mathbf{A}_{-(s_1:s_t) \cup (j)}^{-1}$  for all  $j \in [k^*]$ . Next, we

use Lemma 36 to get the lower bound for the trace term in the denominator, and get

$$\begin{aligned}
(35) &\leq \sqrt{\frac{c_1^2 \left\| \mathbf{A}_{-(s_1:s_t) \cup [k^*]}^{-1} \right\|_2^2 \cdot n^{1+2\epsilon}}{\left( \left(1 - \frac{c}{n}\right)^{k^*-1} \text{Tr} \left( \mathbf{A}_{-(s_1:s_t) \cup [k^*]}^{-1} \right) - c_1 \left\| \mathbf{A}_{-(s_1:s_t) \cup [k^*]}^{-1} \right\|_2 \cdot n^{\frac{1}{2}+\epsilon} \right)^2}} \\
&= \sqrt{\frac{1}{\left( \left(1 - \frac{c}{n}\right)^{k^*-1} \frac{\text{Tr} \left( \mathbf{A}_{-(s_1:s_t) \cup [k^*]}^{-1} \right)}{c_1 \left\| \mathbf{A}_{-(s_1:s_t) \cup [k^*]}^{-1} \right\|_2 \cdot n^{\frac{1}{2}+\epsilon}} - 1 \right)^2}} \quad (36)
\end{aligned}$$

Finally, we apply Lemma 35 to show that eigenvalues in  $\mathbf{A}_{-(s_1:s_t) \cup [k^*]}^{-1}$  are identical up to a constant.

$$(36) \leq \sqrt{\frac{1}{\left( \left(1 - \frac{c}{n}\right)^{k^*-1} \frac{n^{\frac{1}{2}-\epsilon}}{c_2} - 1 \right)^2}} \leq \sqrt{\frac{c_3}{n^{1-2\epsilon}}},$$

Since  $\mathbf{z}_{s_\ell}$  is independent to  $\tilde{\mathbf{C}}_j$  for  $j \in [k^*] \cap \mathcal{S}^c$  and  $\ell \in [k^*]$ , following the same procedure, we can show  $\sqrt{\mathbf{z}_{s_\ell}^\top \tilde{\mathbf{C}}_j \mathbf{z}_{s_\ell}}$  is upper bounded in the same rate. As a result, from Equation (34) we have

$$\begin{aligned}
|\hat{\theta}_j| &\leq \frac{1}{\sqrt{\lambda_j}} \left( 1 + \sum_{\ell=1}^t \left| \frac{\lambda_{s_\ell} \mathbf{z}_{s_\ell}^\top \mathbf{A}_{-s_1:s_\ell}^{-1} \tilde{\mathbf{y}}_{s_{\ell-1}}}{1 + \lambda_{s_\ell} \mathbf{z}_{s_\ell}^\top \mathbf{A}_{-s_1:s_\ell}^{-1} \mathbf{z}_{s_\ell}} \right| \right) \sqrt{\frac{c_3}{n^{1-2\epsilon}}} \\
&\leq \frac{1}{\sqrt{\lambda_j}} (1 + tc) \sqrt{\frac{c_3}{n^{1-2\epsilon}}},
\end{aligned}$$

where we apply Lemma 17 in the last inequality. Next, starting from Equation (34) again, we show the upper bound of  $|\hat{\theta}_j|$  for  $k^* < j \leq d, j \in \mathcal{S}^c$ . By the Sherman-Morrison-Woodbury identity, we have  $\hat{\mathbf{y}}^\top \tilde{\mathbf{C}}_j \hat{\mathbf{y}}$  as

$$\begin{aligned}
\sqrt{\hat{\mathbf{y}}^\top \tilde{\mathbf{C}}_j \hat{\mathbf{y}}} &= \sqrt{\hat{\mathbf{y}}^\top \mathbf{A}_{-s_1:s_t}^{-1} (\lambda_j^2 \mathbf{z}_j \mathbf{z}_j^\top) \mathbf{A}_{-s_1:s_t}^{-1} \hat{\mathbf{y}}} \\
&= \sqrt{\frac{\lambda_j^2 \hat{\mathbf{y}}^\top \mathbf{A}_{-(s_1:s_t) \cup (j)}^{-1} \mathbf{z}_j \mathbf{z}_j^\top \mathbf{A}_{-(s_1:s_t) \cup (j)}^{-1} \hat{\mathbf{y}}}{\left( 1 + \lambda_j \mathbf{z}_j^\top \mathbf{A}_{-(s_1:s_t) \cup (j)}^{-1} \mathbf{z}_j \right)^2}} \\
&\leq \sqrt{\lambda_j^2 \hat{\mathbf{y}}^\top \mathbf{A}_{-(s_1:s_t) \cup (j)}^{-1} \mathbf{z}_j \mathbf{z}_j^\top \mathbf{A}_{-(s_1:s_t) \cup (j)}^{-1} \hat{\mathbf{y}}}. \quad (37)
\end{aligned}$$

Next, we apply the Hanson-Wright inequality (Lemma 32) to obtain

$$\begin{aligned}
(37) &\leq \sqrt{\lambda_j^2 \left( 2c_1 \left\| \mathbf{A}_{-(s_1:s_t) \cup (j)}^{-1} \right\|_2 \cdot n^{\frac{1}{2}+\epsilon} \right)^2} \\
&\leq \sqrt{\lambda_j^2 \left( 2c_1 \left\| \mathbf{A}_{-(s_1:s_t) \cup (j) \cup [k^*]}^{-1} \right\|_2 \cdot n^{\frac{1}{2}+\epsilon} \right)^2} \\
&= \sqrt{\lambda_j^2 \left( \frac{2c_1 \cdot n^{\frac{1}{2}+\epsilon}}{\mu_n(\mathbf{A}_{-(s_1:s_t) \cup (j) \cup [k^*]})} \right)^2}, \quad (38)
\end{aligned}$$

where in the second inequality, we use the property that  $\mathbf{A}_{-(s_1:s_t) \cup (j)}^{-1} \preceq \mathbf{A}_{-(s_1:s_t) \cup (j) \cup [k^*]}^{-1}$  for  $k^* < j \leq d, j \in \mathcal{S}^c$ . Finally, we apply Lemma 34 to show that  $\Sigma_{-(s_1:s_t) \cup (j) \cup [k^*]}$  satisfies  $r_0(\Sigma_{-(s_1:s_t) \cup (j) \cup [k^*]}) \geq bn$ . Therefore, by Lemma 10 in Bartlett et al. (2020), we have

$$\mu_n(\mathbf{A}_{-(s_1:s_t) \cup (j) \cup [k^*]}) \geq \frac{1}{c} \left( \sum_{k=k^*+1, k \in \mathcal{S}^c, k \neq j} \lambda_k \right).$$



Applying the inequality to Equation (38), we get

$$(38) \leq \sqrt{\lambda_j^2 \left( \frac{c_2 \cdot n^{1+2\epsilon}}{\left( \sum_{k=k^*+1, k \in \mathcal{S}^c, k \neq j} \lambda_k \right)^2} \right)}. \quad (39)$$

Next, we want to upper bound  $\frac{1}{\sum_{k=k^*+1, k \in \mathcal{S}^c, k \neq j} \lambda_k}$  by  $\frac{c}{\sum_{k=k^*+1, k \in \mathcal{S}^c} \lambda_k}$ , and we have

$$\begin{aligned} \frac{1}{\sum_{k=k^*+1, k \in \mathcal{S}^c, k \neq j} \lambda_k} &= \left( \frac{1}{\sum_{k=k^*+1, k \in \mathcal{S}^c} \lambda_k} \right) \left( \frac{\sum_{k=k^*+1, k \in \mathcal{S}^c} \lambda_k}{\sum_{k=k^*+1, k \in \mathcal{S}^c, k \neq j} \lambda_k} \right) \\ &= \left( \frac{1}{\sum_{k=k^*+1, k \in \mathcal{S}^c} \lambda_k} \right) \left( 1 + \frac{\lambda_j}{\sum_{k=k^*+1, k \in \mathcal{S}^c, k \neq j} \lambda_k} \right) \end{aligned} \quad (40)$$

According to the effective rank of  $\Sigma$ , for  $k^* < j \leq d, j \in \mathcal{S}^c$ , we have  $\sum_{k=k^*+1, k \in \mathcal{S}^c} \lambda_k \geq b\lambda_{k^*+1}n \geq b\lambda_jn$ . By deducting  $\lambda_j$  on both side, we get

$$\sum_{k=k^*+1, k \in \mathcal{S}^c, k \neq j} \lambda_k \geq b\lambda_jn - \lambda_j.$$

Therefore, we can write

$$(40) \leq \left( \frac{1}{\sum_{k=k^*+1, k \in \mathcal{S}^c} \lambda_k} \right) \left( 1 + \frac{\lambda_j}{b\lambda_jn - \lambda_j} \right) = \left( \frac{1}{\sum_{k=k^*+1, k \in \mathcal{S}^c} \lambda_k} \right) \left( 1 + \frac{1}{bn - 1} \right).$$

Substituting this inequality into Equation (39), we finish the upper bound for  $\sqrt{\hat{\mathbf{y}}^\top \tilde{\mathbf{C}}_j \hat{\mathbf{y}}}$ . Follow this procedure, we can show  $\sqrt{\mathbf{z}_{s_\ell}^\top \tilde{\mathbf{C}}_j \mathbf{z}_{s_\ell}}$  is upper bounded by the same term. As a result, we have for  $k^* < j \leq d, j \in \mathcal{S}^c$ ,

$$\begin{aligned} |\hat{\theta}_j| &\leq \frac{1}{\sqrt{\lambda_j}} \left( 1 + \sum_{\ell=1}^t \left| \frac{\lambda_{s_\ell} \mathbf{z}_{s_\ell}^\top \mathbf{A}_{-s_1:s_\ell}^{-1} \tilde{\mathbf{y}}_{s_{\ell-1}}}{1 + \lambda_{s_\ell} \mathbf{z}_{s_\ell}^\top \mathbf{A}_{-s_1:s_\ell}^{-1} \mathbf{z}_{s_\ell}} \right| \right) \sqrt{\lambda_j^2 \frac{c_2 n^{1+2\epsilon}}{\left( \sum_{k=k^*+1, k \in \mathcal{S}^c} \lambda_k \right)^2}} \\ &\leq (1 + tc) \sqrt{\frac{c_2 n^{1+2\epsilon}}{\left( \sum_{k=k^*+1, k \in \mathcal{S}^c} \lambda_k \right)^2}}, \end{aligned}$$

where we apply Lemma 17 in the last inequality. This completes the proof of the lemma.  $\square$

Equipped with Lemma 26, we can now prove Theorem 14. While it restricts to the case where  $\boldsymbol{\theta}^*$  is supported only in the top  $k^*$  indices of the covariance spectrum, we handle support recovery outside the top  $k^*$  indices in Section F.3.

*Proof.* (Theorem 14)

We show that as long as one of the following conditions hold, the support recovery is guaranteed: (1)  $t$  is known and

$$\lambda_j \ll \frac{\lambda_q \cdot n^{1-2\epsilon}}{t^2} \text{ and } \lambda_j \ll \frac{\left( \tilde{\lambda}_1 r_0 \left( \Sigma_{-[k^*]} \right) \right)^2}{t^2 \cdot (n^{1+2\epsilon}) \cdot \lambda_\ell}$$

hold for all  $j \in \mathcal{S}, q \in \mathcal{S}^c \cap [k^*], \ell > k^*$  or (2)  $\Sigma$  is known. To start with, we first show the support recovery under condition (1). By assumption that  $\boldsymbol{\theta}^*$  is supported only in the top  $k^*$  indices of the covariance spectrum,

we have  $j \in [k^*]$  for  $j \in \mathcal{S}$ . According to Lemma 26, we have the lower bound of  $\hat{\theta}$  as

$$|\hat{\theta}_j| \geq \sqrt{\frac{2\beta_j}{\lambda_j \pi}} \frac{\lambda_j \left( \frac{n}{c_1 \tilde{\lambda}_1 r_0(\Sigma_{-\mathcal{S}})} \right)}{1 + \lambda_j \left( \frac{c_2 n}{\tilde{\lambda}_1 r_0(\Sigma_{-\mathcal{S}})} \right)} = \sqrt{\frac{2\beta_j}{\lambda_j \pi}} \frac{1}{\frac{c_1 \tilde{\lambda}_1 r_0(\Sigma_{-\mathcal{S}})}{\lambda_j n} + c_1 c_2} \geq \sqrt{\frac{2\beta_j}{\lambda_j \pi}} \frac{1}{\frac{c_1 \tilde{\lambda}_1 r_0(\Sigma_{-\mathcal{S}})}{\tilde{\lambda}_1 r_0(\Sigma_{-\mathcal{S}})} + c_1 c_2} = \frac{c_3}{\sqrt{\lambda_j}}, \quad (41)$$

where we apply  $\tilde{\lambda}_1 r_0(\Sigma_{-\mathcal{S}}) = \sum_{k=k^*+1, k \in \mathcal{S}^c}^d \lambda_k \leq \lambda_{k^*} n \leq \lambda_j n$  for  $j \in [k^*]$  by the definition of  $k^*$ . As a result, support is lower bounded in the rate of  $\frac{1}{\sqrt{\lambda_j}}$ . Next, for non-support indices  $j \in \mathcal{S}^c \cap [k^*]$ , by Lemma 26, we have the upper bound by

$$|\hat{\theta}_j| \leq t \sqrt{\frac{c}{\lambda_j n^{1-2\epsilon}}}. \quad (42)$$

Therefore, for  $j \in \mathcal{S}^c \cap [k^*]$ , non-support is upper bounded in the rate of  $\frac{1}{\sqrt{\lambda_j}} \cdot \left( \frac{t}{\sqrt{n^{1-2\epsilon}}} \right)$ . Lastly, for non-support indices  $j > k^*$ , by Lemma 26, we have the upper bound by

$$|\hat{\theta}_j| \leq t \sqrt{\frac{cn^{1+2\epsilon}\lambda_j}{\left( \tilde{\lambda}_1 r_0(\Sigma_{-\mathcal{S}}) \right)^2}}. \quad (43)$$

Therefore, for  $j > k^*$ , non-support is upper bounded in the rate of  $\frac{1}{\sqrt{\lambda_j}} \cdot \sqrt{\frac{t^2 n^{1+2\epsilon} \lambda_j^2}{\left( \tilde{\lambda}_1 r_0(\Sigma_{-\mathcal{S}}) \right)^2}}$ . As a result, we need the following conditions to guarantee the lower bound of the support indices is larger than the upper bound of the non-support indices:

$$\begin{aligned} \frac{1}{\sqrt{\lambda_j}} &\gg \frac{t}{\sqrt{\lambda_k \cdot n^{1-2\epsilon}}} \Leftrightarrow \frac{\lambda_k \cdot n^{1-2\epsilon}}{t^2} \gg \lambda_j \\ \frac{1}{\sqrt{\lambda_j}} &\gg t \sqrt{\frac{n^{1+2\epsilon} \lambda_\ell}{\left( \tilde{\lambda}_1 r_0(\Sigma_{-\mathcal{S}}) \right)^2}} \Leftrightarrow \frac{\left( \tilde{\lambda}_1 r_0(\Sigma_{-\mathcal{S}}) \right)^2}{t^2 \cdot (n^{1+2\epsilon}) \cdot \lambda_\ell} \gg \lambda_j \end{aligned}$$

for all  $j \in \mathcal{S}$  and  $k \in \mathcal{S}^c \cap [k^*]$  and  $\ell > k^*$ . In this way, since we also know  $t$ , we can achieve support recovery by choosing the largest  $t$  indices of  $\hat{\theta}$ .

On the other hand, for condition (2), if we have access to all the eigenvalues  $\{\lambda_j\}_{j=1}^d$  of  $\Sigma$ , we can determine whether  $\hat{\theta}_j$  is a support by checking if  $|\hat{\theta}_j| = \frac{c}{\sqrt{\lambda_j}}$  for some constant  $c > 0$  according to the support lower bound in Equation (41). For non-support rate, we can further upper bound Equation (43) by  $t \sqrt{\frac{c}{\lambda_j n^{1-2\epsilon}}}$  since

$$t \sqrt{\frac{cn^{1+2\epsilon}\lambda_j}{\left( \tilde{\lambda}_1 r_0(\Sigma_{-\mathcal{S}}) \right)^2}} \leq t \sqrt{\frac{cn^{1+2\epsilon}\lambda_j}{(b\lambda_j n)^2}} = \frac{c_1 t}{\sqrt{\lambda_j n^{1-2\epsilon}}}, \quad (44)$$

where we apply  $\tilde{\lambda}_1 r_0(\Sigma_{-\mathcal{S}}) = \sum_{k=k^*+1, k \in \mathcal{S}^c}^d \lambda_k \geq b\lambda_j n$  for  $j > k^*$ . As a result, according to the non-support upper bound in Equation (42) and Equation (44),  $\hat{\theta}_j$  can be classified as non-support if  $|\hat{\theta}_j| = o\left(\frac{t}{\sqrt{\lambda_j n^{1-2\epsilon}}}\right)$ , which decays at a rate proportional to  $\frac{t}{\sqrt{n^{1-2\epsilon}}}$ . This completes the proof of the theorem.  $\square$

## F.2 Spiked and polynomial decay covariance: support inside the top $k^*$ indices

In this section, we characterize different covariance matrices wherein  $\theta^*$  is supported only in the top  $k^*$  indices of the covariance spectrum and we demonstrate the support identification guarantee (Theorem 14) is satisfied.

Corollary 27 demonstrates the characterization in spiked covariance defined in Definition 3, and Corollary 28 and Corollary 29 show the characterization in polynomial decay covariance defined in Definition 4. Note that we do not allow the case  $k^* = 0$  (e.g., spiked covariance with  $q > 1 - r$ ) in this section. The analysis closely follows Lemmas 32 and 34 of Muthukumar et al. (2021).

**Corollary 27.** *Under Assumptions 2 and 5 with  $t \ll n^{\frac{1}{2}-\epsilon}$  and spiked covariance matrix (Definition 3), we assume support are all in the top  $k^*$  indices of the covariance spectrum. By substituting spiked parameters in Definition 3, for  $q < (1 - r)$ , we have*

$$|\hat{\theta}_j| \begin{cases} \gg n^{\frac{-p+q+r}{2}} & \text{for } j \in \mathcal{S} \\ \ll n^{\frac{-p+q+r}{2}} & \text{for } j \in [k^*] \cap \mathcal{S}^c \\ \ll n^{\frac{-2p+2}{2}} & \text{for } k^* < j \leq d \end{cases}.$$

Therefore, if  $p > 2 - q - r$ , we can pick a threshold between bounds to distinguish support and non-support such as  $T = n^{\frac{-p+q+r}{2}}$ .

*Proof.* Recall the spiked covariance in Definition 3 such that

$$\lambda_j := \begin{cases} \frac{ad}{s} & j \in [s] \\ \frac{(1-a)d}{d-s} & \text{otherwise.} \end{cases}$$

First, we need to make sure the conditions in Theorem 14 scenario (1) are satisfied such that

$$\begin{aligned} \lambda_j &\ll \frac{\lambda_q \cdot n^{1-2\epsilon}}{t^2}, \\ \lambda_j &\ll \frac{\left(\tilde{\lambda}_1 r_0 (\Sigma_{-[k^*]})\right)^2}{t^2 \cdot (n^{1+2\epsilon}) \cdot \lambda_\ell}, \end{aligned}$$

for all  $j \in \mathcal{S}$ ,  $q \in \mathcal{S}^c \cap [k^*]$ , and  $\ell > k^*$ . For the first condition, we show it holds because

$$\frac{\lambda_q \cdot n^{1-2\epsilon}}{t^2} = \frac{\frac{ad}{s} \cdot n^{1-2\epsilon}}{t^2} \gg \frac{ad}{s} = \lambda_j,$$

since  $\lambda_j = \lambda_q$  for  $j \in \mathcal{S}$ ,  $q \in \mathcal{S}^c \cap [k^*]$ , and  $t \ll n^{\frac{1}{2}-\epsilon}$ . For the second condition, we can show the right-hand side as

$$\frac{\left(\tilde{\lambda}_1 r_0 (\Sigma_{-[k^*]})\right)^2}{t^2 \cdot (n^{1+2\epsilon}) \cdot \lambda_\ell} = \frac{\left((d-s) \cdot \frac{(1-a)d}{(d-s)}\right)^2}{t^2 \cdot (n^{1+2\epsilon}) \cdot \frac{(1-a)d}{(d-s)}} = \frac{(1-a)d \cdot (d-s) \cdot n^{1-2\epsilon}}{t^2 \cdot n^2} > \frac{1}{c} n^{2p-2}$$

for all  $j \in \mathcal{S}$ ,  $q \in \mathcal{S}^c \cap [k^*]$ , and  $\ell > k^*$ . On the left-hand side, we have

$$\lambda_j = \frac{ad}{s} = n^{p-q-r}.$$

Since the corollary assumes  $p > 2 - q - r$ , the second condition holds. Next, we show the precise threshold

value by using Lemma 26. We have the lower bound for support indices for  $j \in \mathcal{S}$  as

$$\begin{aligned}
|\hat{\theta}_j| &\geq \sqrt{\frac{2\beta_j}{\lambda_j \pi}} \frac{\lambda_j \left( \frac{n}{c_1 \tilde{\lambda}_1 r_0 (\Sigma_{-\bar{\mathcal{S}}})} \right)}{1 + \lambda_j \left( \frac{c_2 n}{\tilde{\lambda}_1 r_0 (\Sigma_{-\bar{\mathcal{S}}})} \right)} \\
&= \sqrt{\frac{2\beta_j}{\pi}} \sqrt{\frac{s}{ad}} \left( \frac{\frac{ad}{s} \left( \frac{n}{c_1 (d-|\bar{\mathcal{S}}|) \frac{(1-a)d}{(d-s)}} \right)}{1 + \frac{ad}{s} \left( \frac{c_2 n}{(d-|\bar{\mathcal{S}}|) \frac{(1-a)d}{(d-s)}} \right)} \right) \\
&= \sqrt{\frac{2\beta_j}{\pi}} n^{\frac{-p+q+r}{2}} \frac{n^{p-q-r} \left( \frac{n}{c_1 (n^p-|\bar{\mathcal{S}}|) \frac{(1-n^{-q})n^p}{(n^p-n^r)}} \right)}{1 + n^{p-q-r} \left( \frac{c_2 n}{(n^p-|\bar{\mathcal{S}}|) \frac{(1-n^{-q})n^p}{(n^p-n^r)}} \right)} \\
&\geq \sqrt{\frac{2\beta_j}{\pi}} n^{\frac{-p+q+r}{2}} \cdot \frac{\frac{1}{c_3} \cdot n^{(1-r)-q}}{1 + c_4 n^{(1-r)-q}},
\end{aligned}$$

where we substitute  $\lambda_j = \frac{ad}{s} = n^{p-q-r}$  and  $\tilde{\lambda}_1 r_0 (\Sigma_{-\bar{\mathcal{S}}}) = (d-|\bar{\mathcal{S}}|) \frac{(1-a)d}{d-s}$ . When  $q < (1-r)$ , the  $n^{(1-r)-q}$  term dominates in the fraction part. Hence, we have

$$|\hat{\theta}_j| \geq \sqrt{\frac{2\beta_j}{\pi}} n^{\frac{-p+q+r}{2}} \cdot c_{10} \left( 1 + cn^{q-(1-r)} \right)^{-1}.$$

For non-support upper bound, we need to consider indices in  $j \in [k^*] \cap \mathcal{S}^c$  and indices in  $j > k^*$ . According to Lemma 26, for  $j \in [k^*] \cap \mathcal{S}^c$ , we have  $\lambda_j = \frac{ad}{s} = n^{p-q-r}$  and we get

$$|\hat{\theta}_j| \leq t \sqrt{\frac{c}{\lambda_j n^{1-2\epsilon}}} \leq c_1 t n^{\frac{-p+q+r-1+2\epsilon}{2}} < c_1 n^{\frac{-p+q+r}{2}},$$

since  $t \ll n^{\frac{1}{2}-\epsilon}$ . On the other hand, for  $j > k^*$ , according to Lemma 26, we substitute  $\lambda_j = \frac{(1-a)d}{d-s} = \frac{(1-n^{-q}) \cdot n^p}{n^p-n^r}$  and  $\tilde{\lambda}_1 r_0 (\Sigma_{-\bar{\mathcal{S}}}) = (d-|\bar{\mathcal{S}}|) \frac{(1-a)d}{d-s}$  and we have

$$\begin{aligned}
|\hat{\theta}_j| &\leq t \sqrt{\frac{c}{\lambda_j} \frac{n^{1+2\epsilon} \lambda_j^2}{(\tilde{\lambda}_1 r_0 (\Sigma_{-\bar{\mathcal{S}}}))^2}} \\
&= t \sqrt{\frac{c(d-s)}{(1-a)d} \left( \frac{n^{1+2\epsilon} \left( \frac{(1-a)d}{d-s} \right)^2}{\left( (d-|\bar{\mathcal{S}}|) \frac{(1-a)d}{d-s} \right)^2} \right)} \\
&= t \sqrt{\frac{c(n^p-n^r)}{(1-n^{-q}) \cdot n^p} \left( \frac{n^{1+2\epsilon}}{(n^p-|\bar{\mathcal{S}}|)^2} \right)} \\
&\leq c_1 t n^{\frac{-2p+1+2\epsilon}{2}} \\
&< c_1 n^{\frac{-2p+2}{2}},
\end{aligned}$$

since  $t \ll n^{\frac{1}{2}-\epsilon}$ . As a result, since we assume  $p > 2 - q - r$ , we have support lower bound larger than the non-support upper bound, and the support identification is guaranteed.  $\square$

**Corollary 28.** *Under Assumptions 2 and 5 with  $t \ll n^{\frac{1}{2}-\epsilon}$  and polynomial decay covariance in Definition 4 with  $u = 1, v = 2$ , we assume support are all in the top  $k^*$  indices of the covariance spectrum. By substituting*

$\lambda_j = \frac{1}{j \cdot \ln^2(j+1)}$ , we have

$$|\hat{\theta}_j| \begin{cases} \gg \frac{1}{\sqrt{\lambda_j}} & \text{for } j \in \mathcal{S} \\ \ll \frac{t}{\sqrt{\lambda_j \cdot n^{1-2\epsilon}}} & \text{for } j \in [k^*] \cap \mathcal{S}^c \\ \ll \frac{t}{\sqrt{\lambda_j \cdot n^{1-2\epsilon}}} & \text{for } k^* < j \leq d \end{cases}.$$

If we get access to all  $\{\lambda_j\}_{j=1}^d$ , we can therefore distinguish support and non-support by examining each  $\hat{\theta}_j$  has order  $\omega\left(\frac{1}{\sqrt{\lambda_j}}\right)$  or decay in a rate of  $o\left(\frac{t}{\sqrt{\lambda_j \cdot n^{1-2\epsilon}}}\right)$ .

*Proof.* Since the conditions in the first scenario in Theorem 14 are not satisfied, we need to assume  $\Sigma$  is known. Next, we show the rate of supports and non-supports based on Lemma 26. We have

$$\begin{aligned} |\hat{\theta}_j| &\geq \sqrt{\frac{2\beta_j}{\lambda_j \pi}} \frac{\lambda_j \left( \frac{n}{c_1 \tilde{\lambda}_1 r_0(\Sigma_{-\mathcal{S}})} \right)}{1 + \lambda_j \left( \frac{c_2 n}{\tilde{\lambda}_1 r_0(\Sigma_{-\mathcal{S}})} \right)}, & \text{for } j \in \mathcal{S}, \\ |\hat{\theta}_j| &\leq t \sqrt{\frac{c}{\lambda_j n^{1-2\epsilon}}}, & \text{for } j \in [k^*] \cap \mathcal{S}^c, \\ |\hat{\theta}_j| &\leq t \sqrt{\frac{n^{1+2\epsilon} \lambda_j}{\left( \tilde{\lambda}_1 r_0(\Sigma_{-\mathcal{S}}) \right)^2}}, & \text{for } k^* < j \leq d, j \in \mathcal{S}^c. \end{aligned}$$

For polynomial decay covariance, we have  $\lambda_j = \frac{1}{j \cdot \ln^2(j+1)}$  for  $j \in [d]$ , and also  $\sum_{j=1}^{\infty} \lambda_j = \sum_{j=1}^{\infty} \frac{1}{j \cdot \ln^2(j+1)} = \mathcal{O}(1)$ . Therefore, for  $j \in \mathcal{S}$ , we have

$$|\hat{\theta}_j| \geq \sqrt{\frac{2\beta_j}{\lambda_j \pi}} \frac{\lambda_j \left( \frac{n}{c_1 \tilde{\lambda}_1 r_0(\Sigma_{-\mathcal{S}})} \right)}{1 + \lambda_j \left( \frac{c_2 n}{\tilde{\lambda}_1 r_0(\Sigma_{-\mathcal{S}})} \right)} = \sqrt{\frac{2\beta_j}{\lambda_j \pi}} \frac{1}{\frac{c_1 \tilde{\lambda}_1 r_0(\Sigma_{-\mathcal{S}})}{\lambda_j n} + c_1 c_2} \geq \sqrt{\frac{2\beta_j}{\lambda_j \pi}} \frac{1}{\frac{c_1 \tilde{\lambda}_1 r_0(\Sigma_{-\mathcal{S}})}{\tilde{\lambda}_1 r_0(\Sigma_{-\mathcal{S}})} + c_1 c_2} = \frac{c_3}{\sqrt{\lambda_j}}$$

where we apply  $\tilde{\lambda}_1 r_0(\Sigma_{-\mathcal{S}}) = \sum_{j=k^*+1}^d \lambda_j \leq \lambda_{k^*} n \leq \lambda_j n$  for  $j \in [k^*]$  by the definition of  $k^*$ . For non-support,  $j \in [k^*] \cap \mathcal{S}^c$ , we have

$$|\hat{\theta}_j| \leq t \sqrt{\frac{c}{\lambda_j n^{1-2\epsilon}}}.$$

For  $j > k^*, j \in \mathcal{S}^c$ , we have

$$|\hat{\theta}_j| \leq t \sqrt{\frac{c}{\lambda_j} \frac{n^{1+2\epsilon} \lambda_j^2}{\left( \tilde{\lambda}_1 r_0(\Sigma_{-\mathcal{S}}) \right)^2}} \leq t \sqrt{\frac{c}{\lambda_j} \frac{n^{1+2\epsilon} \lambda_j^2}{(b \lambda_j n)^2}} = t \sqrt{\frac{c_1}{\lambda_j n^{1-2\epsilon}}},$$

where we apply  $\tilde{\lambda}_1 r_0(\Sigma_{-\mathcal{S}}) = \sum_{j=k^*+1}^d \lambda_j \geq b \lambda_j n$  for  $j > k^*$ . □

**Corollary 29.** Under Assumptions 2 and 5 with  $t \ll n^{\frac{1}{2}-\epsilon}$  and polynomial decay covariance matrix (Definition 4) with  $u \in (0, 1), v = 0$ , we assume support are all in the top  $k^*$  indices of the covariance spectrum. By substituting  $d = n^p$  and  $\lambda_j = \frac{1}{j^u}$ , for  $p \cdot (1 - u) < 1$ , we have

$$|\hat{\theta}_j| \begin{cases} \gg \frac{1}{\sqrt{\lambda_j}} & \text{for } j \in \mathcal{S} \\ \ll \frac{t}{\sqrt{\lambda_j \cdot n^{1-2\epsilon}}} & \text{for } j \in [k^*] \cap \mathcal{S}^c \\ \ll \frac{t}{\sqrt{\lambda_j \cdot n^{1-2\epsilon}}} & \text{for } k^* < j \leq d \end{cases}.$$

If we get access to all  $\{\lambda_j\}_{j=1}^d$ , we can therefore distinguish support and non-support by examining whether each  $\hat{\theta}_j$  has order  $\omega\left(\frac{1}{\sqrt{\lambda_j}}\right)$  or decays in a rate of  $o\left(\frac{t}{\sqrt{\lambda_j \cdot n^{1-2\epsilon}}}\right)$ .

*Proof.* Since the conditions in the first scenario in Theorem 14 are not satisfied, we need to assume  $\Sigma$  is known. Next, we show the rate of supports and non-supports based on Lemma 26. We have

$$\begin{aligned} |\hat{\theta}_j| &\geq \sqrt{\frac{2\beta_j}{\lambda_j \pi}} \frac{\lambda_j \left( \frac{n}{c_1 \tilde{\lambda}_1 r_0(\Sigma_{-\mathcal{S}})} \right)}{1 + \lambda_j \left( \frac{c_2 n}{\tilde{\lambda}_1 r_0(\Sigma_{-\mathcal{S}})} \right)}, & \text{for } j \in \mathcal{S}, \\ |\hat{\theta}_j| &\leq t \sqrt{\frac{c}{\lambda_j n^{1-2\epsilon}}}, & \text{for } j \in [k^*] \cap \mathcal{S}^c, \\ |\hat{\theta}_j| &\leq t \sqrt{\frac{c}{\lambda_j} \frac{n^{1+2\epsilon} \lambda_j^2}{\left( \tilde{\lambda}_1 r_0(\Sigma_{-\mathcal{S}}) \right)^2}}, & \text{for } k^* < j \leq d, j \in \mathcal{S}^c. \end{aligned}$$

For polynomial decay covariance, we have  $\lambda_j = \frac{1}{j^u}$  for  $j \in [d]$ . We have the lower bound for support indices for  $j \in \mathcal{S}$  as

$$\begin{aligned} |\hat{\theta}_j| &\geq \sqrt{\frac{2\beta_j}{\lambda_j \pi}} \frac{\lambda_j \left( \frac{n}{c_1 \tilde{\lambda}_1 r_0(\Sigma_{-\mathcal{S}})} \right)}{1 + \lambda_j \left( \frac{c_2 n}{\tilde{\lambda}_1 r_0(\Sigma_{-\mathcal{S}})} \right)} \\ &= \frac{1}{\sqrt{\lambda_j}} \sqrt{\frac{2\beta_j}{\pi}} \left( \frac{\frac{1}{j^u} \left( \frac{n}{c_1 \sum_{k>k^*, k \in \mathcal{S}^c} \frac{1}{k^u}} \right)}{1 + \frac{1}{j^u} \left( \frac{c_2 n}{\sum_{k>k^*, k \in \mathcal{S}^c} \frac{1}{k^u}} \right)} \right) \\ &= \frac{1}{\sqrt{\lambda_j}} \sqrt{\frac{2\beta_j}{\pi}} \left( \frac{1}{\frac{c_1 \sum_{k>k^*, k \in \mathcal{S}^c} \frac{1}{k^u}}{\frac{1}{j^u} n} + c_1 c_2} \right) \\ &\geq \frac{1}{\sqrt{\lambda_j}} \sqrt{\frac{2\beta_j}{\pi}} \left( \frac{1}{\frac{c_1 \sum_{k>k^*, k \in \mathcal{S}^c} \frac{1}{k^u}}{\sum_{k>k^*, k \in \mathcal{S}^c} \frac{1}{k^u}} + c_1 c_2} \right) \\ &= \frac{c_3}{\sqrt{\lambda_j}}, \end{aligned}$$

where we substitute  $\lambda_j = \frac{1}{j^u}$  and  $\tilde{\lambda}_1 r_0(\Sigma_{-\mathcal{S}}) = \sum_{j>k^*, j \in \mathcal{S}^c} \frac{1}{j^u}$ , and by the definition of  $k^*$ , we apply  $\frac{1}{j^u} n \geq \sum_{k>k^*, k \in \mathcal{S}^c} \frac{1}{k^u}$  for  $j \in [k^*]$  in the last inequality. For non-support,  $j \in [k^*] \cap \mathcal{S}^c$ , we have

$$|\hat{\theta}_j| \leq t \sqrt{\frac{c}{\lambda_j n^{1-2\epsilon}}}.$$

For  $j > k^*, j \in \mathcal{S}^c$ , we have

$$|\hat{\theta}_j| \leq t \sqrt{\frac{c}{\lambda_j} \frac{n^{1+2\epsilon} \lambda_j^2}{\left( \tilde{\lambda}_1 r_0(\Sigma_{-\mathcal{S}}) \right)^2}} \leq t \sqrt{\frac{c}{\lambda_j} \frac{n^{1+2\epsilon} \lambda_j^2}{(b \lambda_j n)^2}} = t \sqrt{\frac{c_1}{\lambda_j n^{1-2\epsilon}}},$$

where we apply  $\tilde{\lambda}_1 r_0(\Sigma_{-\mathcal{S}}) = \sum_{k=k^*+1, k \in \mathcal{S}^c}^d \lambda_k \geq b \lambda_j n$  for  $j > k^*$ . □

### F.3 Spiked and polynomial decay covariance: support outside the top $k^*$ indices

In this section, we show corresponding results to Corollary 27 and Corollary 29 in the case where  $\theta^*$  is supported outside the top  $k^*$  indices of the covariance spectrum. Note that this is necessary to handle scenarios where  $k^* = 0$ , including spiked covariance with  $q > 1 - r$  (Corollary 30) and polynomial decay covariance with  $u \in [0, 1]$ ,  $v = 0$  and  $p \cdot (1 - u) > 1$  (Corollary 31). We will see that the results in this section require  $t \ll n^{\frac{1}{4}-\epsilon}$  and stronger conditions in  $p$  than those in Section F.2.

**Corollary 30.** *Under Assumptions 2 and 5 with  $t \ll n^{\frac{1}{4}-\epsilon}$  and spiked covariance matrix (Definition 3), for  $q < (1 - r)$ , if we have  $\max\{1.5 - q - r, 1\} < p < 2.5 - q - r$ , we can recover support outside of the top  $k^*$  indices of the covariance spectrum such that*

$$|\hat{\theta}_j| \begin{cases} \gg n^{\frac{-p+q+r}{2}} & \text{for } j \in [k^*] \cap \mathcal{S} \\ \gg n^{\frac{-2p+2}{2}} & \text{for } j > k^*, j \in \mathcal{S} \\ \ll n^{\frac{-p+q+r-0.5}{2}} & \text{for } j \in [k^*] \cap \mathcal{S}^c \\ \ll n^{\frac{-2p+1.5}{2}} & \text{for } k^* < j \leq d, j \in \mathcal{S}^c \end{cases}.$$

Therefore, we can pick a threshold between bounds to distinguish support and non-support such as  $T = \min\{n^{\frac{-p+q+r}{2}}, n^{\frac{-2p+2}{2}}\}$ . On the other hand, for  $q > (1 - r)$  and  $\max\{-0.5 + q + r, 1\} < p < 0.5 + q + r$ , we have

$$|\hat{\theta}_j| \begin{cases} \gg n^{\frac{-p+(1-r)-q+1}{2}} & \text{for } j \in [s] \cap \mathcal{S} \\ \gg n^{\frac{-2p+2}{2}} & \text{for } j > s, j \in \mathcal{S} \\ \ll n^{\frac{-p+(1-r)-q+0.5}{2}} & \text{for } j \in [s] \cap \mathcal{S}^c \\ \ll n^{\frac{-2p+1.5}{2}} & \text{for } s < j \leq d, j \in \mathcal{S}^c \end{cases}.$$

Therefore, we can pick a threshold between bounds to distinguish support and non-support such as  $T = \min\{n^{\frac{-p+(1-r)-q+1}{2}}, n^{\frac{-2p+2}{2}}\}$ .

*Proof.* In the first part, for  $q < (1 - r)$ , we already showed the support lower bound for  $j \in [k^*]$  and non-support upper bound for  $j \in [k^*] \cap \mathcal{S}^c$  and  $k^* < j \leq d, j \in \mathcal{S}^c$  in Corollary 27. Note that we apply a different upper bound of  $t$  for non-support upper bound and get different rates. We still need to show the support lower bound for  $j > k^*, j \in \mathcal{S}$ . Hence, for  $j > k^*, j \in \mathcal{S}$ , according to Lemma 26, by substituting  $\lambda_j = \frac{(1-a)d}{d-s} = \frac{(1-n^{-q})n^p}{n^p - n^r}$  and  $\tilde{\lambda}_1 r_0(\Sigma_{-\mathcal{S}}) = (d - |\mathcal{S}|) \frac{(1-a)d}{d-s}$ , we have

$$\begin{aligned} |\hat{\theta}_j| &\geq \sqrt{\frac{2\beta_j}{\lambda_j \pi}} \frac{\lambda_j \left( \frac{n}{c_1 \tilde{\lambda}_1 r_0(\Sigma_{-\mathcal{S}})} \right)}{1 + \lambda_j \left( \frac{c_2 n}{\tilde{\lambda}_1 r_0(\Sigma_{-\mathcal{S}})} \right)} \\ &= \sqrt{\frac{2\beta_j}{\pi}} \sqrt{\frac{d-s}{(1-a)d}} \left( \frac{\frac{(1-a)d}{d-s} \left( \frac{n}{c_1 (d-|\mathcal{S}|) \frac{(1-a)d}{(d-s)}} \right)}{1 + \frac{(1-a)d}{d-s} \left( \frac{c_2 n}{(d-|\mathcal{S}|) \frac{(1-a)d}{(d-s)}} \right)} \right) \\ &= \sqrt{\frac{2\beta_j}{\pi}} \sqrt{\frac{n^p - n^r}{(1-n^{-q})n^p}} \left( \frac{\frac{n}{c_1 (n^p - |\mathcal{S}|)}}{1 + \left( \frac{c_2 n}{(n^p - |\mathcal{S}|)} \right)} \right) \\ &\geq \sqrt{\frac{2\beta_j}{\pi}} c_3 n^{\frac{2-2p}{2}}. \end{aligned}$$

As a result, to recover support outside of top  $k^*$  indices of the covariance spectrum, we need that the lower bound of the support is larger than the upper bound of non-support. Hence, we need the following conditions

$$\begin{aligned} n^{\frac{-2p+2}{2}} &\gg n^{\frac{-p+q+r-0.5}{2}} \Leftrightarrow 2.5 - q - r > p, \\ n^{\frac{-p+q+r}{2}} &\gg n^{\frac{-2p+1.5}{2}} \Leftrightarrow p > 1.5 - q - r. \end{aligned}$$



The proof of the first part is done.

For the second part, we have the support lower bound from Corollary 27 as

$$|\hat{\theta}_j| \geq \sqrt{\frac{2\beta_j}{\pi}} n^{-\frac{p+q+r}{2}} \cdot \frac{\frac{1}{c_8} \cdot n^{(1-r)-q}}{1 + c_9 n^{(1-r)-q}}.$$

Since  $q > (1-r)$ , the numerator part decays to zero as

$$|\hat{\theta}_j| \geq \sqrt{\frac{2\beta_j}{\pi}} c n^{-\frac{p+(1-r)-q+1}{2}}.$$

Moreover, we have  $r_0(\Sigma) > bn$  and  $k^* = 0$ , therefore; for all  $j \in \mathcal{S}^c$ , according to Lemma 26, we have the same non-support upper bound as

$$|\hat{\theta}_j| \leq t \sqrt{\frac{c}{\lambda_j} \frac{n^{1+2\epsilon} \lambda_j^2}{\left(\tilde{\lambda}_1 r_0(\Sigma_{-\bar{s}})\right)^2}}.$$

Next, we substitute different values of  $\lambda_j$  for  $j \in [s]$  and  $j > s$ . We also assume there are  $t_1$  support in the top  $s$  indices of the covariance spectrum, where  $t_1 \leq \min\{s, t\}$ . For  $j \in [s] \cap \mathcal{S}^c$ , we have

$$\begin{aligned} |\hat{\theta}_j| &\leq t \sqrt{\frac{c}{\lambda_j} \frac{n^{1+2\epsilon} \lambda_j^2}{\left(\tilde{\lambda}_1 r_0(\Sigma_{-\bar{s}})\right)^2}} \\ &= t \sqrt{c \frac{s}{ad} \left( \frac{n^{1+2\epsilon} \left(\frac{ad}{s}\right)^2}{\left((s-t_1) \frac{ad}{s} + (d-s-(t-t_1)) \frac{(1-a)d}{d-s}\right)^2} \right)} \\ &= t \sqrt{c n^{-p+q+r} \left( \frac{n^{1+2\epsilon} (n^{p-q-r})^2}{\left((n^r-t_1) n^{p-q-r} + (n^p-n^r-(t-t_1)) \frac{(1-n^{-q})n^p}{n^p-n^r}\right)^2} \right)} \\ &\leq t \sqrt{c_1 n^{-p+q+r} \left( \frac{n^{1+2\epsilon} (n^{p-q-r})^2}{n^{2p}} \right)} \\ &= c_2 t n^{\frac{-p+(1-r)-q+2\epsilon}{2}} \\ &< c_2 t n^{\frac{-p+(1-r)-q+0.5}{2}}, \end{aligned}$$

where we substitute  $t \ll n^{\frac{1}{4}-\epsilon}$ . For  $j > s, j \in \mathcal{S}^c$ , we have the same upper bound as  $q < (1-r)$  case in Corollary 27 such that

$$|\hat{\theta}_j| \leq c_1 t n^{\frac{-2p+1+2\epsilon}{2}} < c_1 n^{\frac{-2p+1.5}{2}},$$

where we apply  $t \ll n^{\frac{1}{4}-\epsilon}$ . To recover support outside of top  $k^*$  indices of the covariance spectrum, we need that the lower bound of the support be larger than the upper bound of non-support. Hence, we need the following conditions

$$\begin{aligned} n^{\frac{-2p+2}{2}} &\gg n^{\frac{-p+(1-r)-q+0.5}{2}} \Leftrightarrow 0.5 + q + r > p, \\ n^{\frac{-p+(1-r)-q+1}{2}} &\gg n^{\frac{-2p+1.5}{2}} \Leftrightarrow p > -0.5 + q + r. \end{aligned}$$

The proof is done. □

**Corollary 31.** Under Assumptions 2 and 5 with  $t \ll n^{\frac{1}{4}-\epsilon}$  and polynomial decay covariance in Definition 4 with  $u \in [0, 1)$  and  $v = 0$ , for  $p \cdot (1 - u) > 1$ , by substituting  $\lambda_j = j^{-u}$ , we have

$$|\hat{\theta}_j| \begin{cases} \gg n \cdot d^{\frac{u-2}{2}} & \text{for } j \in \mathcal{S} \\ \ll n^{\frac{1.5}{2}} \cdot d^{\frac{2u-2}{2}} & \text{for } j \in \mathcal{S}^c \end{cases}.$$

Therefore, if  $p \cdot (1 - u) > \max\{1, p - \frac{1}{2}\}$ , we can pick  $T = n \cdot d^{\frac{u-2}{2}}$  as threshold to distinguish support and non-support. Note that when  $u = 0$ ,  $\Sigma$  degenerates to isotropic covariance.

*Proof.* First, we already showed that polynomial decay covariance (Definition 4) with  $u \in [0, 1)$ ,  $v = 0$  and  $p \cdot (1 - u) > 1$  implies  $k^* = 0$  in Corollary 18. Next, we show the lower bound for support indices  $j \in \mathcal{S}$  and upper bound for non-support indices  $j \in \mathcal{S}^c$ . Based on Lemma 26, we have

$$\begin{aligned} |\hat{\theta}_j| &\geq \sqrt{\frac{2\beta_j}{\lambda_j \pi}} \frac{\lambda_j \left( \frac{n}{c_1 \tilde{\lambda}_1 r_0(\Sigma_{-S})} \right)}{1 + \lambda_j \left( \frac{c_2 n}{\tilde{\lambda}_1 r_0(\Sigma_{-S})} \right)}, & \text{for } j \in \mathcal{S} \\ |\hat{\theta}_j| &\leq t \sqrt{\frac{c}{\lambda_j} \frac{n^{1+2\epsilon} \lambda_j^2}{\left( \tilde{\lambda}_1 r_0(\Sigma_{-S}) \right)^2}}, & \text{for } j \in \mathcal{S}^c, \end{aligned}$$

where  $\Sigma_{-S}$  has  $k^* = 0$  still by Lemma 34. Therefore, for  $j \in \mathcal{S}$ , we have

$$|\hat{\theta}_j| \geq \sqrt{\frac{2\beta_j}{\lambda_j \pi}} \frac{\lambda_j \left( \frac{n}{c_1 \tilde{\lambda}_1 r_0(\Sigma_{-S})} \right)}{1 + \lambda_j \left( \frac{c_2 n}{\tilde{\lambda}_1 r_0(\Sigma_{-S})} \right)} = \sqrt{\frac{2\beta_j}{\lambda_j \pi}} \frac{\lambda_j n}{c_1 \tilde{\lambda}_1 r_0(\Sigma_{-S}) + c_1 c_2 \lambda_j n} = \sqrt{\frac{2\beta_j}{\pi}} \frac{\sqrt{\lambda_j} n}{c_1 \sum_{k=1, k \in \mathcal{S}^c}^d \lambda_k + c_1 c_2 \lambda_j n}.$$

Since  $\sum_{k=1, k \in \mathcal{S}^c}^d \lambda_k \leq \sum_{k=1}^d \lambda_k$  and  $\lambda_j n \leq \sum_{k=1}^d \lambda_k$ , we can upper bound the denominator by  $\sum_{k=1}^d \lambda_k$  and get

$$\sqrt{\frac{2\beta_j}{\pi}} \frac{\sqrt{\lambda_j} n}{c_1 \sum_{k=1, k \in \mathcal{S}^c}^d \lambda_k + c_1 c_2 \lambda_j n} \geq \sqrt{\frac{2\beta_j}{\pi}} \frac{\sqrt{\lambda_j} n}{c_3 \sum_{k=1}^d \lambda_k} \geq \sqrt{\frac{2\beta_j}{\pi}} \frac{\sqrt{\lambda_d} n}{c_3 d^{1-u}} = c_4 n \cdot d^{\frac{u-2}{2}},$$

where we apply equation (18a) in the last inequality. For non-support,  $j \in \mathcal{S}^c$ , we have

$$|\hat{\theta}_j| \leq t \sqrt{\frac{c}{\lambda_j} \frac{n^{1+2\epsilon} \lambda_j^2}{\left( \tilde{\lambda}_1 r_0(\Sigma_{-S}) \right)^2}} = t \sqrt{\frac{c}{\left( \sum_{k=1, k \in \mathcal{S}^c}^d \lambda_k \right)^2} \frac{n^{1+2\epsilon} \lambda_j}{\lambda_j}} \leq t \sqrt{\frac{c}{((d-t)\lambda_d)^2} \frac{n^{1+2\epsilon} \lambda_1}{\lambda_1}} = t \sqrt{\frac{c \cdot n^{1+2\epsilon} d^{2u}}{(d-t)^2}} < c_1 n^{\frac{1.5}{2}} d^{u-1},$$

where we apply  $\lambda_1 \geq \lambda_j$  for all  $j \in [d]$  and  $\sum_{k=1, k \in \mathcal{S}^c}^d \lambda_k \geq (d-t)\lambda_d$ . In the last inequality, we apply  $t \ll n^{\frac{1}{4}-\epsilon}$ . As a result, we need

$$n \cdot d^{\frac{u-2}{2}} \gg n^{\frac{1.5}{2}} d^{u-1} \Leftrightarrow n^{\frac{1}{2}} \gg d^u \Leftrightarrow n^{\frac{1}{2}} \gg n^{p \cdot u} \Leftrightarrow \frac{1}{2} > p \cdot u$$

to ensure the lower bound for support indices  $j \in \mathcal{S}$  is larger than the upper bound for non-support indices  $j \in \mathcal{S}^c$ . Ultimately, we can combine the condition  $p \cdot (1 - u) > 1$  and  $\frac{1}{2} > p \cdot u$  into  $p \cdot (1 - u) > \max\{1, p - \frac{1}{2}\}$ . The proof is done.  $\square$

## G Survival and contamination auxiliary lemmas

In this section, we provide proofs of Lemmas 15, 16, and 17 from Appendix B. While these are essentially extensions of results of Wang et al. (2023), they are not entirely straightforward due to our analysis of  $t$ -sparsity. For example, in Lemma 15, Wang et al. (2023) analyze  $\mathbb{E}[\text{sign}(\mathbf{z}) \mathbf{z}^\top] = \sqrt{\frac{2}{\pi}} \mathbf{I}$  where  $\mathbf{z}$  is an

independent standard Gaussian vector representing the regression labels. On the other hand, given a support set  $\mathcal{S}$ , we must analyze the more complicated term  $\mathbb{E} \left[ \text{sign} \left( \sum_{j \in \mathcal{S}} a_j \mathbf{z}_j \right) \mathbf{z}_{s_1}^\top \right]$ ; this leads to an interesting quantification of the relative contribution of each index of  $\boldsymbol{\theta}^*$ , which we denote by  $\beta$ .

*Proof.* (Lemma 15)

According to the definition of  $Q_{s_t}$  in Equation (2), we have

$$Q_{s_t} = \mathbf{z}_{s_1}^\top \mathbf{A}_{-s_1:s_t}^{-1} \hat{\mathbf{y}} = \mathbf{z}_{s_1}^\top \mathbf{A}_{-s_1:s_t}^{-1} \text{sign} \left( \sum_{j \in \mathcal{S}} a_j \mathbf{z}_j \right)$$

Since  $\mathbf{z}_{s_1}$  and  $\text{sign} \left( \sum_{j \in \mathcal{S}} a_j \mathbf{z}_j \right)$  are independent to  $\mathbf{A}_{-s_1:s_t}^{-1}$ , by applying the parallelogram rule and the Hanson-Wright inequality in Lemma 32, we get

$$\begin{aligned} Q_{s_t} &\leq \mathbb{E} [\mathbf{z}_{s_1}^\top \mathbf{A}_{-s_1:s_t}^{-1} \hat{\mathbf{y}}] + 2c_1 \|\mathbf{A}_{-s_1:s_t}^{-1}\|_2 \cdot n^{\frac{1}{2}+\epsilon}, \\ Q_{s_t} &\geq \mathbb{E} [\mathbf{z}_{s_1}^\top \mathbf{A}_{-s_1:s_t}^{-1} \hat{\mathbf{y}}] - 2c_1 \|\mathbf{A}_{-s_1:s_t}^{-1}\|_2 \cdot n^{\frac{1}{2}+\epsilon}, \end{aligned}$$

with probability at least  $1 - 2e^{-n^{2\epsilon}}$ . Next, we calculate the value of  $\mathbb{E} [\mathbf{z}_{s_1}^\top \mathbf{A}_{-s_1:s_t}^{-1} \hat{\mathbf{y}}]$ .

$$\begin{aligned} \mathbb{E} [\mathbf{z}_{s_1}^\top \mathbf{A}_{-s_1:s_t}^{-1} \hat{\mathbf{y}}] &= \text{Tr} (\mathbf{A}_{-s_1:s_t}^{-1} \mathbb{E} [\hat{\mathbf{y}} \mathbf{z}_{s_1}^\top]) \\ &= \text{Tr} \left( \mathbf{A}_{-s_1:s_t}^{-1} \mathbb{E} \left[ \text{sign} \left( \sum_{j \in \mathcal{S}} a_j z_{j,1} \right) z_{s_1,1} \right] \cdot \mathbf{I} \right) \\ &= \mathbb{E} \left[ \text{sign} \left( \sum_{j \in \mathcal{S}} a_j z_{j,1} \right) z_{s_1,1} \right] \text{Tr} (\mathbf{A}_{-s_1:s_t}^{-1}). \end{aligned}$$

We now derive the value of  $\mathbb{E} \left[ \text{sign} \left( \sum_{j \in \mathcal{S}} a_j z_{j,1} \right) z_{s_1,1} \right]$ . We denote  $p := z_{s_1,1} \sim \mathcal{N}(0, 1)$  and  $q := \sum_{j \in \mathcal{S}, j \neq s_1} a_j z_{j,1} \sim \mathcal{N}(0, \sigma_q^2)$  where  $\sigma_q^2 := \sum_{j \in \mathcal{S}, j \neq s_1} a_j^2$ . We then have

$$\begin{aligned} \mathbb{E} \left[ \text{sign} \left( \sum_{j \in \mathcal{S}} a_j z_{j,1} \right) z_{s_1,1} \right] &= \mathbb{E} \left[ \text{sign} \left( a_{s_1} z_{s_1,1} + \sum_{j \in \mathcal{S}, j \neq s_1} a_j z_{j,1} \right) z_{s_1,1} \right] \\ &= \mathbb{E}_p [\mathbb{E}_q [\text{sign} (a_{s_1} p + q)] \cdot p] \\ &= \mathbb{E}_p [(1 - \Phi_q (-a_{s_1} p) - \Phi_q (-a_{s_1} p)) \cdot p] \\ &= 2\mathbb{E}_p [\Phi_q (a_{s_1} p) \cdot p], \end{aligned}$$

where  $\Phi_q$  is the cdf of  $q$  and we use the property that  $1 - \Phi_q(-a_{s_1}p) = \Phi_q(a_{s_1}p)$ . Then, we have

$$\begin{aligned}
\mathbb{E}_p[\Phi_q(a_{s_1}p) \cdot p] &= \mathbb{E}_p[\mathbb{E}_q[\mathbf{1}_{q \leq a_{s_1}p} | p] \cdot p] \\
&= \mathbb{E}_{p,q}[\mathbf{1}_{q \leq a_{s_1}p} \cdot p] \\
&= \text{sign}(a_{s_1}) \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi\sigma_q^2}} e^{\frac{-q^2}{2\sigma_q^2}} \left( \int_{\frac{q}{|a_{s_1}|}}^{+\infty} \frac{1}{\sqrt{2\pi}} p e^{\frac{-p^2}{2}} dp \right) dq \\
&= \text{sign}(a_{s_1}) \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi\sigma_q^2}} e^{\frac{-q^2}{2\sigma_q^2}} \left( \frac{1}{\sqrt{2\pi}} e^{\frac{-q^2}{2a_{s_1}^2}} \right) dq \\
&= \frac{\text{sign}(a_{s_1})}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi\sigma_q^2}} e^{\frac{-(\sum_{j \in \mathcal{S}} a_j^2)q^2}{2\sigma_q^2 a_{s_1}^2}} dq \\
&= \frac{a_{s_1}}{\sqrt{2\pi \left( \sum_{j \in \mathcal{S}} a_j^2 \right)}} \\
&= \text{sign}(a_{s_1}) \sqrt{\frac{\beta_{s_1}}{2\pi}},
\end{aligned}$$

where the second equality uses the law of total expectation, and we substitute  $\beta_{s_\ell} := \frac{\lambda_{s_\ell} \theta_{s_\ell}^{*2}}{\sum_{j \in \mathcal{S}} \lambda_j \theta_j^{*2}} = \frac{a_{s_\ell}^2}{\sum_{j \in \mathcal{S}} a_j^2}$  in the last equality. The proof is done by substituting the expectation values in the Hanson-Wright inequalities.  $\square$

*Proof.* (Lemma 16)

Recall the definition of  $Q_{s_\ell}$  in Equation (2) such that  $Q_{s_\ell} := \mathbf{z}_{s_1}^\top \mathbf{A}_{-s_1:s_\ell}^{-1} \hat{\mathbf{y}}$  for  $1 \leq \ell \leq t$ . Therefore, we can write  $Q_{s_1}$  as

$$\begin{aligned}
Q_{s_1} &= \mathbf{z}_{s_1}^\top \mathbf{A}_{-s_1:s_1}^{-1} \hat{\mathbf{y}} \\
&= \mathbf{z}_{s_1}^\top \left( \mathbf{A}_{-s_1:s_2} + \lambda_{s_2} \mathbf{z}_{s_2} \mathbf{z}_{s_2}^\top \right)^{-1} \hat{\mathbf{y}} \\
&= \mathbf{z}_{s_1}^\top \left( \mathbf{A}_{-s_1:s_2}^{-1} - \frac{\lambda_{s_2} \mathbf{A}_{-s_1:s_2}^{-1} \mathbf{z}_{s_2} \mathbf{z}_{s_2}^\top \mathbf{A}_{-s_1:s_2}^{-1}}{1 + \lambda_{s_2} \mathbf{z}_{s_2}^\top \mathbf{A}_{-s_1:s_2}^{-1} \mathbf{z}_{s_2}} \right) \hat{\mathbf{y}} \\
&= \mathbf{z}_{s_1}^\top \left( \mathbf{A}_{-s_1:s_t}^{-1} - \sum_{\ell=2}^t \frac{\lambda_{s_\ell} \mathbf{A}_{-s_1:s_\ell}^{-1} \mathbf{z}_{s_\ell} \mathbf{z}_{s_\ell}^\top \mathbf{A}_{-s_1:s_\ell}^{-1}}{1 + \lambda_{s_\ell} \mathbf{z}_{s_\ell}^\top \mathbf{A}_{-s_1:s_\ell}^{-1} \mathbf{z}_{s_\ell}} \right) \hat{\mathbf{y}} \\
&= Q_{s_t} - \sum_{\ell=2}^t \frac{\lambda_{s_\ell} \mathbf{z}_{s_1}^\top \mathbf{A}_{-s_1:s_\ell}^{-1} \mathbf{z}_{s_\ell} \mathbf{z}_{s_\ell}^\top \mathbf{A}_{-s_1:s_\ell}^{-1} \hat{\mathbf{y}}}{1 + \lambda_{s_\ell} \mathbf{z}_{s_\ell}^\top \mathbf{A}_{-s_1:s_\ell}^{-1} \mathbf{z}_{s_\ell}} \\
&= Q_{s_t} \left( 1 - \sum_{\ell=2}^t \frac{\lambda_{s_\ell} \mathbf{z}_{s_1}^\top \mathbf{A}_{-s_1:s_\ell}^{-1} \mathbf{z}_{s_\ell} \mathbf{z}_{s_\ell}^\top \mathbf{A}_{-s_1:s_\ell}^{-1} \hat{\mathbf{y}}}{Q_{s_t} (1 + \lambda_{s_\ell} \mathbf{z}_{s_\ell}^\top \mathbf{A}_{-s_1:s_\ell}^{-1} \mathbf{z}_{s_\ell})} \right) \\
&= Q_{s_t} \left( 1 - \text{sign}(a_{s_1}) \sum_{\ell=2}^t \frac{\lambda_{s_\ell} \mathbf{z}_{s_1}^\top \mathbf{A}_{-s_1:s_\ell}^{-1} \mathbf{z}_{s_\ell} \mathbf{z}_{s_\ell}^\top \mathbf{A}_{-s_1:s_\ell}^{-1} \hat{\mathbf{y}}}{|Q_{s_t}| (1 + \lambda_{s_\ell} \mathbf{z}_{s_\ell}^\top \mathbf{A}_{-s_1:s_\ell}^{-1} \mathbf{z}_{s_\ell})} \right)
\end{aligned}$$

where we apply the Sherman-Morrison-Woodbury identity recursively over  $\ell = 2, \dots, t$ . In the last equality, we use that fact that the sign of  $Q_{s_t}$  is controlled by  $a_{s_1}$  for large enough  $n$  from Lemma 15. According to

the sign of  $Q_{s_t}$ , we have the upper and lower bound of  $Q_{s_1}$  by

$$\begin{aligned} Q_{s_1} &\leq Q_{s_t} \left( 1 + \text{sign}(a_{s_1}) \left| \sum_{\ell=2}^t \frac{\lambda_{s_\ell} \mathbf{z}_{s_1}^\top \mathbf{A}_{-s_1:s_\ell}^{-1} \mathbf{z}_{s_\ell} \mathbf{z}_{s_\ell}^\top \mathbf{A}_{-s_1:s_\ell}^{-1} \hat{\mathbf{y}}}{|Q_{s_t}| (1 + \lambda_{s_\ell} \mathbf{z}_{s_\ell}^\top \mathbf{A}_{-s_1:s_\ell}^{-1} \mathbf{z}_{s_\ell})} \right| \right) \\ Q_{s_1} &\geq Q_{s_t} \left( 1 - \text{sign}(a_{s_1}) \left| \sum_{\ell=2}^t \frac{\lambda_{s_\ell} \mathbf{z}_{s_1}^\top \mathbf{A}_{-s_1:s_\ell}^{-1} \mathbf{z}_{s_\ell} \mathbf{z}_{s_\ell}^\top \mathbf{A}_{-s_1:s_\ell}^{-1} \hat{\mathbf{y}}}{|Q_{s_t}| (1 + \lambda_{s_\ell} \mathbf{z}_{s_\ell}^\top \mathbf{A}_{-s_1:s_\ell}^{-1} \mathbf{z}_{s_\ell})} \right| \right). \end{aligned}$$

Hence, it remains to bound the absolute value term above. We then have

$$\begin{aligned} \left| \sum_{\ell=2}^t \frac{\lambda_{s_\ell} \mathbf{z}_{s_1}^\top \mathbf{A}_{-s_1:s_\ell}^{-1} \mathbf{z}_{s_\ell} \mathbf{z}_{s_\ell}^\top \mathbf{A}_{-s_1:s_\ell}^{-1} \hat{\mathbf{y}}}{|Q_{s_t}| (1 + \lambda_{s_\ell} \mathbf{z}_{s_\ell}^\top \mathbf{A}_{-s_1:s_\ell}^{-1} \mathbf{z}_{s_\ell})} \right| &\leq \sum_{\ell=2}^t \frac{\lambda_{s_\ell} \|\mathbf{z}_{s_1}^\top \mathbf{A}_{-s_1:s_\ell}^{-1} \mathbf{z}_{s_\ell}\| \|\mathbf{z}_{s_\ell}^\top \mathbf{A}_{-s_1:s_\ell}^{-1} \hat{\mathbf{y}}\|}{|Q_{s_t}| (1 + \lambda_{s_\ell} \mathbf{z}_{s_\ell}^\top \mathbf{A}_{-s_1:s_\ell}^{-1} \mathbf{z}_{s_\ell})} \\ &\leq \sum_{\ell=2}^t \frac{\|\mathbf{z}_{s_1}^\top \mathbf{A}_{-s_1:s_\ell}^{-1} \mathbf{z}_{s_\ell}\| \|\mathbf{z}_{s_\ell}^\top \mathbf{A}_{-s_1:s_\ell}^{-1} \hat{\mathbf{y}}\|}{|Q_{s_t}| (\mathbf{z}_{s_\ell}^\top \mathbf{A}_{-s_1:s_\ell}^{-1} \mathbf{z}_{s_\ell})} \\ &= \sum_{\ell=2}^t \underbrace{\frac{\|\mathbf{z}_{s_1}^\top \mathbf{A}_{-s_1:s_\ell}^{-1} \mathbf{z}_{s_\ell}\|}{|Q_{s_t}|}}_{:=T_1} \underbrace{\frac{\|\mathbf{z}_{s_\ell}^\top \mathbf{A}_{-s_1:s_\ell}^{-1} \hat{\mathbf{y}}\|}{\mathbf{z}_{s_\ell}^\top \mathbf{A}_{-s_1:s_\ell}^{-1} \mathbf{z}_{s_\ell}}}_{:=T_2}, \end{aligned}$$

where in the second inequality we deduct 1 in the denominator. Next, we aim to upper bound  $T_1$  and  $T_2$  respectively. According to the bounds of  $Q_{s_t}$  in Lemma 15, we can have the bounds of  $|Q_{s_t}|$  as

$$\sqrt{\frac{2\beta_{s_1}}{\pi}} \text{Tr}(\mathbf{A}_{-s_1:s_t}^{-1}) - 2c_1 \|\mathbf{A}_{-s_1:s_t}^{-1}\|_2 \cdot n^{\frac{1}{2}+\epsilon} \leq |Q_{s_t}| \leq \sqrt{\frac{2\beta_{s_1}}{\pi}} \text{Tr}(\mathbf{A}_{-s_1:s_t}^{-1}) + 2c_1 \|\mathbf{A}_{-s_1:s_t}^{-1}\|_2 \cdot n^{\frac{1}{2}+\epsilon}. \quad (45)$$

For  $T_1$ , we can apply Hanson-Wright inequality (Lemma 32) in the numerator and apply the lower bound of  $|Q_{s_t}|$  in Equation (45) in the denominator, and we get

$$\begin{aligned} T_1 &\leq \frac{2c_1 \|\mathbf{A}_{-s_1:s_\ell}^{-1}\|_2 \cdot n^{\frac{1}{2}+\epsilon}}{\sqrt{\frac{2\beta_{s_1}}{\pi}} \text{Tr}(\mathbf{A}_{-s_1:s_t}^{-1}) - 2c_1 \|\mathbf{A}_{-s_1:s_t}^{-1}\|_2 \cdot n^{\frac{1}{2}+\epsilon}} \\ &\leq \frac{2c_1 \|\mathbf{A}_{-(s_1:s_t) \cup [k^*]}^{-1}\|_2 \cdot n^{\frac{1}{2}+\epsilon}}{\sqrt{\frac{2\beta_{s_1}}{\pi}} \text{Tr}(\mathbf{A}_{-s_1:s_t}^{-1}) - 2c_1 \|\mathbf{A}_{-(s_1:s_t) \cup [k^*]}^{-1}\|_2 \cdot n^{\frac{1}{2}+\epsilon}} \\ &\leq \frac{2c_1 \|\mathbf{A}_{-(s_1:s_t) \cup [k^*]}^{-1}\|_2 \cdot n^{\frac{1}{2}+\epsilon}}{\sqrt{\frac{2\beta_{s_1}}{\pi}} (1 - \frac{c}{n})^{k^*} \text{Tr}(\mathbf{A}_{-(s_1:s_t) \cup [k^*]}^{-1}) - 2c_1 \|\mathbf{A}_{-(s_1:s_t) \cup [k^*]}^{-1}\|_2 \cdot n^{\frac{1}{2}+\epsilon}} \\ &= \frac{1}{\sqrt{\frac{2\beta_{s_1}}{\pi}} (1 - \frac{c}{n})^{k^*} \frac{\text{Tr}(\mathbf{A}_{-(s_1:s_t) \cup [k^*]}^{-1})}{2c_1 \|\mathbf{A}_{-(s_1:s_t) \cup [k^*]}^{-1}\|_2} - 1}, \end{aligned} \quad (46)$$

where the second inequality follows by  $\mathbf{A}_{-s_1:s_\ell}^{-1} \preceq \mathbf{A}_{-s_1:s_t}^{-1} \preceq \mathbf{A}_{-(s_1:s_t) \cup [k^*]}^{-1}$  for all  $\ell \leq t$ , and the third inequality follows by the trace lower bound in Lemma 36. Next, Lemma 35 ensures eigenvalues of  $\mathbf{A}_{-(s_1:s_t) \cup [k^*]}^{-1}$  are identical up to a constant such that

$$\frac{\text{Tr}(\mathbf{A}_{-(s_1:s_t) \cup [k^*]}^{-1})}{\|\mathbf{A}_{-(s_1:s_t) \cup [k^*]}^{-1}\|_2} \geq \frac{n}{c}. \quad (47)$$

Substitute Equation (47) into Equation (46), we get

$$T_1 \leq \frac{1}{\sqrt{\frac{2\beta_{s_1}}{\pi}} (1 - \frac{c}{n})^{k^*} \frac{n^{\frac{1}{2}-\epsilon}}{c_2} - 1} = \frac{1}{c_3 n^{\frac{1}{2}-\epsilon} - 1}.$$

For  $T_2$ , we use the sub-multiplicative matrix norm and get

$$\begin{aligned}
T_2 &\leq \frac{\|\mathbf{z}_{s_\ell}\|_2 \|\hat{\mathbf{y}}\|_2 \|\mathbf{A}_{-s_1:s_\ell}^{-1}\|_2}{\mathbf{z}_{s_\ell}^\top \mathbf{A}_{-s_1:s_\ell}^{-1} \mathbf{z}_{s_\ell}} \\
&\leq \frac{\|\mathbf{z}_{s_\ell}\|_2 \|\hat{\mathbf{y}}\|_2 \|\mathbf{A}_{-s_1:s_\ell}^{-1}\|_2}{\text{Tr}(\mathbf{A}_{-s_1:s_\ell}^{-1}) - c_1 \|\mathbf{A}_{-s_1:s_\ell}^{-1}\|_2 \cdot n^{\frac{1}{2}+\epsilon}} \\
&\leq \frac{\|\mathbf{z}_{s_\ell}\|_2 \|\hat{\mathbf{y}}\|_2 \|\mathbf{A}_{-s_1:s_\ell}^{-1}\|_2}{\left(1 - \frac{c}{n}\right)^{k^*+t-\ell} \text{Tr}(\mathbf{A}_{-(s_1:s_t)\cup[k^*]}^{-1}) - c_1 \|\mathbf{A}_{-s_1:s_\ell}^{-1}\|_2 \cdot n^{\frac{1}{2}+\epsilon}} \\
&\leq \frac{\|\mathbf{z}_{s_\ell}\|_2 \|\hat{\mathbf{y}}\|_2 \|\mathbf{A}_{-(s_1:s_t)\cup[k^*]}^{-1}\|_2}{\left(1 - \frac{c}{n}\right)^{k^*+t-\ell} \text{Tr}(\mathbf{A}_{-(s_1:s_t)\cup[k^*]}^{-1}) - c_1 \|\mathbf{A}_{-(s_1:s_t)\cup[k^*]}^{-1}\|_2 \cdot n^{\frac{1}{2}+\epsilon}}, \tag{48}
\end{aligned}$$

where the second inequality follows the Hanson-Wright inequality in Lemma 32. In the third inequality, we use the trace lower bound in Lemma 36. The last inequality follows by  $\mathbf{A}_{-(s_1:s_t)\cup[k^*]}^{-1} \succeq \mathbf{A}_{-s_1:s_\ell}^{-1}$  for  $\ell \leq t$ . Next, we apply the tightness of eigenvalues of  $\mathbf{A}_{-(s_1:s_t)\cup[k^*]}^{-1}$  in Equation (47) again and get

$$T_2 \leq \frac{\|\mathbf{z}_{s_\ell}\|_2 \|\hat{\mathbf{y}}\|_2}{\left(1 - \frac{c}{n}\right)^{k^*+t-\ell} \frac{n}{c} - c_1 n^{\frac{1}{2}+\epsilon}} \leq \frac{cn}{\left(1 - \frac{c}{n}\right)^{k^*+t-\ell} \frac{n}{c} - c_1 n^{\frac{1}{2}+\epsilon}} \leq c_4,$$

where the last inequality follows Lemma 33 such that  $\|\mathbf{z}_{s_\ell}\|_2 \leq c\sqrt{n}$ . Put together the upper bound of  $T_1$  and  $T_2$ , the proof is complete.  $\square$

*Proof.* (Lemma 17)

We show that  $\left| \frac{\lambda_{s_\ell} \mathbf{z}_{s_\ell}^\top \mathbf{A}_{-s_1:s_\ell}^{-1} \check{\mathbf{y}}_{s_\ell-1}}{1 + \lambda_{s_\ell} \mathbf{z}_{s_\ell}^\top \mathbf{A}_{-s_1:s_\ell}^{-1} \mathbf{z}_{s_\ell}} \right| \leq c$  for  $1 \leq \ell \leq t$  by induction. Recall the definition of  $\check{\mathbf{y}}$  as  $\check{\mathbf{y}}_{s_\ell} := \check{\mathbf{y}}_{s_\ell-1} - \frac{\lambda_{s_\ell} \mathbf{z}_{s_\ell}^\top \mathbf{A}_{-s_1:s_\ell}^{-1} \check{\mathbf{y}}_{s_\ell-1}}{1 + \lambda_{s_\ell} \mathbf{z}_{s_\ell}^\top \mathbf{A}_{-s_1:s_\ell}^{-1} \mathbf{z}_{s_\ell}} \mathbf{z}_{s_\ell}$  and  $\check{\mathbf{y}}_{s_0} = \hat{\mathbf{y}}$  for  $1 \leq \ell \leq t$ . For the base case  $\ell = 1$ , we have

$$\frac{|\lambda_{s_1} \mathbf{z}_{s_1}^\top \mathbf{A}_{-s_1:s_1}^{-1} \hat{\mathbf{y}}|}{1 + \lambda_{s_1} \mathbf{z}_{s_1}^\top \mathbf{A}_{-s_1:s_1}^{-1} \mathbf{z}_{s_1}} \leq \frac{|\mathbf{z}_{s_1}^\top \mathbf{A}_{-s_1:s_1}^{-1} \hat{\mathbf{y}}|}{\mathbf{z}_{s_1}^\top \mathbf{A}_{-s_1:s_1}^{-1} \mathbf{z}_{s_1}} \tag{49}$$

$$\begin{aligned}
&\leq \frac{\|\mathbf{z}_{s_1}\|_2 \|\mathbf{A}_{-s_1:s_1}^{-1}\|_2 \|\hat{\mathbf{y}}\|_2}{\text{Tr}(\mathbf{A}_{-s_1:s_1}^{-1}) - c_1 \|\mathbf{A}_{-s_1:s_1}^{-1}\|_2 \cdot n^{\frac{1}{2}+\epsilon}} \\
&\leq \frac{\|\mathbf{z}_{s_1}\|_2 \|\mathbf{A}_{-(s_1:s_t)\cup[k^*]}^{-1}\|_2 \|\hat{\mathbf{y}}\|_2}{\text{Tr}(\mathbf{A}_{-s_1:s_1}^{-1}) - c_1 \|\mathbf{A}_{-(s_1:s_t)\cup[k^*]}^{-1}\|_2 \cdot n^{\frac{1}{2}+\epsilon}}, \tag{50}
\end{aligned}$$

where in the second inequality, we use the sub-multiplicative of matrix norm in the numerator and Hanson-Wright inequality (Lemma 32) in the denominator. In the last inequality, we use the fact that  $\mathbf{A}_{-(s_1:s_t)\cup[k^*]}^{-1} \succeq \mathbf{A}_{-s_1:s_1}^{-1}$ . Next, we apply Lemma 36 and get  $\text{Tr}(\mathbf{A}_{-s_1:s_1}^{-1}) \geq \left(1 - \frac{c}{n}\right)^{k^*+t-1} \text{Tr}(\mathbf{A}_{-(s_1:s_t)\cup[k^*]}^{-1})$ . Therefore,

$$\begin{aligned}
(50) &\leq \frac{\|\mathbf{z}_{s_1}\|_2 \|\mathbf{A}_{-(s_1:s_t)\cup[k^*]}^{-1}\|_2 \|\hat{\mathbf{y}}\|_2}{\left(1 - \frac{c}{n}\right)^{k^*+t-1} \text{Tr}(\mathbf{A}_{-(s_1:s_t)\cup[k^*]}^{-1}) - c_1 \|\mathbf{A}_{-(s_1:s_t)\cup[k^*]}^{-1}\|_2 \cdot n^{\frac{1}{2}+\epsilon}} \\
&= \frac{\|\mathbf{z}_{s_1}\|_2 \|\hat{\mathbf{y}}\|_2}{\left(1 - \frac{c}{n}\right)^{k^*+t-1} \frac{\text{Tr}(\mathbf{A}_{-(s_1:s_t)\cup[k^*]}^{-1})}{\|\mathbf{A}_{-(s_1:s_t)\cup[k^*]}^{-1}\|_2} - c_1 n^{\frac{1}{2}+\epsilon}} \\
&\leq \frac{cn}{\left(1 - \frac{c}{n}\right)^{k^*+t-1} \frac{n}{c} - c_1 n^{\frac{1}{2}+\epsilon}} \\
&\leq c,
\end{aligned}$$

where in the second inequality we apply Lemma 33 to get  $\|\mathbf{z}_{s_1}\|_2 \leq c\sqrt{n}$  and we apply Lemma 35 to show eigenvalues of  $\mathbf{A}_{-(s_1:s_\ell) \cup [k^*]}^{-1}$  are identical up to a constant. The base case  $\ell = 1$  is proved. Next, we assume  $\left| \frac{\lambda_{s_j} \mathbf{z}_{s_j}^\top \mathbf{A}_{-s_1:s_j}^{-1} \tilde{\mathbf{y}}_{s_{j-1}}}{1 + \lambda_{s_j} \mathbf{z}_{s_j}^\top \mathbf{A}_{-s_1:s_j}^{-1} \mathbf{z}_{s_j}} \right| \leq c$  is true for  $1 \leq j \leq \ell - 1$ , and we show for  $j = \ell$  the statement holds. We have the  $j = \ell$  case

$$\begin{aligned}
\left| \frac{\lambda_{s_\ell} \mathbf{z}_{s_\ell}^\top \mathbf{A}_{-s_1:s_\ell}^{-1} \tilde{\mathbf{y}}_{s_{\ell-1}}}{1 + \lambda_{s_\ell} \mathbf{z}_{s_\ell}^\top \mathbf{A}_{-s_1:s_\ell}^{-1} \mathbf{z}_{s_\ell}} \right| &\leq \frac{|\mathbf{z}_{s_\ell}^\top \mathbf{A}_{-s_1:s_\ell}^{-1} \tilde{\mathbf{y}}_{s_{\ell-1}}|}{\mathbf{z}_{s_\ell}^\top \mathbf{A}_{-s_1:s_\ell}^{-1} \mathbf{z}_{s_\ell}} \\
&= \frac{\left| \mathbf{z}_{s_\ell}^\top \mathbf{A}_{-s_1:s_\ell}^{-1} \left( \hat{\mathbf{y}} - \sum_{j=1}^{\ell-1} \frac{\lambda_{s_j} \mathbf{z}_{s_j}^\top \mathbf{A}_{-s_1:s_j}^{-1} \tilde{\mathbf{y}}_{s_{j-1}}}{1 + \lambda_{s_j} \mathbf{z}_{s_j}^\top \mathbf{A}_{-s_1:s_j}^{-1} \mathbf{z}_{s_j}} \mathbf{z}_{s_j} \right) \right|}{\mathbf{z}_{s_\ell}^\top \mathbf{A}_{-s_1:s_\ell}^{-1} \mathbf{z}_{s_\ell}} \\
&\leq \frac{|\mathbf{z}_{s_\ell}^\top \mathbf{A}_{-s_1:s_\ell}^{-1} \hat{\mathbf{y}}|}{\mathbf{z}_{s_\ell}^\top \mathbf{A}_{-s_1:s_\ell}^{-1} \mathbf{z}_{s_\ell}} + \sum_{j=1}^{\ell-1} \frac{\left| \frac{\lambda_{s_j} \mathbf{z}_{s_j}^\top \mathbf{A}_{-s_1:s_j}^{-1} \tilde{\mathbf{y}}_{s_{j-1}}}{1 + \lambda_{s_j} \mathbf{z}_{s_j}^\top \mathbf{A}_{-s_1:s_j}^{-1} \mathbf{z}_{s_j}} \mathbf{z}_{s_\ell}^\top \mathbf{A}_{-s_1:s_\ell}^{-1} \mathbf{z}_{s_j} \right|}{\mathbf{z}_{s_\ell}^\top \mathbf{A}_{-s_1:s_\ell}^{-1} \mathbf{z}_{s_\ell}} \\
&\leq \frac{|\mathbf{z}_{s_\ell}^\top \mathbf{A}_{-s_1:s_\ell}^{-1} \hat{\mathbf{y}}|}{\mathbf{z}_{s_\ell}^\top \mathbf{A}_{-s_1:s_\ell}^{-1} \mathbf{z}_{s_\ell}} + \sum_{j=1}^{\ell-1} \frac{c |\mathbf{z}_{s_\ell}^\top \mathbf{A}_{-s_1:s_\ell}^{-1} \mathbf{z}_{s_j}|}{\mathbf{z}_{s_\ell}^\top \mathbf{A}_{-s_1:s_\ell}^{-1} \mathbf{z}_{s_\ell}},
\end{aligned}$$

where we upper bound the term by taking absolute value individually, and in the last inequality we apply the induction assumption for  $1 \leq j \leq \ell - 1$ . For the first term, we can achieve a constant upper bound by following the exact procedure in the base case  $\ell = 1$ . For the second term, we can use the Hanson-Wright inequality (Lemma 32) and show

$$\begin{aligned}
\frac{|\mathbf{z}_{s_\ell}^\top \mathbf{A}_{-s_1:s_\ell}^{-1} \mathbf{z}_{s_j}|}{\mathbf{z}_{s_\ell}^\top \mathbf{A}_{-s_1:s_\ell}^{-1} \mathbf{z}_{s_\ell}} &\leq \frac{c_1 \|\mathbf{A}_{-s_1:s_\ell}^{-1}\|_2 \cdot n^{\frac{1}{2}+\epsilon}}{\text{Tr}(\mathbf{A}_{-s_1:s_\ell}^{-1}) - c_1 \|\mathbf{A}_{-s_1:s_\ell}^{-1}\|_2 \cdot n^{\frac{1}{2}+\epsilon}} \\
&\leq \frac{c_1 \|\mathbf{A}_{-(s_1:s_\ell) \cup [k^*]}^{-1}\|_2 \cdot n^{\frac{1}{2}+\epsilon}}{\text{Tr}(\mathbf{A}_{-s_1:s_\ell}^{-1}) - c_1 \|\mathbf{A}_{-(s_1:s_\ell) \cup [k^*]}^{-1}\|_2 \cdot n^{\frac{1}{2}+\epsilon}} \\
&\leq \frac{c_1 \|\mathbf{A}_{-(s_1:s_t) \cup [k^*]}^{-1}\|_2 \cdot n^{\frac{1}{2}+\epsilon}}{(1 - \frac{c}{n})^{k^*+t-\ell} \text{Tr}(\mathbf{A}_{-(s_1:s_t) \cup [k^*]}^{-1}) - c_1 \|\mathbf{A}_{-(s_1:s_t) \cup [k^*]}^{-1}\|_2 \cdot n^{\frac{1}{2}+\epsilon}} \\
&= \frac{1}{(1 - \frac{c}{n})^{k^*+t-\ell} \frac{\text{Tr}(\mathbf{A}_{-(s_1:s_t) \cup [k^*]}^{-1})}{c_1 \|\mathbf{A}_{-(s_1:s_t) \cup [k^*]}^{-1}\|_2 \cdot n^{\frac{1}{2}+\epsilon}} - 1} \\
&\leq \frac{1}{(1 - \frac{c}{n})^{k^*+t-\ell} \frac{n^{\frac{1}{2}-\epsilon}}{c} - 1} \\
&\leq \frac{c}{n^{\frac{1}{2}-\epsilon}},
\end{aligned}$$

where we use the fact that  $\mathbf{A}_{-(s_1:s_t) \cup [k^*]}^{-1} \succeq \mathbf{A}_{-s_1:s_\ell}^{-1}$  for  $\ell \leq t$  in the second inequality and we use Lemma 36 to get the trace lower bound in the third inequality. Finally, we again apply Lemma 35 to show eigenvalues in  $\mathbf{A}_{-(s_1:s_t) \cup [k^*]}^{-1}$  are identical up to a constant. Putting together the bounds for the first term and the second term, we complete the induction proof for  $j = \ell - 1$  case. The proof is done.  $\square$

## H Background lemmas

In this section, we provide statements and/or proofs of some miscellaneous lemmas.



The first lemma is the Hanson-Wright inequality, which demonstrates the quadratic term  $\mathbf{z}^\top \mathbf{M} \mathbf{z}$  of sub-Gaussian random vector  $\mathbf{z}$  concentrates around its expectation.

**Lemma 32.** (*Hanson-Wright Inequalities, Rudelson and Vershynin (2013)*) Let  $\mathbf{z}$  be a random vector with i.i.d. sub-Gaussian entries  $\mathbf{z}_i$  such that  $\mathbb{E}[\mathbf{z}_i] = 0$  and  $\|\mathbf{z}_i\|_{\psi_2} \leq 1$ . There exists a universal constant  $c > 0$  such that for any positive semi-definite matrix  $\mathbf{M}$  and for every  $t \geq 0$ , we have

$$\mathbb{P} [|\mathbf{z}^\top \mathbf{M} \mathbf{z} - \mathbb{E}[\mathbf{z}^\top \mathbf{M} \mathbf{z}]| > t] \leq 2 \exp \left\{ -c \min \left\{ \frac{t^2}{\|\mathbf{M}\|_F^2}, \frac{t}{\|\mathbf{M}\|_2} \right\} \right\}.$$

Note that  $\|\mathbf{M}\|_F^2 \leq n \|\mathbf{M}\|_2^2$  and we substitute  $t = c_1 \|\mathbf{M}\|_2 \cdot n^{\frac{1}{2}+\epsilon}$  where  $c_1^2 = \frac{1}{c}$  and  $\epsilon \in (0, \frac{1}{4})$  to get

$$|\mathbf{z}^\top \mathbf{M} \mathbf{z} - \mathbb{E}[\mathbf{z}^\top \mathbf{M} \mathbf{z}]| \leq c_1 \|\mathbf{M}\|_2 \cdot n^{\frac{1}{2}+\epsilon},$$

with probability at least  $1 - 2e^{-n^{2\epsilon}}$ . Again, note that  $\|\mathbf{M}\|_2 \leq \text{Tr}(\mathbf{M})$  and  $\|\mathbf{M}\|_F^2 = \text{Tr}(\mathbf{M}^2) \leq (\text{Tr}(\mathbf{M}))^2$ , and we substitute  $t = \frac{1}{c} \cdot \text{Tr}(\mathbf{M}) \cdot (\ln n)$  to get

$$\mathbf{z}^\top \mathbf{M} \mathbf{z} \leq \mathbb{E}[\mathbf{z}^\top \mathbf{M} \mathbf{z}] + \frac{1}{c} \text{Tr}(\mathbf{M}) \cdot (\ln n) \leq \left(1 + \frac{1}{c}\right) \cdot \text{Tr}(\mathbf{M}) \cdot (\ln n),$$

with probability at least  $1 - \frac{1}{n}$ . Note that the probabilities are over  $\mathbf{z}$  only and  $\mathbf{M}$  is positive semi-definite and is independent to  $\mathbf{z}$ .

The next lemma restates a bound on the squared norm of a Gaussian random vector.

**Lemma 33.** Let  $\mathbf{z} \sim \mathcal{N}(\boldsymbol{\mu}, \mathbf{I}_n)$  and for  $\delta \in (0, 1)$  and  $c > 1$ , we have

$$\frac{n}{c} = n(1 - \delta) \leq \|\mathbf{z}\|_2^2 \leq n(1 + \delta) = cn,$$

with probability at least  $1 - 2e^{-n\delta^2}$ .

The following lemma guarantees that if  $\boldsymbol{\Sigma}$  exhibits a heavy tail such that  $r_{k^*}(\boldsymbol{\Sigma}) \geq bn$ , it retains a heavy tail even after removing  $t \ll n^{\frac{1}{2}}$  components from the tail.

**Lemma 34.** For any data covariance matrix  $\boldsymbol{\Sigma}$  satisfying  $k^* \leq \frac{n}{c}$  such that  $r_{k^*}(\boldsymbol{\Sigma}) \geq bn$ , for any set of indices  $\mathcal{S}$  such that  $\mathcal{S} = \{j \mid k^* < j \leq d\}$  and  $|\mathcal{S}| = t \ll n^{\frac{1}{2}} \ll d$ , we have  $r_{k^*}(\boldsymbol{\Sigma}_{-s_1:s_t}) \geq bn - t \geq b_2 n$ .

*Proof.* By the definition of effective ranks,  $\boldsymbol{\Sigma}$  satisfies

$$r_{k^*}(\boldsymbol{\Sigma}) = \frac{\sum_{j=k^*+1}^d \lambda_j}{\lambda_{k^*+1}} \geq bn.$$

By removing  $t$  components whose index is larger than  $k^*$ , and denote  $j^* = \min\{j \mid k^* < j \leq k^*+1+t, j \in \mathcal{S}^c\}$ , we have

$$r_{k^*}(\boldsymbol{\Sigma}_{-s_1:s_t}) = \frac{\sum_{j=k^*+1, j \in \mathcal{S}^c}^d \lambda_j}{\lambda_{j^*}} \geq \frac{\sum_{j=k^*+1, j \in \mathcal{S}^c}^d \lambda_j}{\lambda_{k^*+1}} = \frac{\sum_{j=k^*+1}^d \lambda_j}{\lambda_{k^*+1}} - \frac{\sum_{j \in \mathcal{S}} \lambda_j}{\lambda_{k^*+1}} \geq bn - t,$$

where the first inequality follows  $\lambda_{k^*+1} \geq \lambda_{j^*}$  and the last inequality follows the lemma assumption on  $\mathcal{S}$  and  $\lambda_{k^*+1} \geq \lambda_j$  for all  $j \in \mathcal{S}$ . As a result, since  $t \ll n^{\frac{1}{2}}$ , the proof is complete.  $\square$

Next, we apply Lemma 34 to demonstrate that the eigenvalues of the tail of  $\boldsymbol{\Sigma}$ , after removing  $t \ll n^{\frac{1}{2}}$  components, remain identical up to a constant factor.

**Lemma 35.** For any data covariance matrix  $\Sigma$  satisfying  $k^* \leq \frac{n}{c}$  such that  $r_{k^*}(\Sigma) \geq bn$ , for any set of indices  $\mathcal{S}$  such that  $\mathcal{S} = \{j \mid k^* < j \leq d\}$  and  $|\mathcal{S}| = t \ll n^{\frac{1}{2}-\epsilon} \ll d$ , we have  $r_0(\Sigma_{-[k^*] \cup (s_1:s_t)}) = r_{k^*}(\Sigma_{-s_1:s_t}) \geq bn$ . Therefore, we have

$$\frac{1}{cn} \leq \frac{\left\| \mathbf{A}_{-[k^*] \cup (s_1:s_t)}^{-1} \right\|_2}{\text{Tr} \left( \mathbf{A}_{-[k^*] \cup (s_1:s_t)}^{-1} \right)} \leq \frac{c}{n},$$

for  $c \geq 1$  with probability at least  $1 - 2e^{-\frac{n}{\sqrt{c}}}$ .

*Proof.* According to Lemma 34, we have  $r_{k^*}(\Sigma_{-s_1:s_t}) = \frac{\sum_{j=k^*+1, j \in \mathcal{S}^c} \lambda_j}{\lambda_{j^*}} \geq bn$ , where we denote  $j^* = \min\{j \mid k^* < j \leq k^* + 1 + t, j \in \mathcal{S}^c\}$ . Furthermore, by re-indexing eigenvalues, we denote  $\{\tilde{\lambda}_j\}_{j=1}^{d-k^*-t}$  the eigenvalues of the leave- $k^*$  and  $t$ -out covariance matrix  $\Sigma_{-(s_1:s_t) \cup [k^*]}$ , and we have  $r_0(\Sigma_{-(s_1:s_t) \cup [k^*]}) = \frac{\sum_{j=1}^{d-k^*-t} \tilde{\lambda}_j}{\tilde{\lambda}_1} = r_{k^*}(\Sigma_{-s_1:s_t}) \geq bn$ . Based on Lemma 10 in Bartlett et al. (2020), for  $\Sigma_{-(s_1:s_t) \cup [k^*]}$ , we have

$$\frac{1}{c} \tilde{\lambda}_1 r_0(\Sigma_{-(s_1:s_t) \cup [k^*]}) \leq \mu_n(\mathbf{A}_{-(s_1:s_t) \cup [k^*]}) \leq \mu_1(\mathbf{A}_{-(s_1:s_t) \cup [k^*]}) \leq c \tilde{\lambda}_1 r_0(\Sigma_{-(s_1:s_t) \cup [k^*]}),$$

with probability at least  $1 - 2e^{-\frac{n}{c}}$ . Therefore, we have the bounds for  $\left\| \mathbf{A}_{-[k^*] \cup (s_1:s_t)}^{-1} \right\|_2$  as

$$\begin{aligned} \left\| \mathbf{A}_{-[k^*] \cup (s_1:s_t)}^{-1} \right\|_2 &= \frac{1}{\mu_n(\mathbf{A}_{-[k^*] \cup (s_1:s_t)})} \leq \frac{c}{\tilde{\lambda}_1 r_0(\Sigma_{-(s_1:s_t) \cup [k^*]})} \\ \left\| \mathbf{A}_{-[k^*] \cup (s_1:s_t)}^{-1} \right\|_2 &= \frac{1}{\mu_n(\mathbf{A}_{-[k^*] \cup (s_1:s_t)})} \geq \frac{1}{\mu_1(\mathbf{A}_{-[k^*] \cup (s_1:s_t)})} \geq \frac{1}{c \tilde{\lambda}_1 r_0(\Sigma_{-(s_1:s_t) \cup [k^*]})}. \end{aligned}$$

Similarly, for  $\text{Tr}(\mathbf{A}_{-[k^*] \cup (s_1:s_t)}^{-1})$ , we have

$$\begin{aligned} \text{Tr}(\mathbf{A}_{-[k^*] \cup (s_1:s_t)}^{-1}) &= \sum_{i=1}^n \frac{1}{\mu_i(\mathbf{A}_{-[k^*] \cup (s_1:s_t)})} \leq \frac{n}{\mu_n(\mathbf{A}_{-[k^*] \cup (s_1:s_t)})} \leq \frac{cn}{\tilde{\lambda}_1 r_0(\Sigma_{-(s_1:s_t) \cup [k^*]})} \\ \text{Tr}(\mathbf{A}_{-[k^*] \cup (s_1:s_t)}^{-1}) &= \sum_{i=1}^n \frac{1}{\mu_i(\mathbf{A}_{-[k^*] \cup (s_1:s_t)})} \geq \frac{n}{\mu_1(\mathbf{A}_{-[k^*] \cup (s_1:s_t)})} \geq \frac{n}{c \tilde{\lambda}_1 r_0(\Sigma_{-(s_1:s_t) \cup [k^*]})}. \end{aligned}$$

By substituting these bounds into  $\frac{\left\| \mathbf{A}_{-[k^*] \cup (s_1:s_t)}^{-1} \right\|_2}{\text{Tr}(\mathbf{A}_{-[k^*] \cup (s_1:s_t)}^{-1})}$ , the proof is done.  $\square$

The following lemma extends Lemma 25 from Wang et al. (2023) to show the trace bounds when removing distinct components in  $\Sigma$ , whereas the original result only provided the lower bound for removing the top  $k$  components.

**Lemma 36** (From Lemma 25 in Wang et al. (2023)). For any data covariance matrix  $\Sigma$  satisfying  $k^* \leq \frac{n}{c}$  such that  $r_{k^*}(\Sigma) \geq bn$ , for any set of indices  $\mathcal{S}$  such that  $\mathcal{S} = \{j \mid k^* < j \leq d\}$  and  $|\mathcal{S}| = t \ll n^{\frac{1}{2}} \ll d$ , for any  $0 \leq \ell_1 \leq \ell_2 \leq t$  and  $0 \leq k_1 \leq k_2 \leq k^*$  and sufficiently large  $n$ , we have

$$\text{Tr}(\mathbf{A}_{-(s_1:s_{\ell_2}) \cup [k_2]}^{-1}) \geq \text{Tr}(\mathbf{A}_{-(s_1:s_{\ell_1}) \cup [k_1]}^{-1}) \geq \left(1 - \frac{c}{n}\right)^{k_2 - k_1 + \ell_2 - \ell_1} \text{Tr}(\mathbf{A}_{-(s_1:s_{\ell_2}) \cup [k_2]}^{-1}),$$

with a probability at least  $1 - 2(k_2 - k_1 + \ell_2 - \ell_1)e^{-\frac{n}{c}}$ .

*Proof.* For the first inequality, it directly holds since  $\mathbf{A}_{-(s_1:s_{\ell_1}) \cup [k_1]} \succeq \mathbf{A}_{-(s_1:s_{\ell_2}) \cup [k_2]}$  implies  $\mathbf{A}_{-(s_1:s_{\ell_2}) \cup [k_2]}^{-1} \succeq \mathbf{A}_{-(s_1:s_{\ell_1}) \cup [k_1]}^{-1}$ . Next, according to Lemma 25 in Wang et al. (2023), by removing the top component in  $\Sigma$ , we have

$$\text{Tr}(\mathbf{A}_{-(s_1:s_{\ell_1}) \cup [k_1]}^{-1}) \geq \left(1 - \frac{c}{n}\right)^{k_2 - k_1} \text{Tr}(\mathbf{A}_{-(s_1:s_{\ell_1}) \cup [k_2]}^{-1}).$$

Next, following the proof steps in Lemma 25 in Wang et al. (2023), by the Sherman-Morrison-Woodbury identity, we have

$$\begin{aligned}
\text{Tr} \left( \mathbf{A}_{-(s_1:s_{\ell_1}) \cup [k_2]}^{-1} \right) &= \text{Tr} \left( \mathbf{A}_{-(s_1:s_{\ell_1+1}) \cup [k_2]}^{-1} \right) - \frac{\lambda_{s_{\ell_1+1}} \text{Tr} \left( \mathbf{A}_{-(s_1:s_{\ell_1+1}) \cup [k_2]}^{-1} \mathbf{z}_{s_{\ell_1+1}} \mathbf{z}_{s_{\ell_1+1}}^\top \mathbf{A}_{-(s_1:s_{\ell_1+1}) \cup [k_2]}^{-1} \right)}{1 + \lambda_{s_{\ell_1+1}} \mathbf{z}_{s_{\ell_1+1}}^\top \mathbf{A}_{-(s_1:s_{\ell_1+1}) \cup [k_2]}^{-1} \mathbf{z}_{s_{\ell_1+1}}} \\
&= \text{Tr} \left( \mathbf{A}_{-(s_1:s_{\ell_1+1}) \cup [k_2]}^{-1} \right) - \frac{\lambda_{s_{\ell_1+1}} \mathbf{z}_{s_{\ell_1+1}}^\top \mathbf{A}_{-(s_1:s_{\ell_1+1}) \cup [k_2]}^{-2} \mathbf{z}_{s_{\ell_1+1}}}{1 + \lambda_{s_{\ell_1+1}} \mathbf{z}_{s_{\ell_1+1}}^\top \mathbf{A}_{-(s_1:s_{\ell_1+1}) \cup [k_2]}^{-1} \mathbf{z}_{s_{\ell_1+1}}} \\
&\geq \text{Tr} \left( \mathbf{A}_{-(s_1:s_{\ell_1+1}) \cup [k_2]}^{-1} \right) - \left\| \mathbf{A}_{-(s_1:s_{\ell_1+1}) \cup [k_2]}^{-1} \right\|_2 \cdot \frac{\lambda_{s_{\ell_1+1}} \mathbf{z}_{s_{\ell_1+1}}^\top \mathbf{A}_{-(s_1:s_{\ell_1+1}) \cup [k_2]}^{-1} \mathbf{z}_{s_{\ell_1+1}}}{1 + \lambda_{s_{\ell_1+1}} \mathbf{z}_{s_{\ell_1+1}}^\top \mathbf{A}_{-(s_1:s_{\ell_1+1}) \cup [k_2]}^{-1} \mathbf{z}_{s_{\ell_1+1}}} \\
&\geq \text{Tr} \left( \mathbf{A}_{-(s_1:s_{\ell_1+1}) \cup [k_2]}^{-1} \right) - \left\| \mathbf{A}_{-(s_1:s_{\ell_1+1}) \cup [k_2]}^{-1} \right\|_2, \tag{51}
\end{aligned}$$

where in the first inequality, we use the property that  $\mathbf{x}^\top \mathbf{M}^2 \mathbf{x} \leq \|\mathbf{M}\|_2 \cdot \mathbf{x}^\top \mathbf{M} \mathbf{x}$ . Next, we have

$$\frac{\left\| \mathbf{A}_{-(s_1:s_{\ell_1+1}) \cup [k_2]}^{-1} \right\|_2}{\text{Tr} \left( \mathbf{A}_{-(s_1:s_{\ell_1+1}) \cup [k_2]}^{-1} \right)} \leq \frac{\left\| \mathbf{A}_{-(s_1:s_{\ell_1+1}) \cup [k^*]}^{-1} \right\|_2}{\text{Tr} \left( \mathbf{A}_{-(s_1:s_{\ell_1+1}) \cup [k_2]}^{-1} \right)} \leq \frac{\left\| \mathbf{A}_{-(s_1:s_{\ell_1+1}) \cup [k^*]}^{-1} \right\|_2}{(1 - \frac{c}{n})^{k^* - k_2} \text{Tr} \left( \mathbf{A}_{-(s_1:s_{\ell_1+1}) \cup [k^*]}^{-1} \right)} \leq \frac{c}{(1 - \frac{c}{n})^{k^* - k_2} n},$$

where in the first inequality, we apply the property that  $\mathbf{A}_{-(s_1:s_{\ell_1+1}) \cup [k_2]}^{-1} \preceq \mathbf{A}_{-(s_1:s_{\ell_1+1}) \cup [k^*]}^{-1}$ , in the second inequality, we use Lemma 25 in Wang et al. (2023), and in the last inequality, we apply Lemma 35. As a result, from Equation (51), by dividing  $\text{Tr} \left( \mathbf{A}_{-(s_1:s_{\ell_1+1}) \cup [k_2]}^{-1} \right)$  on both sides, we have

$$\frac{\text{Tr} \left( \mathbf{A}_{-(s_1:s_{\ell_1}) \cup [k_2]}^{-1} \right)}{\text{Tr} \left( \mathbf{A}_{-(s_1:s_{\ell_1+1}) \cup [k_2]}^{-1} \right)} \geq 1 - \frac{\left\| \mathbf{A}_{-(s_1:s_{\ell_1+1}) \cup [k_2]}^{-1} \right\|_2}{\text{Tr} \left( \mathbf{A}_{-(s_1:s_{\ell_1+1}) \cup [k_2]}^{-1} \right)} \geq 1 - \frac{c}{(1 - \frac{c}{n})^{k^* - k_2} n} \geq 1 - \frac{c_1}{n},$$

By applying the steps  $\ell_2 - \ell_1$  times, the proof is complete.  $\square$

## I Additional simulations

In this section, we present additional simulations of our few-shot postprocessing algorithm. In general, we find that our postprocessing algorithm can recover  $t$ -sparse signal in cases where the classification and regression tasks fail — even including the worst-case scenario of isotropic covariance.

Our simulations were run on an Nvidia A5000 GPU with 24GB VRAM, but this level of compute is not necessary. Our code is available at <https://github.com/tmlabonte/taskshift>.

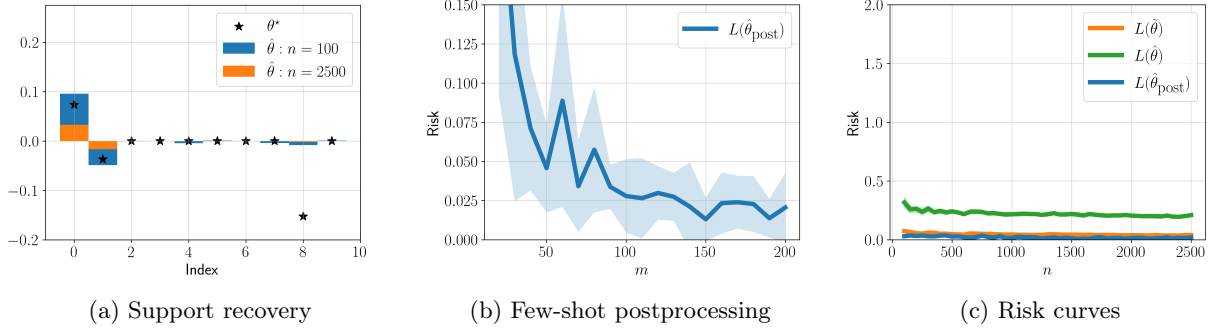


Figure 3: **Task shift for spiked covariance with  $q < 1 - r$  and signal outside the spike.** We set  $p = 1.5$ ,  $q = 0.5$ , and  $r = 0.25$  so that  $q < 1 - r$ . Moreover, we add an additional signal component which lies outside the covariance spike for any  $n \leq 2500$ . Our task shift algorithm correctly recovers the support and generalizes well; note that the decay of the component outside the spike (index 8) is faster than those in the spike (indices 1-2), but still slower than those outside the support (indices 3-7 and 9-10). The true signal  $\theta^*$  is 3-sparse with  $a_1 = 1$ ,  $a_2 = -0.5$ , and  $a_8 = -0.15$  (see Assumption 5). Our postprocessing algorithm uses top- $t$  support recovery and least-squares on noisy  $m$ -shot regression data. We plot the mean and standard deviation over 10 draws of the training dataset  $\mathbf{X}$ .

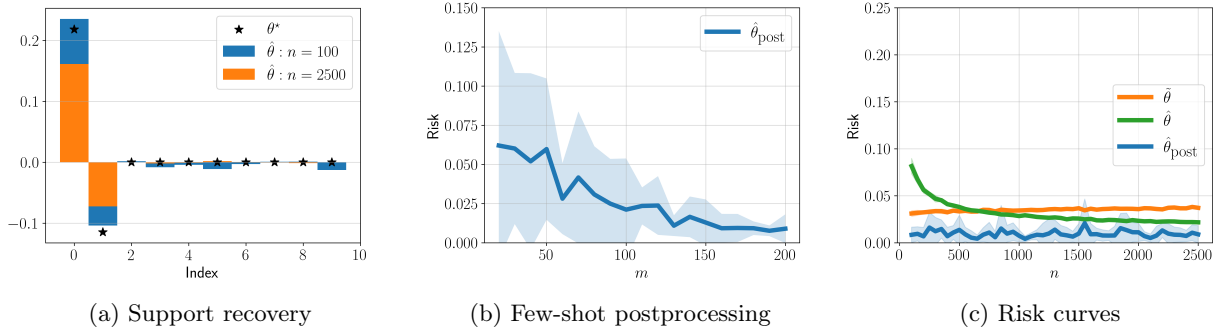


Figure 4: **Task shift for polynomial covariance with  $u = 0.25, v = 0$ .** Our task shift estimator generalizes for polynomial covariance models. The true signal  $\theta^*$  is 2-sparse with  $a_1 = 0.2$  and  $a_2 = -0.1$  (see Assumption 5), and we set  $d = n^{1.5}$ . Note that this parameterization satisfies the conditions of Corollary 31. Our postprocessing algorithm uses top- $t$  support recovery and least-squares on noisy  $m$ -shot regression data. We plot the mean and standard deviation over 10 draws of the training dataset  $\mathbf{X}$ .

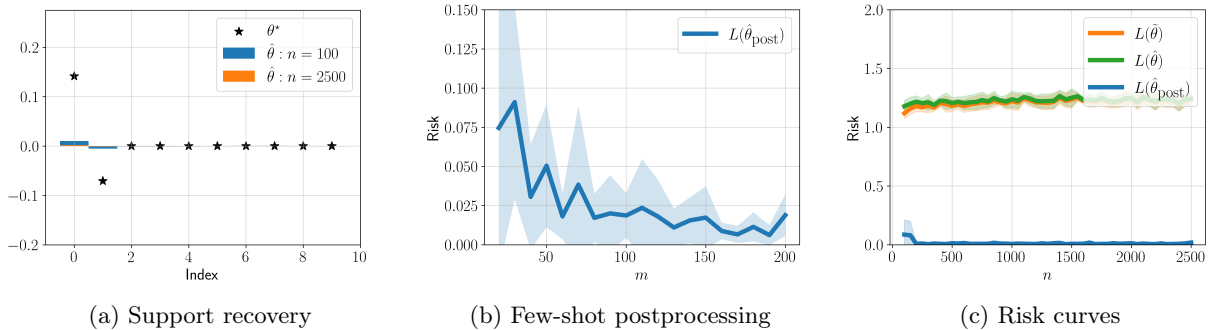
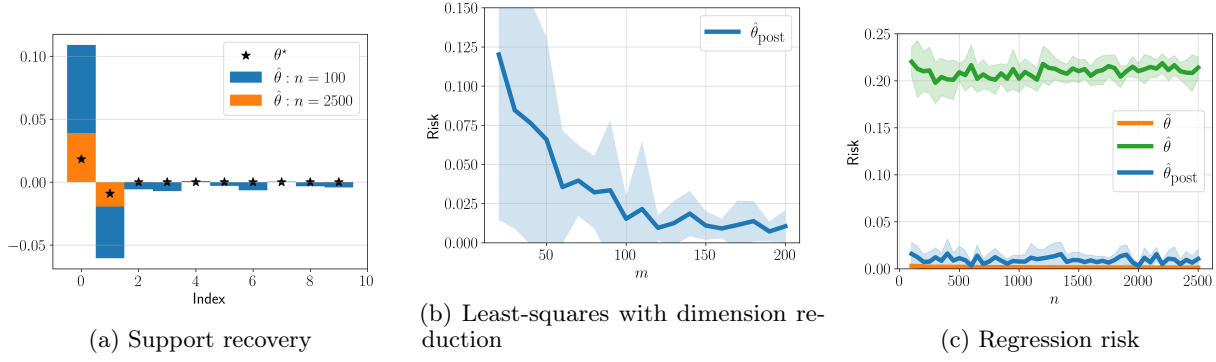
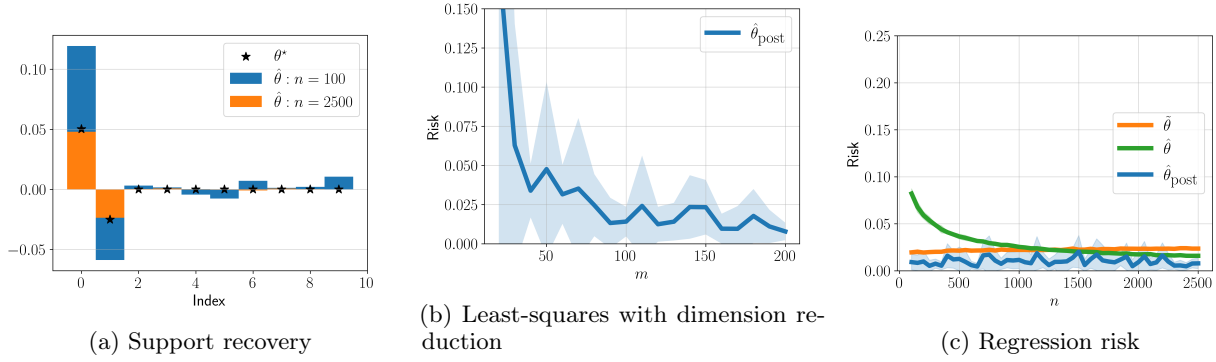


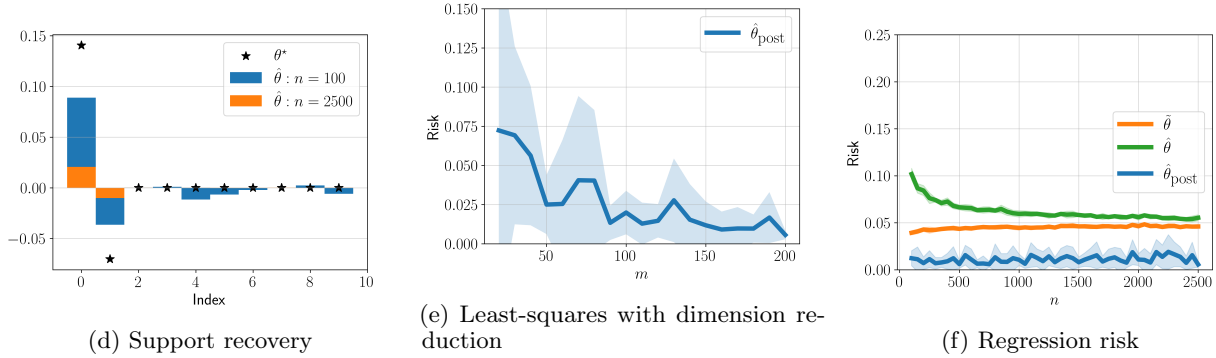
Figure 5: **Task shift for isotropic covariance  $\Sigma = 50I$ .** Our task shift estimator generalizes even in worst-case scenarios for minimum  $\ell_2$ -norm interpolation such as isotropic covariance. The true signal  $\theta^*$  is 2-sparse with  $a_1 = 1$  and  $a_2 = -0.5$  (see Assumption 5), and we set  $d = n^{1.5}$ . Our postprocessing algorithm uses top- $t$  support recovery and least-squares on noisy  $m$ -shot regression data. We plot the mean and standard deviation over 10 draws of the training dataset  $\mathbf{X}$ .



(i) **Task shift for spiked covariance when classification and regression generalize.** We set  $p = 1.5$ ,  $q = 0.3$ , and  $r = 0.5$  so that  $0 < q < 1 - r$ . In this regime, the classification MNI  $\hat{\theta}$  generalizes on the original classification problem and the regression MNI  $\tilde{\theta}$  generalizes on the original regression problems.



(ii) **Task shift for spiked covariance when classification generalizes but regression does not.** We set  $p = 1.5$ ,  $q = 0.6$ , and  $r = 0.5$  so that  $1 - r < q < (1 - r) + (p - 1)/2$ . In this regime, the classification MNI  $\hat{\theta}$  generalizes on the original classification problem, but the regression MNI  $\tilde{\theta}$  does not generalize on the original regression problems.



(iii) **Task shift for spiked covariance when neither classification nor regression generalize.** We set  $p = 1.5$ ,  $q = 0.9$ , and  $r = 0.5$  so that  $(1 - r) + (p - 1)/2 < q < p - r$ . In this regime, the classification MNI  $\hat{\theta}$  does not generalize on the original classification problem, and the regression MNI  $\tilde{\theta}$  does not generalize on the original regression problems.

Figure 6: **Postprocessing achieves task shift in the three regimes of Muthukumar et al. (2021).** The left column demonstrates the survival of  $t$ -sparse signal support components in the classification MNI  $\hat{\theta}$  while non-support components decay quickly. The middle column shows the  $O(\frac{t}{m})$  regression error of least-squares with reduction to  $t$  dimensions using  $m$  regression samples under standard Gaussian noise. Finally, the right column displays the regression risk of the classification MNI, regression MNI, and our postprocessed predictor. The signal  $\theta^*$  is 2-sparse with  $a_1 = 0.2$  and  $a_2 = -0.1$  (see Assumption 5). The middle column fixes  $n = 2500$ . We plot the mean and standard deviation over 10 draws of the training dataset  $\mathbf{X}$ .