

# 000 001 QUANTIFYING CROSS-DOMAIN KNOWLEDGE DISTIL- 002 LATION IN THE PRESENCE OF DOMAIN SHIFT 003 004

005 **Anonymous authors**

006 Paper under double-blind review

## 007 008 ABSTRACT 009

011 Cross-domain knowledge distillation often suffers from domain shift. Although  
012 domain adaptation methods have shown strong empirical success in addressing  
013 this issue, their theoretical foundations remain underdeveloped. In this paper, we  
014 study knowledge distillation in a teacher–student framework for regularized linear  
015 regression and derive high-dimensional asymptotic excess risk for the student es-  
016 timator, accounting for both covariate shift and model shift. This asymptotic anal-  
017 ysis enables a precise characterization of the performance gain in cross-domain  
018 knowledge distillation. Our results demonstrate that, even under substantial shifts  
019 between the source and target domains, it remains feasible to identify an imita-  
020 tion parameter for which the student model outperforms the student-only base-  
021 line. Moreover, we show that the student’s generalization performance exhibits  
022 the double descent phenomenon.

## 024 1 INTRODUCTION

026 The success of modern machine learning tasks typically requires the availability of large-scale la-  
027 beled datasets. However, collecting labeled data for a new target task is often challenging and ex-  
028 pensive. When data in the target domain is scarce, it is possible to leverage labeled data from related  
029 source domains. Knowledge distillation (KD) (Hinton et al., 2015), originally proposed for model  
030 compression, is a popular technique that transfers knowledge from a capable teacher model trained  
031 on a source domain to a smaller student model. This is achieved by guiding the student model to  
032 mimic the teacher model’s outputs. The extra information in the teacher’s predictions often improves  
033 the student model’s performance when target domain data is limited. KD has recently achieved re-  
034 markable success across several fields including image classification (Radford et al., 2021; Li et al.,  
035 2024), speech recognition (Mingote et al., 2020), and language models (Gu et al., 2023; Agarwal  
036 et al., 2024).

037 We denote the source domain data and target domain data as  $(\mathbf{X}_1, \mathbf{y}_1)$  and  $(\mathbf{X}_2, \mathbf{y}_2)$ , respectively.  
038 This work focuses on the following cross-domain KD process: a teacher model is first trained on the  
039 source domain data, and its predicted labels for the target domain inputs are then used to supervise  
040 the training of the student model by minimizing the per-sample objective function,

$$041 \quad \mathcal{L}(\xi) = \xi \ell(y_2^t, y_2^s) + (1 - \xi) \ell(y_2, y_2^s), \quad (1)$$

042 where  $\ell$  denotes the loss function,  $y_2$  is the ground-truth label,  $y_2^t$  is the teacher’s predicted label,  
043 and  $y_2^s$  denotes the student’s prediction. The weight parameter  $\xi$ , known as the imitation parameter  
044 (Lopez-Paz et al., 2015), balances the contributions of the teacher’s predictions and the observed  
045 labels during training. **We summarize the key findings of this paper in the following informal lemma.**

046 **Lemma 1. (informal)** *Under mild conditions, the excess risk in linear regression with quadratic*  
047 *loss admits a unique minimizer  $\xi^*$ , which can be negative.*

049 Cross-domain KD often suffers from a *shift* between the source and target domains. For instance,  
050 the source domain may consist of standard American English speech, while a region-specific voice  
051 assistant must handle local dialects. Another example is a face detection model trained on images  
052 of light-skinned individuals (source domain) being applied to images of dark-skinned individuals  
053 (target domain). Learning a discriminative predictor under such domain shifts between source and  
target domains is known as domain adaptation (Ganin et al., 2016). While much of the literature on

054 domain adaptation has focused on improving the performance of KD, relatively little is understood  
 055 about when – and how effectively – the student model can learn from the teacher in the presence of  
 056 domain shift.

057 Recently, [Emrullah Ildiz et al. \(2025\)](#) analyzed the weak-to-strong (W2S) generalization of linear  
 058 models in a cross-domain setting, and identified the form of the optimal surrogate model. However,  
 059 their analysis relies on the condition that the covariance matrices of the source and target domains are  
 060 jointly diagonalizable, which limits its ability to capture the influence of eigenvectors. Moreover,  
 061 their results are restricted to the setting  $\xi = 1$  (i.e., pure teacher supervision), leaving the trade-  
 062 off between distillation and learning from observed student data unexplored. Motivated by these  
 063 limitations, we take a step toward a more complete understanding of the performance gains of cross-  
 064 domain KD for linear regression.

065 In this paper, we present a theoretical analysis of cross-domain KD in the context of linear mod-  
 066 els, leveraging tools from random matrix theory. For ridge regression, we study two cases: (i) a  
 067 deterministic-parameter setting, in which the teacher and student parameter vectors are non-random;  
 068 and (ii) a random-parameter setting, in which a shared parameter vector is drawn from a prior distri-  
 069 bution. We also extend our analysis to the ridgeless regression setting. All proofs of the theoretical  
 070 results are provided in the appendix. We summarize our contributions as follows:

- 071 • **High-dimensional risk characterization.** We derive precise high-dimensional asymptotic-  
 072 ics for the risk of cross-domain KD via a bias–variance decomposition. Our results  
 073 reveal how the excess risk depends on the parameter vectors and the input distributions in  
 074 both domains, generalizing the student-only setting of [Hastie et al. \(2022\)](#).
- 075 • **Utility of cross-domain KD.** (“*Stones from other mountains can polish jade*”). Intuitively,  
 076 large domain shifts between the teacher’s and student’s training data might limit – or even  
 077 harm – the value of teacher supervision for the student’s generalization. Surprisingly, our  
 078 analysis shows that even under substantial domain discrepancies, it is still possible to find  
 079 an  $\xi \in \mathbb{R}$  such that the student model can outperform the student-only baseline. The  
 080 existence of such  $\xi$  depends on the geometry of the models and the covariance matrices of  
 081 both domains.
- 082 • **Double descent phenomenon.** We observe that the excess risk, as a function of  
 083 the dimension-to-sample-size ratio, exhibits the double-descent phenomenon in KD for  
 084 teacher-student model – previously documented by [Hastie et al. \(2022\)](#); [Nakkiran et al.](#)  
 085 (2021) in student-only models, and by [Moniri & Hassani \(2025\)](#) for  $\xi = 1$  under no do-  
 086 main shift with isotropic covariance.

## 088 1.1 RELATED WORKS

089 **Theory of KD.** In recent years, a growing body of work has sought to understand the effects of KD.  
 090 The theoretical understanding of distillation began with [Phuong & Lampert \(2019\)](#), who initially  
 091 investigated linear student networks. [Wei et al. \(2021\)](#); [Borup & Andersen \(2021\)](#); [Das & Sanghavi](#)  
 092 ([2023](#)); [Pareek et al. \(2024\)](#); [Jeong & Chung \(2025\)](#) theoretically studied self-distillation, a variant  
 093 of KD in which the student model has the same architecture as the teacher and is trained on the  
 094 same data. [Menon et al. \(2021\)](#) showed that a “Bayes teacher” providing true class probabilities can  
 095 reduce the variance of the student’s objective, leading to improved performance. [Harutyunyan et al.](#)  
 096 ([2023](#)) proposed a framework that highlighted a delicate interplay among the teacher’s accuracy, the  
 097 student’s margin with respect to the teacher predictions, and the complexity of the teacher predictions.  
 098 From an information-theoretic perspective, [Dissanayake et al. \(2025\)](#) quantified and explained  
 099 the transferred knowledge and knowledge left to distill for a downstream task.

100 **Cross-domain KD and domain adaptation.** Many studies have explored various methods to ad-  
 101 dress the domain shift problem in the field of KD. Empirical works include [Su & Maji \(2016\)](#);  
 102 [Kundu et al. \(2019\)](#); [Asami et al. \(2017\)](#); [Li et al. \(2023\)](#); [Xu et al. \(2024\)](#); [Tang et al. \(2025\)](#). [Ye et al. \(2024\)](#) proposed the **Maximum Conditional Mutual Information** method, which enables the  
 103 teacher model to capture more contextual information to generate more accurate estimates of the  
 104 Bayes conditional probability distribution. The emergence of large language models (LLMs) has  
 105 brought new advancements, such as distillation across vastly different architectures and scalable  
 106 cross-domain transfer. For more details, readers may refer to [Fedus et al. \(2022\)](#); [Ouyang et al.](#)  
 107 ([2022](#)); [Yang et al. \(2024\)](#). From a theoretical perspective, [Emrullah Ildiz et al. \(2025\)](#) focused

108 on the setting where the student is trained using only the teacher’s predictions, and analyzed the  
 109 conditions under which the student can outperform the teacher in cross-domain KD.  
 110

111 **Weak-to-strong generalization.** Weak-to-strong (W2S) generalization (Burns et al., 2024), which  
 112 concerns using predictions generated by a weaker teacher model to train a more powerful student  
 113 model, is closely connected to KD. Emrullah Ildiz et al. (2025) provided an analysis of ridgeless  
 114 regression and proved that when using a weak model as the surrogate (teacher), W2S training can  
 115 provably outperform training with true labels. Charikar et al. (2024) assumed that the models are  
 116 selected over a convex set, and quantified the gain of the weak-label trained strong model over  
 117 the weak model. Wu & Sahai (2025) explored W2S generalization for classification in a spiked  
 118 covariance model. Medvedev et al. (2025) explained how W2S generalization can arise in random  
 119 feature models described by two-layer networks. Theoretical research in this area has continued to  
 120 grow, see Dong et al. (2025); Shin et al. (2025); Moniri & Hassani (2025); Oh et al. (2025), for  
 121 example.

## 122 1.2 NOTATIONS

123 We use  $\|\cdot\|$  to denote the spectral norm for matrices and the Euclidean norm for vectors, and  $\|\cdot\|_F$   
 124 for the Frobenius norm of a matrix. Standard big-O and small-o notations are employed. Moreover,  
 125 we denote  $x_n = o_{a.s.}(a_n)$ , if  $x_n/a_n \rightarrow 0$  almost surely. For any sequences  $a_n \geq 0$  and  $b_n \geq 0$ , we  
 126 write  $a_n \lesssim b_n$  if  $a_n = O(b_n)$ , and  $a_n \sim b_n$  if both  $a_n \lesssim b_n$  and  $b_n \lesssim a_n$ . We use  $\delta(\cdot)$  to denote the  
 127 indicator function, which takes the value 1 if the condition  $\cdot$  holds, and 0 otherwise. Throughout the  
 128 paper,  $c$  and  $C$  denote constants that may vary from line to line. For a random variable  $x$ , we use  
 129  $x \sim D$  to indicate that  $x$  follows the distribution  $D$ .  
 130

## 131 2 PRELIMINARIES

### 132 2.1 PROBLEM SETUP

133 Suppose there are  $N_1$  covariates  $\{\mathbf{x}_j^{(1)}\}_{j=1}^{N_1}$  drawn i.i.d. from an  $M$ -dimensional source distribution  
 134  $D_1$  and  $N_2$  covariates  $\{\mathbf{x}_j^{(2)}\}_{j=1}^{N_2}$  drawn i.i.d. from an  $M$ -dimensional target distribution  $D_2$ . We  
 135 consider a linear regression task specified by an unknown parameter vector  $\beta_i \in \mathbb{R}^M$ :

$$136 \quad y_j^{(i)} = \beta_i^\top \mathbf{x}_j^{(i)} + \varepsilon_j^{(i)}, \quad i = 1, 2, \quad 1 \leq j \leq N_i,$$

137 where  $\varepsilon_j^{(i)} \in \mathbb{R}$  is a zero-mean random noise term with variance  $\sigma^2$ . For  $i = 1, 2$  and  $z \in \mathbb{C} \setminus \mathbb{R}^+$ ,  
 138 define

$$139 \quad \mathbf{X}_i = (\mathbf{x}_1^{(i)}, \dots, \mathbf{x}_{N_i}^{(i)}) \in \mathbb{R}^{M \times N_i}, \quad \mathbf{y}_i = (y_1^{(i)}, \dots, y_{N_i}^{(i)})^\top \in \mathbb{R}^{N_i},$$

$$140 \quad \mathbf{Q}_i(z) = \left( \frac{1}{N_i} \mathbf{X}_i \mathbf{X}_i^\top - z \mathbf{I}_M \right)^{-1}, \quad \varepsilon_i = (\varepsilon_1^{(i)}, \dots, \varepsilon_{N_i}^{(i)})^\top \in \mathbb{R}^{N_i}.$$

141 We refer to the case where  $D_1 \neq D_2$  as a *covariate shift*, and the case where  $\beta_1 \neq \beta_2$  as a *model  
 142 shift*.

143 **Teacher Model:** The teacher model is finetuned on  $\{(\mathbf{x}_j^{(1)}, y_j^{(1)})\}_{j=1}^{N_1}$ :

$$144 \quad \beta_t = \arg \min_{\beta} \left( \frac{1}{N_1} \|\mathbf{y}_1 - \mathbf{X}_1^\top \beta\|^2 + \lambda_t \|\beta\|^2 \right) = \frac{1}{N_1} \mathbf{Q}_1(-\lambda_t) \mathbf{X}_1 \mathbf{y}_1, \quad (2)$$

145 where  $\lambda_t > 0$  is the teacher regularization parameter. The risk of  $\beta_t$  when  $M \sim N_1$  in the high-  
 146 dimensional setting has been studied extensively in the literature such as Dobriban & Wager (2018);  
 147 Hastie et al. (2022).

148 **Student Model Trained with Cross-Domain KD:** We use the pre-trained teacher model together  
 149 with covariates  $\{\mathbf{x}_j^{(2)}\}_{j=1}^{N_2}$  to generate predictions:  
 150

$$151 \quad \mathbf{y}_2^t = (y_1^t, \dots, y_{N_2}^t)^\top = (\mathbf{x}_1^{(2)}, \dots, \mathbf{x}_{N_2}^{(2)})^\top \beta_t.$$

162 The student model is finetuned on the target domain data  $\{(\mathbf{x}_j^{(2)}, y_j^{(2)})\}_{j=1}^{N_2}$  and the teacher's predictions  $\{(\mathbf{x}_j^{(2)}, y_j^t)\}_{j=1}^{N_2}$ , using the per-sample objective function defined in equation 1 with an imitation parameter  $\xi$ , as follows:

$$\begin{aligned} 166 \quad \beta_s &= \arg \min_{\beta} \mathcal{L}(\xi) = \arg \min_{\beta} \xi \left( \frac{1}{N_2} \|\mathbf{y}_2^t - \mathbf{X}_2^\top \beta\|^2 \right) + (1 - \xi) \left( \frac{1}{N_2} \|\mathbf{y}_2 - \mathbf{X}_2^\top \beta\|^2 \right) + \lambda_s \|\beta\|^2 \\ 167 \quad &= (\mathbf{X}_2 \mathbf{X}_2^\top + N_2 \lambda_s \mathbf{I}_M)^{-1} (\xi \mathbf{X}_2 \mathbf{y}_2^t + (1 - \xi) \mathbf{X}_2 \mathbf{y}_2), \end{aligned} \quad (3)$$

170 where  $\lambda_s$  is the student regularization parameter. While it is common to restrict  $\xi \in [0, 1]$  (Lopez-  
171 Paz et al., 2015), we do not impose this constraint, in line with Das & Sanghavi (2023); Pareek  
172 et al. (2024). **From equation 3, the parameter  $\xi$  is independent of  $\mathbf{Q}_2$ , making it possible to choose**  
173 **a negative  $\xi$  that achieves better generalization performance.** For the covariates  $\mathbf{x}_j^{(i)}$  and the noise  
174 terms  $\varepsilon_j^{(i)}$ ,  $i = 1, 2$ ,  $1 \leq j \leq N_i$ , we make the following assumptions, which are standard in the  
175 random matrix theory literature (see, e.g., Bai & Silverstein (2010)).

177 **Assumption 1.** *Suppose  $\mathbf{X}_1, \mathbf{X}_2, \varepsilon_1$ , and  $\varepsilon_2$  are mutually independent. Moreover, we assume*

178 (a) *the covariates are generated according to*

$$180 \quad \mathbf{X}_i = (\Sigma_i)^{1/2} \mathbf{Z}_i, \quad \text{for } i = 1, 2,$$

181 where  $\mathbf{Z}_i = (z_{jk}^{(i)})$  is an  $M \times N_i$  random matrix with i.i.d. entries of zero mean and unit variance,  
182 and  $\Sigma_i$  is a positive semi-definite matrix. Furthermore, we assume for all  $p \in \mathbb{N}$ , there is a constant  
183  $C_p$  such that

$$185 \quad \max_{i=1,2} \mathbb{E} |z_{11}^{(i)}|^p \leq C_p. \quad (4)$$

187 (b)  $M \sim N_1 \sim N_2$ .

188 (c)  $\varepsilon_i \in \mathbb{R}^{N_i}$  is a random vector consisting of i.i.d. entries of zero mean, variance  $\sigma^2$ , and for all  
189  $p \in \mathbb{N}$ , there is a constant  $c_p$  such that

$$191 \quad \max_{i=1,2} \mathbb{E} |\varepsilon_1^{(i)}|^p \leq c_p.$$

193 While we allow  $z_{11}^{(1)}$  and  $z_{11}^{(2)}$  to follow different distributions – a form of covariate shift – our theoretical  
194 results do not depend on their specific distributions, provided that the moment conditions in  
195 Assumption 1(a) are satisfied. The requirement that all moments of  $z_{11}^{(i)}$  exist can be relaxed to the  
196 existence of **(8 + c)-th moment** for any positive constant  $c$ , with minor modifications to our proof  
197 and hence we do not pursue this generalization here. The following assumption on the structure of  
198 the covariance matrices is imposed to facilitate theoretical analysis and rule out degenerate cases.

199 **Assumption 2.** *Let  $\tau$  be a small constant. Denote the eigenvalues of  $\Sigma_i$  by  $\sigma_1^i \geq \sigma_2^i \cdots \geq \sigma_M^i \geq 0$ .*

200 (a) *(Boundedness of  $\Sigma_i$ ). We assume that  $\max_{i=1,2} \|\Sigma_i\| = \sigma_1^i < \tau^{-1}$ .*

201 (b) *(Anti-concentration at 0). For  $i = 1, 2$ , the empirical spectral distribution of  $\Sigma_i$  satisfies*

$$205 \quad \frac{1}{M} \sum_{j=1}^M \delta(\sigma_j^i \leq \tau) \leq 1 - \tau.$$

208 Let  $(\mathbf{x}, y)$  be an unseen sample of the target task, that is  $y = \beta_2^\top \mathbf{x} + \varepsilon$ , where  $\mathbf{x} \sim D_2$  and  $\varepsilon$  follows  
209 the same distribution with  $\varepsilon_1^{(2)}$ . Under the mean squared loss, the generalization ability is quantified  
210 by the risk of the estimator  $\beta_s$ :

$$212 \quad \mathbf{R}(\beta_s) = \mathbb{E}_{\mathbf{x}, y} |y - \beta_s^\top \mathbf{x}|^2 = \mathbb{E}_{\mathbf{x}, y} |(\beta_2 - \beta_s)^\top \mathbf{x} + \varepsilon|^2 = \|\Sigma_2^{1/2}(\beta_2 - \beta_s)\|^2 + \sigma^2,$$

213 where  $\mathbb{E}_{\mathbf{x}, y}$  denotes the expectation taken with respect to (w.r.t.) the pair  $(\mathbf{x}, y)$ . The excess test risk  
214 is defined as follows:

$$215 \quad \mathbf{ER}(\beta_s) = \mathbf{R}(\beta_s) - \sigma^2 = \|\Sigma_2^{1/2}(\beta_2 - \beta_s)\|^2. \quad (5)$$

When  $\xi = 0$ ,  $\beta_s$  reduces to the ridge regression estimator for the student only model, and we denote the corresponding excess risk by  $\mathbf{ER}_0$ . Note that  $\mathbf{ER}(\beta_s)$  can be decomposed into bias and variance as  $\mathbf{ER}(\beta_s) = \mathbf{Bias} + \mathbf{Var}$ , where

$$\mathbf{Bias} = \|\Sigma_2^{1/2}(\beta_2 - \mathbb{E}_{x,y}\beta_s)\|^2,$$

$$\mathbf{Var} = \|\Sigma_2^{1/2}(\beta_s - \mathbb{E}_{x,y}\beta_s)\|^2.$$

**One may easily check that  $\mathbf{ER} = O(1)$  almost surely.** In the remainder of this paper, we derive asymptotic expressions for the bias and variance terms to analyze the generalization performance of the student model using tools from random matrix theory.

## 2.2 RANDOM MATRIX THEORY

Before proceeding to the theoretical analysis, we introduce several key quantities from random matrix theory that will appear in our main results. For any distribution  $G$  supported on  $\mathbb{R}^+ = [0, \infty)$ , its Stieltjes transform is defined as

$$m_G(z) = \int \frac{1}{x-z} dG(x), \quad z \notin \text{supp}(G).$$

Next, we define the asymptotic eigenvalue density of random matrices via its Stieltjes transform. This lemma is well-known in the random matrix theory literature (e.g., [Bai & Silverstein \(2010\)](#)).

**Lemma 2.** *Let  $\mathbf{X} = \Sigma^{1/2}\mathbf{Z}$  be a random matrix, where  $\mathbf{Z} = (z_{jk}) \in \mathbb{R}^{M \times N}$ ,  $M \sim N$  satisfies Assumption 1(a), and  $\Sigma$  satisfies Assumption 2. For each  $z \in \mathbb{C} \setminus \mathbb{R}^+$ , there exists a unique  $m \equiv m_M(z) \in \mathbb{C}$  satisfying the equation*

$$z = -\frac{1}{m} + \frac{1}{N} \text{Tr} \frac{\Sigma}{1+m\Sigma} = -\frac{1}{m} - \frac{z}{N} \text{Tr} \Sigma \Pi, \quad \text{with } \Im z \Im m(z) \geq 0, \quad (6)$$

where  $\Pi(z) = -(z + zm\Sigma)^{-1}$ .

## 3 THEORETICAL ANALYSIS

In this section, we analyze the excess risk  $\mathbf{ER}(\beta_s)$  defined in 5 under three distinct settings. In Section 3.1, we consider the case where  $\beta_1$  and  $\beta_2$  are deterministic, with their difference being arbitrary. In Section 3.2, we study the scenario in which  $\beta_1 = \beta_2$  and the common parameter vector is drawn from a prior distribution. Finally, in Section 3.3, we analyze ridgeless regression under the regime where  $M < N_1, N_2$  and the covariance matrices  $\Sigma_1, \Sigma_2$  are invertible.

Before presenting the main results, we first introduce some necessary notation. For  $M, N_i, \Sigma_i$  and  $z < 0$ , the Stieltjes transform determined by Lemma 2 is denoted by  $m_i(z)$ .

Let  $\Pi_i(z) = -(z + zm_i(z)\Sigma_i)^{-1}$ ,  $i = 1, 2$ . We write  $\mathbf{Q}_1 = \mathbf{Q}_1(-\lambda_t)$ ,  $\mathbf{Q}_2 = \mathbf{Q}_2(-\lambda_s)$ ,  $\Pi_1 = \Pi_1(-\lambda_t)$  and  $\Pi_2 = \Pi_2(-\lambda_s)$  for notational simplicity. For any deterministic matrix  $\mathbf{A}$  with bounded spectral norm, we define

$$\mathcal{S}_i(\mathbf{A}) = \mathbf{A} + \frac{\frac{1}{N_i} \text{Tr} \Sigma_i \Pi_i \mathbf{A} \Pi_i}{(1 + \frac{1}{N_i} \text{Tr} \Sigma_i \Pi_i)^2 - \frac{1}{N_i} \text{Tr}(\Sigma_i \Pi_i)^2} \Sigma_i, \quad i = 1, 2.$$

Moreover, when  $\mathbf{A} = \mathbf{I}_M$ , we denote

$$\Pi'_1 = \frac{d}{dz} \Pi_1(z) \Big|_{z=-\lambda_t} = \Pi_1 \mathcal{S}_1(\mathbf{I}_M) \Pi_1, \quad \Pi'_2 = \frac{d}{dz} \Pi_2(z) \Big|_{z=-\lambda_s} = \Pi_2 \mathcal{S}_2(\mathbf{I}_M) \Pi_2,$$

**The other quantities are summarized in Table 1.**

### 3.1 DETERMINISTIC REGRESSION PARAMETERS

We now state our first main result.

270 Table 1: Some notations used in the theoretical results  
271

272	$\mathbf{E}_1 = \mathbf{\Pi}_1 \mathcal{S}_1(\mathbf{\Sigma}_2) \mathbf{\Pi}_1, \quad \mathbf{E}_2 = \mathbf{\Pi}_1 \mathcal{S}_1(\mathbf{\Pi}_2 \mathcal{S}_2(\mathbf{\Sigma}_2) \mathbf{\Pi}_2) \mathbf{\Pi}_1$
273	$\mathbf{E}_3 = \mathbf{\Pi}_1 \mathcal{S}_1(\mathbf{\Sigma}_2 \mathbf{\Pi}_2) \mathbf{\Pi}_1, \quad \mathbf{E}_4 = \mathbf{\Pi}_2 \mathcal{S}_2(\mathbf{\Sigma}_2) \mathbf{\Pi}_2, \quad \mathbf{E}_5 = \mathbf{\Sigma}_2 \mathbf{\Pi}_2$

276 **Theorem 1.** Let  $\gamma = \beta_1 - \beta_2$ . For the deterministic vectors  $\|\beta_1\|$  and  $\|\beta_2\|$ , assume that  
277  $\|\beta_1\|, \|\beta_2\| \leq c$  for some constant  $c$ . Under Assumptions 1-2, the following results hold:

$$278 \quad \widehat{\text{Bias}} = \widehat{\text{Bias}} + o_{a.s.}(1), \quad \widehat{\text{Var}} = \widehat{\text{Var}} + o_{a.s.}(1),$$

280 where

$$281 \quad \widehat{\text{Bias}} = \xi^2 \beta_1^T [\lambda_t^2 \mathbf{E}_1 + \lambda_s^2 \lambda_t^2 \mathbf{E}_2 - 2\lambda_t^2 \lambda_s \mathbf{E}_3] \beta_1 + 2\xi \beta_2^T [\lambda_s^2 \mathbf{E}_4 - \lambda_s \mathbf{E}_5] \gamma \\ 282 \quad + \lambda_s^2 \beta_2^T \mathbf{E}_4 \beta_2 + 2\xi \beta_1^T [\lambda_t \lambda_s \mathbf{\Pi}_1 \mathbf{\Pi}_5 - \lambda_t \lambda_s^2 \mathbf{\Pi}_1 \mathbf{E}_4] \beta_2 + \xi^2 \gamma^T [-2\lambda_s \mathbf{E}_5 + \lambda_s^2 \mathbf{E}_4] \gamma \quad (7) \\ 284 \quad + 2\xi^2 \gamma^T [\lambda_s \lambda_t \mathbf{E}_5 \mathbf{\Pi}_1 - \lambda_t \lambda_s^2 \mathbf{E}_4 \mathbf{\Pi}_1 - \lambda_t \mathbf{\Sigma}_2 \mathbf{\Pi}_1 + \lambda_t \lambda_s \mathbf{E}_5 \mathbf{\Pi}_1] \beta_1,$$

285 and

$$287 \quad \widehat{\text{Var}} = \frac{\xi^2 \sigma^2}{N_1} \text{Tr}[(\mathbf{\Sigma}_2 - 2\lambda_s \mathbf{E}_5 + \lambda_s^2 \mathbf{E}_4)(\mathbf{\Pi}_1 - \lambda_t \mathbf{\Pi}'_1)] + \frac{(1 - \xi)^2 \sigma^2}{N_2} \text{Tr}[\mathbf{E}_5 - \lambda_s \mathbf{\Sigma}_2 \mathbf{\Pi}'_2]. \quad (8)$$

290 This theorem characterizes the dependence of  $\text{ER}(\beta_s)$  on the geometry of  $\mathbf{\Sigma}_1, \beta_1, \mathbf{\Sigma}_2$ , and  $\beta_2$ .  
291 We provide an illustrative example here. Suppose that  $\mathbf{\Sigma}_i$  admits the spectral decomposition  $\mathbf{\Sigma}_i =$   
292  $\mathbf{U}_i \mathbf{\Lambda}_i \mathbf{U}_i^T$ , for  $i = 1, 2$ . Consider the term  $\beta_1^T \mathbf{\Pi}_1 \mathbf{E}_5 \beta_2$ , which can be expressed as

$$294 \quad \beta_1^T (\lambda_t + \lambda_t m_1 \mathbf{\Sigma}_1)^{-1} (\lambda_s + \lambda_s m_2 \mathbf{\Sigma}_2)^{-1} \mathbf{\Sigma}_2 \beta_2 = (\lambda_s \lambda_t)^{-1} \tilde{\beta}_1^T (1 + m_1 \mathbf{\Lambda}_1)^{-1} \mathbf{U}_1^T \mathbf{U}_2 \tilde{\mathbf{\Lambda}}_2 \tilde{\beta}_2, \quad (9)$$

295 where  $\tilde{\mathbf{\Lambda}}_2$  is a diagonal matrix with entries  $\tilde{\Lambda}_{2,jj} = \frac{\Lambda_{2,jj}}{1 + m_2 \Lambda_{2,jj}}$ . The vector  $\tilde{\beta}_i = \mathbf{U}_i \beta_i$  captures  
296 the alignment between  $\beta_i$  and the eigenvectors of  $\mathbf{\Sigma}_i$ . The right-hand side of equation 9 explicitly  
297 reveals how the term  $\beta_1^T \mathbf{\Pi}_1 \mathbf{E}_5 \beta_2$  depends on  $\tilde{\beta}_i$ , the eigenvalues of  $\mathbf{\Sigma}_1$  and  $\mathbf{\Sigma}_2$ , and the eigenvector  
298 overlap  $\mathbf{U}_1^T \mathbf{U}_2$  between the two covariance matrices. In the special case where each  $\beta_i$  is aligned  
299 with an eigenvector of  $\mathbf{\Sigma}_i$  – for simplicity, suppose it corresponds to the first eigenvector – the  
300 expression equation 9 further simplifies to  $\beta_1^T \beta_2 (\lambda_s \lambda_t)^{-1} \frac{\Lambda_{2,11}}{(1 + m_1 \Lambda_{1,11})(1 + m_2 \Lambda_{2,11})}$ , which depends on  
301 the eigenvalues of  $\mathbf{\Sigma}_i$  and the inner product  $\beta_1^T \beta_2$ . This observation extends the results of [Hastie et al. \(2022\)](#), which considers high-dimensional least squares regression within a single domain  
302 (corresponding to  $\xi = 0$  in equation 3).

### 305 3.2 RANDOM REGRESSION PARAMETERS

307 In this section, we assume that the vector  $\beta_1 = \beta_2 = \beta$  is random, and consider the excess risk  
308 under two population covariance matrices,  $\mathbf{\Sigma}_1$  and  $\mathbf{\Sigma}_2$ , which may be equal or distinct. Before  
309 presenting the main result, we introduce the following assumption, commonly used in the literature  
310 ([Dobriban & Wager, 2018; Moniri & Hassani, 2025](#)).

311 **Assumption 3.** The regression parameter vector  $\beta = (\beta_1, \dots, \beta_M)^T \in \mathbb{R}^M$  is random, with each  
312 entry i.i.d., and  $\beta_1$  satisfies

$$314 \quad \mathbb{E} \beta_1 = 0, \quad \mathbb{E} \beta_1^2 = \frac{\tilde{\sigma}^2}{M}, \quad \text{and } \mathbb{E} |\sqrt{M} \beta_1|^p \leq C_p,$$

316 for any  $p \in \mathbb{N}$ , where  $C_p$  is a constant depending only on  $p$ .

317 **Theorem 2.** Suppose Assumptions 1-3 hold. Then the following asymptotic expressions hold:

$$319 \quad \widehat{\text{Bias}} = \frac{\tilde{\sigma}^2}{M} \left[ \xi^2 \lambda_t^2 \text{Tr} \mathbf{\Sigma}_2 \mathbf{\Pi}'_1 + 2\xi \lambda_t \lambda_s \text{Tr} \mathbf{\Pi}_1 \mathbf{\Pi}_2 \mathbf{\Sigma}_2 + \lambda_s^2 \text{Tr} \mathbf{\Sigma}_2 \mathbf{\Pi}'_2 \right. \\ 320 \quad \left. - 2\xi^2 \lambda_t^2 \lambda_s \text{Tr} \mathbf{\Sigma}_2 \mathbf{\Pi}_2 \mathbf{\Pi}'_1 + \xi \lambda_t \lambda_s^2 \text{Tr} [\mathbf{E}_4 (-2\mathbf{\Pi}_1 + \xi \lambda_t \mathbf{\Pi}'_1)] \right] = \text{Bias} + o_{a.s.}(1),$$

323 and  $\widehat{\text{Var}} = \text{Var} + o_{a.s.}$ , which coincides with the expression in Theorem 1.

This theorem extends the result of [Moniri & Hassani \(2025\)](#), which considers the case of no covariate shift, with inputs drawn i.i.d. from  $\mathcal{N}(0, \mathbf{I}_M)$  in the context of W2S generalization (i.e., when  $\xi = 1$ ). Our framework generalizes this analysis by allowing  $\xi \in \mathbb{R}$ , thereby providing a more comprehensive understanding of the trade-off between learning from the teacher and from the observed labels.

Let  $\underline{m}_1(z), \underline{m}_2(z)$  be the Stieltjes transforms of the standard Marchenko-Pastur law with parameters  $M/N_1, M/N_2$ , respectively; their explicit forms are given in equation 28. The following corollary follows immediately from Theorem 2 and the fact that  $\Pi_1 = \underline{m}_1 \mathbf{I}_M, \Pi_2 = \underline{m}_2 \mathbf{I}_M$  (see, e.g., [Alex et al. \(2014\)](#)).

**Corollary 1.** Suppose  $\Sigma_1 = \Sigma_2 = \mathbf{I}_M$ . Write  $\underline{m}_1 = \underline{m}_1(-\lambda_t), \underline{m}_2 = \underline{m}_2(-\lambda_s)$ . Under Assumption 1 and Assumption 3, we have the following expressions:

$$\begin{aligned} \text{Bias} &= \tilde{\sigma}^2 [\xi^2 \lambda_t^2 \underline{m}'_1 + 2\xi \lambda_t \lambda_s \underline{m}_1 \underline{m}_2 + \lambda_s^2 \underline{m}'_2 - 2\xi^2 \lambda_t^2 \lambda_s \underline{m}_2 \underline{m}'_1 \\ &\quad - 2\xi \lambda_t \lambda_s^2 \underline{m}'_2 \underline{m}_1 + \xi^2 \lambda_t^2 \lambda_s^2 \underline{m}'_1 \underline{m}'_2] + o_{a.s.}(1), \end{aligned}$$

and

$$\begin{aligned} \text{Var} &= \xi^2 \sigma^2 \frac{M}{N_1} [\underline{m}_1 - 2\lambda_s \underline{m}_1 \underline{m}_2 + \lambda_s^2 \underline{m}_1 \underline{m}'_2 - \lambda_t \underline{m}'_1 + 2\lambda_t \lambda_s \underline{m}_2 \underline{m}'_1 - \lambda_t \lambda_s^2 \underline{m}'_1 \underline{m}'_2] \\ &\quad + (1 - \xi)^2 \sigma^2 \frac{M}{N_2} [\underline{m}_2 - \lambda_s \underline{m}'_2] + o_{a.s.}(1), \end{aligned}$$

where  $\underline{m}'_1, \underline{m}'_2$  denote the derivatives evaluated at  $z = -\lambda_t$  and  $z = -\lambda_s$ , respectively:

$$\underline{m}'_1 = \frac{d}{dz} \underline{m}_1(z) \Big|_{z=-\lambda_t}, \underline{m}'_2 = \frac{d}{dz} \underline{m}_2(z) \Big|_{z=-\lambda_s}.$$

As previously noted, we do not restrict  $\xi$  to the interval  $[0, 1]$ . It has been shown in [Das & Sanghavi \(2023\)](#) that the optimal value of  $\xi$  may exceed 1. In Corollary 2 below, we present a toy example demonstrating that even when the input data across domains are i.i.d. and in the absence of model shift – i.e., with no domain shift – the limiting optimal value of  $\xi$  can be negative.

**Corollary 2.** Suppose the conditions in Corollary 1 hold. The limiting optimal value of  $\xi < 0$  if

$$\lambda_s \lambda_t \underline{m}_1 \text{SNR} - \frac{M}{N_2} > 0, \quad (10)$$

where  $\text{SNR} = \frac{\tilde{\sigma}^2}{\sigma^2} = \frac{\|\beta\|^2}{\sigma^2} + o_{a.s.}(1)$ .

**Remark 1.** We call the case  $\xi < 0$  **anti-learning against the teacher's supervision**, in contrast to  $\xi > 1$ , which [Das & Sanghavi \(2023\)](#) termed anti-learning the observed (possibly noisy) labels. This corollary provides insight into the selection of  $\xi$ : the sign of the limiting optimal value of  $\xi$  depends not only on parameters  $(\lambda_t, \lambda_s)$  but also on data-related factors (SNR, data dimension, and sample sizes of both domains).

### 3.3 RIDGELESS REGRESSION

In this section, we consider the minimum  $\ell_2$  norm least squares (ridgeless) regression estimator. Specifically, the teacher model is defined by

$$\beta_t = (\mathbf{X}_1 \mathbf{X}_1^\top)^+ \mathbf{X}_1 \mathbf{y}_1,$$

where  $(\mathbf{X}_1 \mathbf{X}_1^\top)^+$  denotes the Moore-Penrose inverse of  $\mathbf{X}_1 \mathbf{X}_1^\top$ . Similarly, the ridgeless estimator of  $\beta_s$  takes the form

$$\beta_s = (\mathbf{X}_2 \mathbf{X}_2^\top)^+ [\xi \mathbf{X}_2 \mathbf{X}_2^\top \beta_t + (1 - \xi) \mathbf{X}_2 \mathbf{y}_2].$$

**Theorem 3.** (1) Suppose  $\beta_1, \beta_2$  are deterministic, and Assumptions 1-2 hold. We further assume

$$\left| \frac{M}{N_i} - 1 \right| \geq \tau, \quad \tau \leq \sigma_{\min}(\Sigma_i) \leq \dots \leq \sigma_{\max}(\Sigma_i) \leq \tau^{-1}, \text{ for } i = 1, 2.$$

Define  $f(\lambda) = \widehat{\text{Bias}}$  and  $g(\lambda) = \widehat{\text{Var}}$ , with  $\lambda = \lambda_s = \lambda_t$ , where the expressions for  $\widehat{\text{Bias}}$  and  $\widehat{\text{Var}}$  are provided in equation 7 and equation 8, respectively. We have

$$\text{Bias} = f(0^+) + o_{a.s.}(1), \quad \text{Var} = g(0^+) + o_{a.s.}(1). \quad (11)$$

(2) Suppose  $\beta = \beta_1 = \beta_2$  are random and Assumptions 1-3 hold. Then, the estimated expressions in equation 11 still holds with  $f(\lambda) = \widehat{\text{Bias}}$  replaced by the  $\widehat{\text{Bias}}$  defined in Theorem 2.

If a matrix  $\mathbf{A}$  is nonsingular,  $\mathbf{A}^+ = \mathbf{A}^{-1}$ . The following corollary gives the characterization of  $\mathbf{ER}(\beta_s)$  in the under-parameterized setting.

**Corollary 3.** Suppose the conditions in Theorem 3 hold and  $\frac{M}{N_1}, \frac{M}{N_2} \leq 1 - \tau$ . The estimator for student model obtained by equation 1 is the averaging estimator:

$$\beta_s = \xi \beta_1^{\text{OLS}} + (1 - \xi) \beta_2^{\text{OLS}}, \text{ where } \beta_i^{\text{OLS}} = (\mathbf{X}_i \mathbf{X}_i^\top)^{-1} \mathbf{X}_i \mathbf{y}_i, i = 1, 2. \quad (12)$$

Adopting the notation  $\gamma = \beta_1 - \beta_2$  in Theorem 1, we have

$$\widehat{\text{Bias}} = \xi^2 \gamma^\top \Sigma_2 \gamma, \quad \widehat{\text{Var}} = (1 - \xi)^2 \sigma^2 \frac{M}{N_2 - M} + \xi^2 \sigma^2 \frac{1}{N_1 - M} \text{Tr} \Sigma_2 \Sigma_1^{-1}.$$

Based on the conclusions of Theorems 1-3, the high-dimensional asymptotic excess risk, regarded as a function of  $\xi$ , is a quadratic function. Given that the excess risk is non-negative, the quadratic function opens upwards. This observation is consistent with Pareek et al. (2024), where self-distillation is considered. Given a  $\xi \in \mathbb{R}$ , the gain of cross-domain KD is characterized by the reduction in excess risk,  $\mathbf{ER}_0 - \mathbf{ER}(\beta_s)$ .

**Proposition 1.** Under the conditions of Theorem 1 and Assumption A.1 for the deterministic case in Appendix B.7, there exists a value of  $\xi \in \mathbb{R}$  such that

$$\min_{\xi \in \mathbb{R}} (\mathbf{ER}(\beta_s) - \mathbf{ER}_0) < 0. \text{ a.s.} \quad (13)$$

Moreover, under the conditions of Theorem 1 and Assumption A.2 for the random case in Appendix B.7, the inequality 13 also holds.

**Remark 2.** This proposition shows that, even in the presence of a significant domain discrepancy, it is possible to find a value of  $\xi \in \mathbb{R}$  such that the student model outperforms the student-only baseline (i.e., training on the observed labels only). We provide further details in Appendix B.7, where we provide closed-form expressions for the optimal  $\xi^*$  under several common settings and demonstrate that covariate shift can, in some cases, be beneficial for KD.

### 3.4 NUMERICAL SIMULATIONS

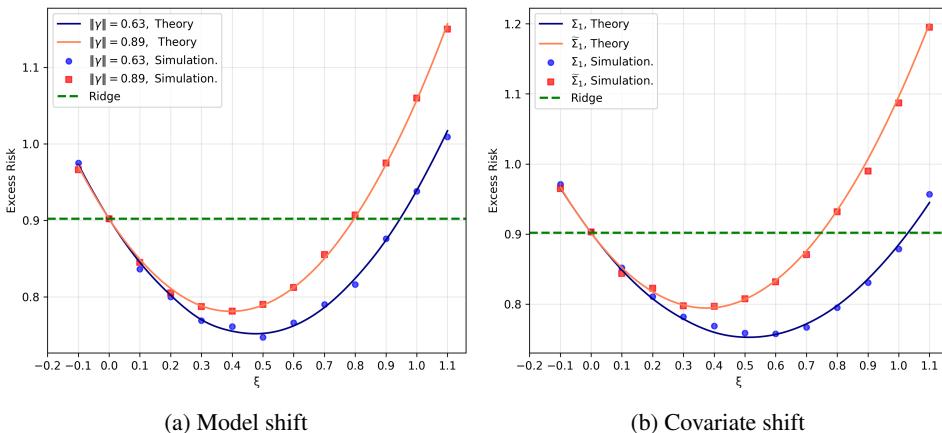


Figure 1: Student's excess risk in the presence of domain shift. Solid lines represent theoretical values, while scattered points denote simulation results (averaged over 100 trials). The dashed green line indicates the theoretical performance for student-only baseline, corresponding to ridge regression trained solely on the target domain data. (a) Settings:  $(\lambda_t, \lambda_s) = (0.1, 0.5)$ ,  $(M, N_1, N_2) = (400, 600, 200)$ ,  $\Sigma_1 = \Sigma_2 = \mathbf{I}_M$ . The vectors  $\beta_2 = (1, \dots, 1)^\top / \sqrt{M}$ ,  $\sigma^2 = 1$ . We label the case  $\|\gamma\| = 0.63$  as  $\gamma = -(2, \dots, 2, 0, \dots, 0)^\top / \sqrt{M}$  with the first  $M/10$  entries equal to  $-2/\sqrt{M}$ , and the case  $\|\gamma\| = 0.89$  with the first  $M/5$  entries equal to  $-2/\sqrt{M}$ . (b) Settings:  $(\lambda_t, \lambda_s) = (0.1, 0.5)$ ,  $\beta_1 = \beta_2 \sim \mathcal{N}(0, M^{-1} \mathbf{I}_M)$ ,  $(M, N_1, N_2) = (600, 200, 300)$ ,  $\Sigma_1 = 4\mathbf{I}_M$ ,  $\tilde{\Sigma}_1 = \text{diag}(d_1, \dots, d_M)$  with  $d_i = 0.64\delta(i \leq M/2) + 0.25\delta(M/2 < i \leq M)$ ,  $\sigma^2 = 1$ .

We plot the excess risk of the student model: (a) under model shift with identical covariate distributions, and (b) under covariate shift with identical parameter vectors, in Figure 1. All theoretical

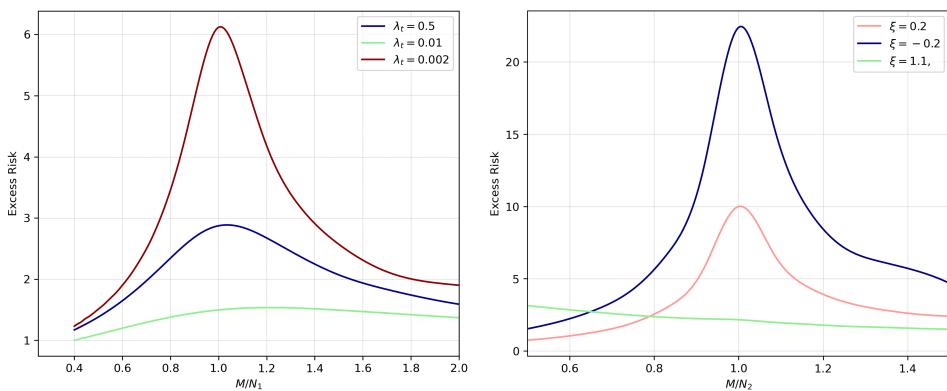
(a) Excess risk as a function of  $\frac{M}{N_1}$  for varying  $\lambda_t$  (b) Excess risk as a function of  $\frac{M}{N_2}$  for varying  $\xi$ 

Figure 2: Non-monotone student excess risk curves. We set  $\Sigma_2 = \mathbf{I}_M$ ,  $\Sigma_1 = \text{diag}(d_1, \dots, d_M)$  where  $d_i = 0.64\delta(i \leq \frac{M}{2}) + 0.25\delta(\frac{M}{2} < i \leq M)$ . (a) Results are shown for fixed  $M = 600$  and  $\lambda_s = 0.05$  with different  $N_1$ . (b) Results are shown for fixed  $N_2$  and  $(\lambda_t, \lambda_s) = (0.05, 0.001)$ , with varying  $M$ .

values of the Stieltjes transform presented in this paper are obtained by solving equation 6. Due to space limitations, the numerical validation of Corollary 3 is provided in Appendix C.2. Simulation results, averaged over 100 independent trials, show good agreement with the theoretical predictions.

Furthermore, we present numerical simulations of  $\mathbf{ER}(\beta_s)$  as a function of  $\lambda_s$  and  $\lambda_t$  for various values of  $\xi$ ; these results are included in Appendix C.3.

## 4 DOUBLE DESCENT OF THE EXCESS RISK

In this section, fixing  $\xi, \lambda_t$  and  $\lambda_s$ , we examine the excess risk as a function of the dimension  $M$  and the sample sizes  $N_1$  and  $N_2$ . We find that the student model exhibits the double descent phenomenon, characterized by a non-monotonic behavior of the excess risk as a function of the ratio of dimension-to-sample-size. This phenomenon is consistent with findings in various linear regression settings (Hastie et al., 2022; Nakkiran et al., 2021; Belkin et al., 2020; Moniri & Hassani, 2024), and has been previously observed by Moniri & Hassani (2025) in the special case of pure teacher supervision without domain shift, where the risk was studied as a function of  $\frac{M}{N_1}$ .

Using our theoretical predictions from Theorem 2, we plot the excess risk of the student model,  $\mathbf{ER} = \mathbf{ER}(\frac{M}{N_1})$ , as a function of  $\frac{M}{N_1}$  in Figure 2(a). The double descent phenomenon is evident for all three values of  $\lambda_t$ . As  $\lambda_t$  decreases, the peak of the risk curve shifts towards  $\frac{M}{N_1} = 1$ . In Figure 2(b), we plot  $\mathbf{ER} = \mathbf{ER}(\frac{M}{N_2})$  against  $\frac{M}{N_2}$ , while allowing  $\frac{M}{N_1}$  to vary simultaneously. We consider different values of  $\xi$  and observe that the double descent phenomenon is most pronounced in the regime of *anti-learning against the teacher's supervision* ( $\xi < 0$ ). In contrast, when  $\xi = 1.1$ , no double descent occurs within the ratio range  $[0.5, 1.5]$ .

## 5 EXTENSION

### 5.1 NONLINEAR CASE

Our theoretical results are initially established for linear models; however, we anticipate that they can be extended to more general settings. To explore this extension, we conduct numerical simulations specifically for nonlinear models here. We assume the source data  $\{(\mathbf{x}_j^{(1)}, y_j^{(1)})\}_{j=1}^{N_1}$  are generated i.i.d. according to  $y_j^{(1)} = f(\mathbf{x}_j^{(1)}) + \varepsilon_j^{(1)}$ , for  $1 \leq j \leq N_1$ . The target data  $\{(\mathbf{x}_j^{(2)}, y_j^{(2)})\}_{j=1}^{N_2}$  are generated according to  $y_j^{(2)} = \tilde{f}(\mathbf{x}_j^{(2)}) + \varepsilon_j^{(2)}$ , for  $1 \leq j \leq N_2$ . Suppose  $\mathbf{x}_j^{(1)} \sim D_1, 1 \leq j \leq N_1$

486 and  $\mathbf{x}_j^{(2)} \sim D_2, 1 \leq j \leq N_2$ . We refer to the case  $D_1 \neq D_2$  as a covariate shift, and the case where  
 487  $f \neq \tilde{f}$  as a model shift.  
 488

489 We consider learning the unknown function using a fully connected two-layer neural network with  
 490  $n$  hidden neurons:  $f_{\text{NN}}(\mathbf{x}) = \mathbf{a}^T \sigma(\mathbf{W}\mathbf{x})$ , where  $\mathbf{W} \in \mathbb{R}^{n \times M}$  is the weight matrix, and  $\sigma(\cdot)$  is  
 491 an activation function applied entrywise. When the random weight matrix  $\mathbf{W}$  is fixed and only  
 492 the second-layer weight  $\mathbf{a}$  is optimized, the model reduces to a kernel regression model, where the  
 493 kernel defined by  $\mathbf{x} \rightarrow \sigma(\mathbf{W}\mathbf{x})$  is referred to as the conjugate kernel (Neal, 2012). The teacher  
 494 model is given by  $f_{\text{NN}}^t(\mathbf{x}) = \mathbf{a}_t^T \sigma(\tilde{\mathbf{W}}_1 \mathbf{x})$ , with

$$495 \quad \mathbf{a}_t = \arg \min_{\mathbf{a}} \left\{ \frac{1}{N_1} \|\mathbf{y}_1 - \sigma(\mathbf{X}_1^T \tilde{\mathbf{W}}_1^T) \mathbf{a}\|^2 + \lambda_t \|\mathbf{a}\|^2 \right\}.$$

496 We use  $f_{\text{NN}}^t(\mathbf{x}_j^{(2)})$  together with the covariates  $\{\mathbf{x}_j^{(2)}\}_{j=1}^{N_2}$  to generate predictions  $\mathbf{y}_2^t$ . Then the  
 497 student model is finetuned on the target domain data and  $\mathbf{y}_2^t$ . The student model takes the form  
 498  $f_{\text{NN}}^s(\mathbf{x}) = \mathbf{a}_s^T \sigma(\mathbf{W}_1 \mathbf{x})$  with

$$501 \quad \mathbf{a}_s = \arg \min_{\mathbf{a}} \xi \left( \frac{1}{N_2} \|\mathbf{y}_2^t - \sigma(\mathbf{X}_2^T \mathbf{W}_1^T) \mathbf{a}\|^2 \right) + (1 - \xi) \left( \frac{1}{N_2} \|\mathbf{y}_2 - \sigma(\mathbf{X}_2^T \mathbf{W}_1^T) \mathbf{a}\|^2 \right) + \lambda_s \|\mathbf{a}\|^2.$$

502 We also examine a setting where the teacher model is a deeper neural network. Specifically, while  
 503 keeping the student model fixed, we let the teacher be a Four-layer fully connected network:

$$504 \quad f_{\text{NN}}^t = \mathbf{a}_t^T \sigma(\tilde{\mathbf{W}}_3 \sigma(\tilde{\mathbf{W}}_2 \sigma(\tilde{\mathbf{W}}_1 \mathbf{x}))),$$

505 where

$$506 \quad \mathbf{a}_t = \arg \min_{\mathbf{a}} \frac{1}{N_1} \|\mathbf{y}_1 - [\sigma(\tilde{\mathbf{W}}_3 \sigma(\tilde{\mathbf{W}}_2 \sigma(\tilde{\mathbf{W}}_1 \mathbf{X}_1)))^T \mathbf{a}\|^2 + \lambda_t \|\mathbf{a}\|^2.$$

507 We set  $D_1 = \mathcal{N}(0, 4\mathbf{I}_M)$  and  $D_2 = \mathcal{N}(0, \mathbf{I}_M)$ . Let  $f(\mathbf{x}) = (\beta^T \mathbf{x})^2 + 1$ ,  $\tilde{f}(\mathbf{x}) = (\beta^T \mathbf{x})^2$ .  
 508 Because  $D_1 \neq D_2$  and  $f \neq \tilde{f}$ , both covariate shift and model shift are present in this setting. More  
 509 details and the numerical results are provided in Appendix C.1.

## 514 5.2 DEPENDENCE BETWEEN DOMAINS

515 In this section, we consider two cases in which  $\mathbf{X}_1$  and  $\mathbf{X}_2$  are not fully independent. Case 1:  
 516 Assume  $\mathbf{X}_1$  exhibits weak dependence on  $\mathbf{X}_2$  in the following sense:  $\mathbf{X}_1 = \alpha \mathbf{X}_2 + \tilde{\mathbf{X}}_1$ , where  
 517  $\alpha \rightarrow 0$  as  $M \rightarrow \infty$ , and  $\tilde{\mathbf{X}}_1$  is independent of  $\mathbf{X}_2$  and takes the form  $\Sigma_1^{1/2} \mathbf{Z}_1$ . It is easy to  
 518 see  $\text{Cov}(\mathbf{x}_j^{(1)}, \mathbf{x}_j^{(2)}) = \text{Cov}(\mathbf{x}_j^{(2)}, \alpha \mathbf{x}_j^{(2)}) = \alpha \Sigma_2$ . Case 2: Suppose  $\mathbf{X}_1$  is a signal-plus-noise data  
 519 matrix:  $\mathbf{X}_1 = \mathbf{X}_2 + \mathbf{A}$ , where  $\mathbf{A}$  is a deterministic signal matrix with  $\|\mathbf{A}\| = o(\sqrt{M})$ . This  
 520 model captures realistic scenarios in domain adaptation where the source and target domains share  
 521 a common underlying data matrix, but differ by a small deterministic shift—such as a faint shared  
 522 signal across features in source domain. In Case 1, our theoretical analysis remains valid. For Case  
 523 2, we obtain a new limiting behavior; the theoretical results and technical details are provided in  
 524 Section B.9.

## 526 6 CONCLUSION

527 In this paper, we present a theoretical analysis of cross-domain KD for linear models using random  
 528 matrix theory. Through the bias-variance decomposition, we precisely characterize the asymptotic  
 529 expressions of excess risk for the student model in the high-dimensional setting. A surprising finding  
 530 is that when the imitation parameter  $\xi$  is allowed to take any real value, cross-domain KD may  
 531 outperform training solely on the target domain – even in the presence of significant discrepancies  
 532 between source and target domains. This highlights the potential of distillation to effectively transfer  
 533 knowledge across highly heterogeneous domains.

534 Our work also points to several promising directions for future research. Our theoretical analysis is  
 535 currently limited to linear models; extending it to more complex architectures, particularly nonlinear  
 536 models, would significantly broaden its applicability. Furthermore, while we observe the double  
 537 descent phenomenon using the established theoretical limits; a rigorous theoretical characterization  
 538 of this behavior in nonlinear models remains an important avenue for future investigation.

## 540 REFERENCES

541 Rishabh Agarwal, Nino Vieillard, Yongchao Zhou, Piotr Stanczyk, Sabela Ramos Garea, Matthieu  
 542 Geist, and Olivier Bachem. On-policy distillation of language models: Learning from self-  
 543 generated mistakes. In *The twelfth international conference on learning representations*, 2024.

544 Bloemendal Alex, László Erdős, Antti Knowles, Horng-Tzer Yau, and Jun Yin. Isotropic local laws  
 545 for sample covariance and generalized Wigner matrices. *Electronic Journal of Probability*, 19  
 546 (none):1 – 53, 2014. doi: 10.1214/EJP.v19-3054.

547 Taichi Asami, Ryo Masumura, Yoshikazu Yamaguchi, Hirokazu Masataki, and Yushi Aono. Domain  
 548 adaptation of dnn acoustic models using knowledge distillation. In *2017 IEEE International  
 549 Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5185–5189, 2017. doi:  
 550 10.1109/ICASSP.2017.7953145.

551 Zhidong Bai and Jack W Silverstein. *Spectral analysis of large dimensional random matrices*,  
 552 volume 20. Springer, 2010.

553 Mikhail Belkin, Daniel Hsu, and Ji Xu. Two models of double descent for weak features. *SIAM  
 554 Journal on Mathematics of Data Science*, 2(4):1167–1180, 2020. doi: 10.1137/20M1336072.

555 Kenneth Borup and Lars N Andersen. Even your teacher needs guidance: Ground-truth targets  
 556 dampen regularization imposed by self-distillation. *Advances in Neural Information Processing  
 557 Systems*, 34:5316–5327, 2021.

558 D. L. Burkholder. Distribution Function Inequalities for Martingales. *The Annals of Probability*, 1  
 559 (1):19 – 42, 1973. doi: 10.1214/aop/1176997023.

560 Collin Burns, Pavel Izmailov, Jan Hendrik Kirchner, Bowen Baker, Leo Gao, Leopold Aschenbren-  
 561 ner, Yining Chen, Adrien Ecoffet, Manas Joglekar, Jan Leike, Ilya Sutskever, and Jeffrey Wu.  
 562 Weak-to-strong generalization: Eliciting strong capabilities with weak supervision. In *Forty-first  
 563 International Conference on Machine Learning*, 2024.

564 Moses Charikar, Chirag Pabbaraju, and Kirankumar Shiragur. Quantifying the gain in weak-to-  
 565 strong generalization. In *The Thirty-eighth Annual Conference on Neural Information Processing  
 566 Systems*, 2024.

567 Romain Couillet and Zhenyu Liao. *Random matrix methods for machine learning*. Cambridge  
 568 University Press, 2022.

569 Rudrajit Das and Sujay Sanghavi. Understanding self-distillation in the presence of label noise. In  
 570 *International Conference on Machine Learning*, pp. 7102–7140. PMLR, 2023.

571 Pasan Dissanayake, Faisal Hamman, Barproda Halder, Ilia Sucholutsky, Qiuyi Zhang, and Sang-  
 572 hamitra Dutta. Quantifying knowledge distillation using partial information decomposition. In  
 573 *International Conference on Artificial Intelligence and Statistics*, pp. 4474–4482. PMLR, 2025.

574 Edgar Dobriban and Stefan Wager. High-dimensional asymptotics of prediction: Ridge regression  
 575 and classification. *The Annals of Statistics*, 46(1):247 – 279, 2018. doi: 10.1214/17-AOS1549.

576 Yijun Dong, Yicheng Li, Yunai Li, Jason D. Lee, and Qi Lei. Discrepancies are virtue: Weak-to-  
 577 strong generalization through lens of intrinsic dimension. In *Forty-second International Confer-  
 578 ence on Machine Learning*, 2025.

579 M Emrullah Ildiz, Halil Alperen Gozeten, Ege Onur Taga, Marco Mondelli, and Samet Oymak.  
 580 High-dimensional analysis of knowledge distillation: Weak-to-strong generalization and scaling  
 581 laws. In *13th International Conference on Learning Representations*, 2025.

582 William Fedus, Barret Zoph, and Noam Shazeer. Switch transformers: Scaling to trillion parameter  
 583 models with simple and efficient sparsity. *Journal of Machine Learning Research*, 23(120):1–39,  
 584 2022.

585 Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François  
 586 Laviolette, Mario March, and Victor Lempitsky. Domain-adversarial training of neural networks.  
 587 *Journal of machine learning research*, 17(59):1–35, 2016.

594 Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. Minilm: Knowledge distillation of large lan-  
 595 guage models. *arXiv preprint arXiv:2306.08543*, 2023.

596

597 Hravir Harutyunyan, Ankit Singh Rawat, Aditya Krishna Menon, Seungyeon Kim, and Sanjiv Ku-  
 598 mar. Supervision complexity and its role in knowledge distillation. In *The Eleventh International  
 599 Conference on Learning Representations*, 2023.

600 Trevor Hastie, Andrea Montanari, Saharon Rosset, and Ryan J. Tibshirani. Surprises in high-  
 601 dimensional ridgeless least squares interpolation. *The Annals of Statistics*, 50(2):949 – 986, 2022.  
 602 doi: 10.1214/21-AOS2133.

603

604 Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv  
 605 preprint arXiv:1503.02531*, 2015.

606

607 Hyeonsu Jeong and Hye Won Chung. Rethinking self-distillation: Label averaging and enhanced  
 608 soft label refinement with partial labels. In *The Thirteenth International Conference on Learning  
 609 Representations*, 2025.

610

611 Jogendra Nath Kundu, Nishank Lakkakula, and R Venkatesh Babu. Um-adapt: Unsupervised multi-  
 612 task adaptation using adversarial cross-task distillation. In *Proceedings of the IEEE/CVF interna-  
 613 tional conference on computer vision*, pp. 1436–1445, 2019.

614

615 Lujun Li, Yufan Bao, Peijie Dong, Chuanguang Yang, Anggeng Li, Wenhan Luo, Qifeng Liu, Wei  
 616 Xue, and Yike Guo. Detkds: Knowledge distillation search for object detectors. In *Forty-first  
 617 International Conference on Machine Learning*, 2024.

618

619 Wei Li, Kefeng Fan, and Huihua Yang. Teacher–student mutual learning for efficient source-free  
 620 unsupervised domain adaptation. *Knowledge-Based Systems*, 261:110204, 2023. ISSN 0950-  
 621 7051. doi: <https://doi.org/10.1016/j.knosys.2022.110204>.

622

623 Zeqin Lin and Guangming Pan. Eigenvector overlaps in large sample covariance matrices and non-  
 624 linear shrinkage estimators. *arXiv preprint arXiv:2404.18173*, 2024.

625

626 David Lopez-Paz, Léon Bottou, Bernhard Schölkopf, and Vladimir Vapnik. Unifying distillation  
 627 and privileged information. *arXiv preprint arXiv:1511.03643*, 2015.

628

629 Marko Medvedev, Kaifeng Lyu, Dingli Yu, Sanjeev Arora, Zhiyuan Li, and Nathan Srebro. Weak-  
 630 to-strong generalization even in random feature networks, provably. In *Forty-second International  
 631 Conference on Machine Learning*, 2025.

632

633 Aditya K Menon, Ankit Singh Rawat, Sashank Reddi, Seungyeon Kim, and Sanjiv Kumar. A  
 634 statistical perspective on distillation. In Marina Meila and Tong Zhang (eds.), *Proceedings of  
 635 the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine  
 636 Learning Research*, pp. 7632–7642. PMLR, 18–24 Jul 2021.

637

638 Victoria Mingote, Antonio Miguel, Dayana Ribas, Alfonso Ortega, and Eduardo Lleida. Knowledge  
 639 distillation and random erasing data augmentation for text-dependent speaker verification. In  
 640 *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing  
 641 (ICASSP)*, pp. 6824–6828. IEEE, 2020.

642

643 Behrad Moniri and Hamed Hassani. Asymptotics of linear regression with linearly dependent data.  
 644 *arXiv preprint arXiv:2412.03702*, 2024.

645

646 Behrad Moniri and Hamed Hassani. On the mechanisms of weak-to-strong generalization: A theo-  
 647 retical perspective. *arXiv preprint arXiv:2505.18346*, 2025.

648

649 Preetum Nakkiran, Prayaag Venkat, Sham M. Kakade, and Tengyu Ma. Optimal regularization can  
 650 mitigate double descent. In *International Conference on Learning Representations*, 2021.

651

652 Radford M Neal. *Bayesian learning for neural networks*, volume 118. Springer Science & Business  
 653 Media, 2012.

654

655 Junsoo Oh, Jerry Song, and Chulhee Yun. From linear to nonlinear: Provable weak-to-strong gen-  
 656 eralization through feature learning. In *High-dimensional Learning Dynamics 2025*, 2025.

648 Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong  
 649 Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to fol-  
 650 low instructions with human feedback. *Advances in neural information processing systems*, 35:  
 651 27730–27744, 2022.

652 Divyansh Pareek, Simon S Du, and Sewoong Oh. Understanding the gains from repeated self-  
 653 distillation. *Advances in Neural Information Processing Systems*, 37:7759–7796, 2024.

654

655 Mary Phuong and Christoph Lampert. Towards understanding knowledge distillation. In *Inter-  
 656 national conference on machine learning*, pp. 5142–5151. PMLR, 2019.

657

658 Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal,  
 659 Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual  
 660 models from natural language supervision. In *International conference on machine learning*, pp.  
 661 8748–8763. PmLR, 2021.

662 Changho Shin, John Cooper, and Frederic Sala. Weak-to-strong generalization through the data-  
 663 centric lens. In *The Thirteenth International Conference on Learning Representations*, 2025.

664

665 Jong-Chyi Su and Subhransu Maji. Adapting models to signal degradation using distillation. *arXiv  
 666 preprint arXiv:1604.00433*, 2016.

667

668 Jialiang Tang, Shuo Chen, Gang Niu, Hongyuan Zhu, Joey Tianyi Zhou, Chen Gong, and Masashi  
 669 Sugiyama. Direct distillation between different domains. In *Computer Vision – ECCV 2024*, pp.  
 154–172, Cham, 2025. Springer Nature Switzerland.

670

671 Colin Wei, Kendrick Shen, Yining Chen, and Tengyu Ma. Theoretical analysis of self-training  
 672 with deep networks on unlabeled data. In *International Conference on Learning Representations*,  
 673 2021.

674

675 David Xing Wu and Anant Sahai. Provable weak-to-strong generalization via benign overfitting. In  
 676 *The Thirteenth International Conference on Learning Representations*, 2025.

677

678 Qing Xu, Min Wu, Xiaoli Li, Kezhi Mao, and Zhenghua Chen. Reinforced cross-domain knowledge  
 679 distillation on time series data. In *The Thirty-eighth Annual Conference on Neural Information  
 680 Processing Systems*, 2024.

681

682 Chuanpeng Yang, Yao Zhu, Wang Lu, Yidong Wang, Qian Chen, Chenlong Gao, Bingjie Yan, and  
 683 Yiqiang Chen. Survey on knowledge distillation for large language models: methods, evaluation,  
 684 and application. *ACM Transactions on Intelligent Systems and Technology*, 2024.

685

686 Fan Yang, Hongyang R. Zhang, Sen Wu, Christopher Re, and Weijie J. Su. Precise high-dimensional  
 687 asymptotics for quantifying heterogeneous transfers. *Journal of Machine Learning Research*, 26  
 688 (113):1–88, 2025.

689

690 Linfeng Ye, Shayan Mohajer Hamidi, Renhao Tan, and EN-HUI YANG. Bayes conditional distribu-  
 691 tion estimation for knowledge distillation based on conditional mutual information. In *The Twelfth  
 692 International Conference on Learning Representations*, 2024. URL <https://openreview.net/forum?id=yV6wwEbtkR>.

693

694

695

696

697

698

699

700

701

702 A THE USE OF LARGE LANGUAGE MODELS (LLMs)  
703704 Qwen3 is used to polish the writing.  
705706 B PROOFS  
707708 B.1 BASIC TOOLS  
709710 Preliminary definitions and auxiliary lemmas used in the proofs of the main results are provided in  
711 this section.  
712713 **Lemma 3.** (Lemma B.26 in [Bai & Silverstein \(2010\)](#)) Let  $\mathbf{C}$  be an  $M \times M$  deterministic matrix  
714 and  $\mathbf{z} \in \mathbb{R}^M$  be a random vector of independent entries. Assume that  $\mathbb{E}x_i = 0$ ,  $\mathbb{E}|x_i|^2 = 1$ , and  
715  $\mathbb{E}|x_i|^\ell \leq C_\ell$ . Then for any  $\ell \geq 1$ ,

716 
$$\mathbb{E}|\mathbf{z}^\top \mathbf{C} \mathbf{z} - \text{Tr}\mathbf{C}|^\ell \leq c_\ell((C_4 \text{Tr}\mathbf{C}\mathbf{C}^*)^{\ell/2} + C_{2\ell} \text{Tr}(\mathbf{C}\mathbf{C}^*)^{\ell/2}),$$
  
717

718 where  $c_\ell$  is a constant depending on  $\ell$  only.  
719720 Before stating the subsequent results, it is convenient to introduce the notion of stochastic domi-  
721 nation.  
722723 **Definition 1.** Let  $\chi = \chi^{(p)}$ ,  $\zeta = \zeta^{(p)}$  be two families of  $p$ -dependent random variables. We say that  
724  $\chi$  is stochastically dominated by  $\zeta$  if for all small  $c > 0$  and large constant  $\ell > 0$ ,

725 
$$\mathbb{P}(|\chi^{(p)}| > p^c |\zeta^{(p)}|) \leq p^{-\ell}$$

726 for all large  $p$ . If  $\chi$  is stochastically dominated by  $\zeta$ , we use the notation  $\chi \prec \zeta$  or  $\chi = O_\prec(\zeta)$ . We  
727 say an event  $\mathcal{E}_p$  holds with high probability if

728 
$$\mathbb{P}(\mathcal{E}_p^C) \leq p^{-\ell} \text{ for any fixed } \ell > 0.$$
  
729

730 **Lemma 4.** (Lemma 22 in [Yang et al. \(2025\)](#)) Let  $\mathbf{Z}$  satisfies Assumption 1(a)-(b). Suppose  $\frac{M}{N} \leq$   
731  $1 - \tau$ . Then we have

732 
$$(\sqrt{M} - \sqrt{N})^2 + O_\prec(\sqrt{N}) \leq \lambda_{\min}(\mathbf{Z}\mathbf{Z}^\top) \leq \lambda_{\max}(\mathbf{Z}\mathbf{Z}^\top) \leq (\sqrt{M} + \sqrt{N})^2 + O_\prec(\sqrt{N}).$$
  
733

734 The upper bound on  $\lambda_{\max}(\mathbf{Z}\mathbf{Z}^\top)$  still holds without the assumption  $M/N \leq 1 - \tau$ .  
735736 **Lemma 5.** (Corollary 25 in [Yang et al. \(2025\)](#)) Suppose  $\varepsilon_1, \dots, \varepsilon_t$  are independent random vectors  
737 satisfying Assumption 1(c). Then, we have that for any deterministic vector  $\mathbf{v} \in \mathbb{R}^N$ ,

738 
$$|\mathbf{v}^\top \varepsilon_i| \prec \sigma \|\mathbf{v}\|, \quad i = 1, \dots, t,$$
  
739

740 and for any deterministic matrix  $\mathbf{B} \in \mathbb{R}^{N \times N}$ ,

741 
$$|\varepsilon_i^\top \mathbf{B} \varepsilon_j - \delta(i = j)\sigma^2 \text{Tr}\mathbf{B}| \prec \sigma^2 \|\mathbf{B}\|_{\text{F}}, \quad \text{for } i, j \in [t].$$
  
742

743 Moreover, for any deterministic vector  $\mathbf{v}$ , we have

744 
$$|\mathbf{v}^\top \varepsilon_i| \prec \sigma \|\mathbf{v}\|, \quad i \in [t].$$
  
745

746 **Definition 2.** Let  $\mathbf{A}_p, \mathbf{B}_p \in \mathbb{R}^{p \times p}$  be sequences of random or deterministic symmetric real matrices.  
747 We say  $\mathbf{A}_p, \mathbf{B}_p$  are equivalent, denoted by  $\mathbf{A}_p \asymp \mathbf{B}_p$ , if

748 
$$\frac{1}{p} \text{Tr}\mathbf{D}_p(\mathbf{A}_p - \mathbf{B}_p) = o_{a.s.}(1) \text{ and } \mathbf{u}^\top(\mathbf{A}_p - \mathbf{B}_p)\mathbf{v} = o_{a.s.}(1)$$
  
749

750 for any sequence of deterministic matrices  $\mathbf{D}_p$  and all deterministic vectors  $\mathbf{u}, \mathbf{v}$  such that

751 
$$\limsup_p \|\mathbf{D}_p\| < \infty, \limsup_p \max\{\|\mathbf{u}\|, \|\mathbf{v}\|\} < \infty.$$
  
752

753 **Lemma 6.** (1) (Theorem 2.6 in [Coullet & Liao \(2022\)](#)) We denote by  $\varrho$  the probability measure  
754 associated with  $m$  determined in Lemma 2. Let  $\mathbf{X} = \Sigma^{1/2} \mathbf{Z} \in \mathbb{R}^{M \times N}$ , where the entries of  $\mathbf{Z}$

756 are i.i.d. centered random variables with unit variance and finite  $8 + c$ -th moment, for any positive  
 757 constant  $c$ . Suppose Assumption 1(b) and Assumption 2 hold. Then for  $z \in \mathbb{C} \setminus \mathbb{R}^+$ , we have  
 758

$$759 \mathbf{Q}(z) \asymp \mathbf{\Pi}(z), \quad \underline{\mathbf{Q}}(z) \asymp m(z)\mathbf{I}_M, \quad (14)$$

760 where

$$763 \mathbf{Q}(z) = \left( \frac{\mathbf{X}\mathbf{X}^\top}{N} - z\mathbf{I}_M \right)^{-1}, \quad \underline{\mathbf{Q}}(z) = \left( \frac{\mathbf{X}^\top\mathbf{X}}{N} - z\mathbf{I}_M \right)^{-1}, \quad \mathbf{\Pi}(z) = -(z\mathbf{I}_M + zm\mathbf{\Sigma})^{-1}.$$

766 (2) When  $\frac{M}{N} < 1 - \tau$ , equation 14 still holds at  $z = 0$ .

768 **Proof:** Let  $\mathbf{u}, \mathbf{v}$  be two deterministic unit vectors and  $f_N(\lambda) = \mathbf{u}^\top \mathbf{Q}(-\lambda) \mathbf{v}$  for  $\lambda < 0$ . Since  
 769  $\lambda_{\min}(\frac{\mathbf{X}\mathbf{X}^\top}{N}) > \frac{1}{2}(1 - \sqrt{M/N})^2$  with high probability, we have  $|f_N(\lambda)| \leq \|\mathbf{Q}(-\lambda)\| \lesssim 1$ ,  
 770  $|f'_N(\lambda)| \leq \|\frac{d}{d\lambda} \mathbf{Q}(-\lambda)\| \lesssim 1$  with high probability. Therefore,  $\{f_N(\lambda)\}$  is equicontinuous with  
 771 high probability. By applying the Arzela-Ascoli theorem,  $f_N$  converges uniformly to its limit  
 772  $f(\lambda) = \mathbf{u}^\top \mathbf{\Pi}(-\lambda) \mathbf{v}$ . By the Moore-Osgood theorem, we can exchange limits and get  
 773

$$774 \lim_{N \rightarrow \infty} f_N(0) = \lim_{N \rightarrow \infty} \lim_{\lambda \rightarrow 0^-} f_N(\lambda) = \lim_{\lambda \rightarrow 0^-} \lim_{N \rightarrow \infty} f_N(\lambda) = \lim_{\lambda \rightarrow 0^-} \mathbf{u}^\top \mathbf{\Pi}(-\lambda) \mathbf{v} = \mathbf{u}^\top \mathbf{\Pi}(0) \mathbf{v}, \text{ a.s.},$$

776 where we use the fact that both  $\mathbf{Q}$  and  $\mathbf{\Pi}$  are analytic in an open neighborhood of 0 with high  
 777 probability. Similarly, we can derive  $\frac{1}{M} \text{Tr} \mathbf{A} \mathbf{Q}(0) = \frac{1}{M} \text{Tr} \mathbf{A} \mathbf{\Pi}(0) + o_{a.s.}(1)$ , which completes the  
 778 proof.

779 Lemma 6 shows that  $\mathbf{\Pi}(z)$  is a deterministic equivalent of  $\mathbf{Q}(z)$ . For technical reasons, we further  
 780 require the following result.

782 **Lemma 7.** Suppose the conditions in Lemma 6 hold.  $\mathbf{A}$  denotes a deterministic  $M \times M$  matrix  
 783 with bounded spectral norm. For any fixed complex numbers  $\tilde{z}_1, \tilde{z}_2 \in \mathbb{C} \setminus \mathbb{R}^+$ , we have for all  
 784 deterministic vectors  $\mathbf{u}, \mathbf{v}$ ,

$$785 \mathbf{u}^\top (\mathbf{Q}(\tilde{z}_1) \mathbf{A} \mathbf{Q}(\tilde{z}_2) - \mathbf{\Pi}(\tilde{z}_1) \mathcal{S}(\mathbf{A}) \mathbf{\Pi}(\tilde{z}_2)) \mathbf{v} = o_{a.s.}(\|\mathbf{u}\| \|\mathbf{v}\|), \quad (15)$$

787 where

$$789 \mathcal{S}(\mathbf{A}) = \mathbf{A} + \frac{\frac{1}{N} \text{Tr} \mathbf{\Sigma} \mathbf{\Pi}(\tilde{z}_1) \mathbf{A} \mathbf{\Pi}(\tilde{z}_2)}{\left(1 + \frac{1}{N} \text{Tr} \mathbf{\Sigma} \mathbf{\Pi}(\tilde{z}_2)\right) \left(1 + \frac{1}{N} \text{Tr} \mathbf{\Sigma} \mathbf{\Pi}(\tilde{z}_1)\right) - \frac{1}{N} \text{Tr} \mathbf{\Sigma} \mathbf{\Pi}(\tilde{z}_1) \mathbf{\Sigma} \mathbf{\Pi}(\tilde{z}_2)} \mathbf{\Sigma}. \quad (16)$$

792 Moreover, for any deterministic matrix  $\mathbf{C} \in \mathbb{R}^{M \times M}$  satisfying  $\|\mathbf{C}\| \leq C$  for some constant  $C$ , we  
 793 have

$$795 \frac{1}{M} \text{Tr} \mathbf{C} [\mathbf{Q}(\tilde{z}_1) \mathbf{A} \mathbf{Q}(\tilde{z}_2) - \mathbf{\Pi}_1(\tilde{z}_1) \mathcal{S}(\mathbf{A}) \mathbf{\Pi}_2(\tilde{z}_2)] = o_{a.s.}(1). \quad (17)$$

797 The proof of this lemma is deferred to Appendix B.8.

799 **Remark 3.** Lemma 7 provides the deterministic equivalent of  $\mathbf{Q}(\tilde{z}_1) \mathbf{A} \mathbf{Q}(\tilde{z}_2)$ . Lin & Pan (2024)  
 800 established the local laws for the  $\mathbf{Q}(\tilde{z}_1) \mathbf{A} \mathbf{Q}(\tilde{z}_2)$ . However, their results require  $\Re \tilde{z}_1, \Re \tilde{z}_2$  to be  
 801 sufficiently close to  $\text{supp}(\rho)$  and  $\Im \tilde{z}_1, \Im \tilde{z}_2$  to be bounded below by  $N^{-1+c}$ , where  $c$  is any fixed  
 802 constant. Lemma 7 extends the result to other regions.

803 **Remark 4.** To relax the moment assumption, we apply a standard truncation argument commonly  
 804 used in random matrix theory (e.g., Yang et al. (2025)). This approach allows us to employ Lemma 4  
 805 under the weaker finite  $(8 + c)$ -th moment condition, introducing only a negligible additional error  
 806 term that depends on  $M$  but does not affect the leading-order asymptotics of our results. Moreover,  
 807 a careful examination of the proofs shows that the same moment condition is also sufficient to es-  
 808 tablish Lemma 7. Consequently, all of our theoretical conclusions remain valid under this relaxed  
 809 assumption.

810 B.2 PROOF OF THEOREM 1  
811812 To simplify notation, we set  $z_1 = -\lambda_t, z_2 = -\lambda_s$ . Recalling equation 2 and equation 3, we get  
813

$$\begin{aligned}
814 \beta_s - \beta_2 &= \frac{1}{N_2} \mathbf{Q}_2 [\xi \mathbf{X}_2 \mathbf{X}_2^\top \beta_t + (1 - \xi) \mathbf{X}_2 (\mathbf{X}_2^\top \beta_2 + \boldsymbol{\varepsilon}_2)] - \beta_2 \\
815 &= \frac{1}{N_2} \mathbf{Q}_2 \left[ \frac{1}{N_1} \xi \mathbf{X}_2 \mathbf{X}_2^\top \mathbf{Q}_1 \mathbf{X}_1 (\mathbf{X}_1^\top \beta_1 + \boldsymbol{\varepsilon}_1) + (1 - \xi) \mathbf{X}_2 (\mathbf{X}_2^\top \beta_2 + \boldsymbol{\varepsilon}_2) \right] - \beta_2 \\
816 &= \xi [(\mathbf{I}_M + z_2 \mathbf{Q}_2) (\mathbf{I}_M + z_1 \mathbf{Q}_1) \beta_1 - \beta_2] + \underbrace{\xi \frac{1}{N_1} (\mathbf{I}_M + z_2 \mathbf{Q}_2) \mathbf{Q}_1 \mathbf{X}_1 \boldsymbol{\varepsilon}_1}_{a_5} \\
817 &\quad + (1 - \xi) [(\mathbf{I}_M + z_2 \mathbf{Q}_2) \beta_2 - \beta_2] + \underbrace{(1 - \xi) \frac{1}{N_2} \mathbf{Q}_2 \mathbf{X}_2 \boldsymbol{\varepsilon}_2}_{a_6} \\
818 &= \xi \boldsymbol{\gamma} + \underbrace{\xi z_2 \mathbf{Q}_2 \boldsymbol{\gamma}}_{a_1} + \underbrace{\xi z_1 \mathbf{Q}_1 \beta_1}_{a_2} + \underbrace{z_2 \mathbf{Q}_2 \beta_2}_{a_3} + \underbrace{\xi z_1 z_2 \mathbf{Q}_2 \mathbf{Q}_1 \beta_1}_{a_4} + \mathbf{a}_5 + \mathbf{a}_6.
\end{aligned} \tag{18}$$

821 By this, we decompose  $\mathbf{ER}(\beta_s)$  as follows:  
822

$$\begin{aligned}
823 \mathbf{ER}(\beta_s) &= (\beta_s - \beta_2)^\top \Sigma_2 (\beta_s - \beta_2) \\
824 &= \xi^2 \boldsymbol{\gamma}^\top \Sigma_2 \boldsymbol{\gamma} + 2 \sum_{i=1}^6 b_i + \sum_{i=1}^6 h_{ii} + \sum_{1 \leq i \neq j \leq 6} h_{ij},
\end{aligned} \tag{19}$$

833 where

$$b_i = \xi \boldsymbol{\gamma}^\top \Sigma_2 \mathbf{a}_i, \quad h_{ii} = \|\Sigma_2^{1/2} \mathbf{a}_i\|^2, \quad h_{ij} = \mathbf{a}_i^\top \Sigma_2 \mathbf{a}_j.$$

836 Next, we compute the limit of each term above.  
837838 Let  $n \in \mathbb{N}^+$ . According to the Definition 1 and the Borel–Cantelli lemma, we have  
839

$$\chi(n) = o_{a.s.}(1) \text{ if } \chi(n) \prec n^{-c}$$

840 for any constant  $c > 0$ . By this, the limits of  $b_1, b_2, b_3, b_4$  can be readily obtained using Lemma 6:  
841

$$\begin{aligned}
842 b_1 &= \xi^2 z_2 \boldsymbol{\gamma}^\top \Sigma_2 \boldsymbol{\gamma} = \xi^2 z_2 \boldsymbol{\gamma}^\top \Pi_2 \Sigma_2 \boldsymbol{\gamma} + o_{a.s.}(1), \\
843 b_2 &= \xi^2 z_1 \boldsymbol{\gamma}^\top \Sigma_2 \mathbf{Q}_1 \beta_1 = \xi^2 z_1 \boldsymbol{\gamma}^\top \Sigma_2 \Pi_1 \beta_1 + o_{a.s.}(1), \\
844 b_3 &= \xi z_2 \beta_2^\top \Sigma_2 \boldsymbol{\gamma} = \xi z_2 \beta_2^\top \Pi_2 \Sigma_2 \boldsymbol{\gamma} + o_{a.s.}(1), \\
845 b_4 &= \xi^2 z_1 z_2 \boldsymbol{\gamma}^\top \Sigma_2 \mathbf{Q}_2 \mathbf{Q}_1 \beta_1 = \xi^2 z_1 z_2 \boldsymbol{\gamma}^\top \Sigma_2 \Pi_2 \Pi_1 \beta_1 + o_{a.s.}(1),
\end{aligned}$$

847 where the last identity is due to  
848

$$849 \boldsymbol{\gamma}^\top \Sigma_2 [\mathbf{Q}_2 \mathbf{Q}_1 - \Pi_2 \Pi_1] \beta_1 = \boldsymbol{\gamma}^\top \Sigma_2 [(\mathbf{Q}_2 - \Pi_2) \mathbf{Q}_1 + \Pi_2 (\mathbf{Q}_1 - \Pi_1)] \beta_1 = o_{a.s.}(1).$$

851 We now consider the terms contributing to  $\mathbf{Var}$ . By Lemma 5 and the identity  
852

$$\frac{1}{N_i} \mathbf{Q}_i \mathbf{X}_i \mathbf{X}_i^\top = \mathbf{I}_M + z_1 \mathbf{Q}_i, \quad i = 1, 2. \tag{20}$$

855 we find that

$$\begin{aligned}
856 |h_{55} - \frac{\xi^2 \sigma^2}{N_1^2} \text{Tr}(\mathbf{I}_M + z_2 \mathbf{Q}_2) \Sigma_2 (\mathbf{I}_M + z_2 \mathbf{Q}_2) \mathbf{Q}_1 \mathbf{X}_1 \mathbf{X}_1^\top \mathbf{Q}_1| \\
857 &\prec \frac{1}{M^2} \|(\mathbf{I}_M + z_2 \mathbf{Q}_2) \Sigma_2 (\mathbf{I}_M + z_2 \mathbf{Q}_2) (\mathbf{Q}_1 + z_1 \mathbf{Q}_1^2)\|_{\mathbb{F}} \lesssim \frac{1}{\sqrt{M}}.
\end{aligned} \tag{21}$$

860 For any deterministic matrix  $\mathbf{C} \in \mathbb{R}^{M \times M}$  satisfying  $\|\mathbf{C}\|$  is bounded, and having the spectral de-  
861 composition  
862

$$\mathbf{C} = \sum_{i=1}^M \lambda_i \mathbf{u}_i \mathbf{v}_i^\top,$$

864 we have by Lemma 6 that  
865

$$\begin{aligned}
866 \quad \frac{1}{M} \text{Tr} \mathbf{C} \mathbf{Q}_1 \mathbf{Q}_2 &= \frac{1}{M} \sum_{i=1}^M \lambda_i \text{Tr} \mathbf{u}_i \mathbf{v}_i^\top \mathbf{Q}_1 \mathbf{Q}_2 \\
867 \\
868 &= \frac{1}{M} \sum_{i=1}^M \lambda_i \mathbf{v}_i^\top \mathbf{Q}_1 \mathbf{Q}_2 \mathbf{u}_i \\
869 \\
870 &= \frac{1}{M} \sum_{i=1}^M \lambda_i \mathbf{v}_i^\top \mathbf{\Pi}_1 \mathbf{\Pi}_2 \mathbf{u}_i + o_{a.s.}(1) \\
871 \\
872 &= \frac{1}{M} \text{Tr} \mathbf{C} \mathbf{\Pi}_1 \mathbf{\Pi}_2 + o_{a.s.}(1).
\end{aligned} \tag{22}$$

873 Similarly, by recalling the notation  $\mathbf{\Pi}_i \mathcal{S}_i(\mathbf{I}_M) \mathbf{\Pi}_i = \mathbf{\Pi}'_i$  for  $i = 1, 2$ , one may check by Lemma 7  
874 that

$$\begin{aligned}
875 \quad \frac{1}{M} \text{Tr} \mathbf{C} \mathbf{Q}_2 \mathbf{Q}_1^2 &= \frac{1}{M} \text{Tr} \mathbf{C} \mathbf{\Pi}_2 \mathbf{\Pi}'_1 + o_{a.s.}(1), \\
876 \quad \frac{1}{M} \text{Tr} \mathbf{C} \mathbf{Q}_1 \mathbf{Q}_2^2 &= \frac{1}{M} \text{Tr} \mathbf{C} \mathbf{\Pi}_1 \mathbf{\Pi}'_2, \\
877 \quad \frac{1}{M} \text{Tr} \mathbf{Q}_2 \mathbf{C} \mathbf{Q}_2 \mathbf{Q}_1^2 &= \frac{1}{M} \text{Tr} \mathbf{\Pi}_2 \mathcal{S}_2(\mathbf{C}) \mathbf{\Pi}_2 \mathbf{\Pi}'_1 + o_{a.s.}(1), \\
878 \quad \frac{1}{M} \text{Tr} \mathbf{Q}_1 \mathbf{C} \mathbf{Q}_1 \mathbf{Q}_2^2 &= \frac{1}{M} \text{Tr} \mathbf{\Pi}_1 \mathcal{S}_1(\mathbf{C}) \mathbf{\Pi}_1 \mathbf{\Pi}'_2 + o_{a.s.}(1).
\end{aligned} \tag{23}$$

879 Then by Lemma 6, Lemma 7, equation 20 and equation 21, for  $\xi \neq 0$ , we have

$$\begin{aligned}
880 \quad \frac{h_{55}}{\xi^2 \sigma^2} &= \frac{1}{N_1} \text{Tr} (\mathbf{I}_M + z_2 \mathbf{Q}_2) \mathbf{\Sigma}_2 (\mathbf{I}_M + z_2 \mathbf{Q}_2) (\mathbf{Q}_1 + z_1 \mathbf{Q}_1^2) \\
881 \\
882 &= \frac{1}{N_1} \text{Tr} [\mathbf{\Sigma}_2 + z_2 \mathbf{\Sigma}_2 \mathbf{Q}_2 + z_2 \mathbf{Q}_2 \mathbf{\Sigma}_2 + z_2^2 \mathbf{Q}_2 \mathbf{\Sigma}_2 \mathbf{Q}_2] [\mathbf{\Pi}_1 + z_1 \mathbf{\Pi}'_1] + o_{a.s.}(1) \\
883 \\
884 &= \frac{1}{N_1} \text{Tr} [\mathbf{\Sigma}_2 + 2z_2 \mathbf{\Sigma}_2 \mathbf{\Pi}_2 + z_2^2 \mathbf{\Pi}_2 \mathcal{S}_2(\mathbf{\Sigma}_2) \mathbf{\Pi}_2] [\mathbf{\Pi}_1 + z_1 \mathbf{\Pi}'_1] \\
885 \\
886 &+ o_{a.s.}(1).
\end{aligned}$$

887 Likewise, we have by Lemma 6 that

$$\begin{aligned}
888 \quad h_{66} &= (1 - \xi)^2 \sigma^2 \frac{1}{N_2} \text{Tr} \mathbf{\Sigma}_2 (\mathbf{I}_M + z_2 \mathbf{Q}_2) \mathbf{Q}_2 \\
889 \\
890 &= (1 - \xi)^2 \sigma^2 \frac{1}{N_2} \text{Tr} \mathbf{\Sigma}_2 (\mathbf{\Pi}_2 + z_2 \mathbf{\Pi}'_2) + o_{a.s.}(1).
\end{aligned} \tag{24}$$

900 Let  $d = \min \{ \text{dist}(z_1, \mathbb{R}^+), \text{dist}(z_2, \mathbb{R}^+) \}$ . According to Lemma 5 and the fact following from  
901 equation 20 that

$$\frac{1}{\sqrt{N_i}} \|\mathbf{Q}_i \mathbf{X}_i\| = \sqrt{\|\mathbf{Q}_i + z_i \mathbf{Q}_i^2\|} \leq \sqrt{d^{-1} + d^{-2}|z_i|} \lesssim 1,$$

902 one has for  $j = 1, 2, 3, 4$ ,

$$\begin{aligned}
903 \quad |h_{5j}| &= |h_{j5}| \prec \frac{\sigma}{M} \|\mathbf{X}_1^\top \mathbf{Q}_1 (\mathbf{I}_M + z_1 \mathbf{Q}_2) \mathbf{\Sigma}_2 \mathbf{a}_j\| \lesssim \frac{1}{\sqrt{M}}, \\
904 \quad |h_{6j}| &= |h_{j6}| \prec \frac{\sigma}{M} \|\mathbf{X}_2^\top \mathbf{Q}_2 \mathbf{\Sigma}_2 \mathbf{a}_j\| \lesssim \frac{1}{\sqrt{M}},
\end{aligned}$$

905 and

$$|b_5 + b_6| \prec \frac{1}{\sqrt{M}}.$$

Using Lemma 5 again, it can be shown that

$$\begin{aligned} |h_{65}| &= |h_{56}| \prec \frac{\sigma^2}{M^2} \|\mathbf{X}_2^\top \mathbf{Q}_2 \boldsymbol{\Sigma}_2 (\mathbf{I}_M + z_2 \mathbf{Q}_2) \mathbf{Q}_1 \mathbf{X}_1\|_{\mathsf{F}} \\ &\lesssim \frac{\sigma^2}{M} \sqrt{\frac{N_2 \|\mathbf{X}_2^\top \mathbf{Q}_2 \boldsymbol{\Sigma}_2 (\mathbf{I}_M + z_2 \mathbf{Q}_2) \mathbf{Q}_1 \mathbf{X}_1\|^2}{M^2}} \lesssim \frac{1}{\sqrt{M}}. \end{aligned}$$

Therefore, we get

$$b_5 + b_6 + h_{65} + h_{56} + \sum_{j=1}^4 (h_{5j} + h_{j5} + h_{6j} + h_{j6}) = o_{a.s.}(1).$$

We now turn to the terms  $h_{ii}$ ,  $i = 1, 2, 3, 4$ . By Lemma 7, we have

$$\begin{aligned}
h_{11} &= \xi^2 z_2^2 \gamma^\top \mathbf{Q}_2 \boldsymbol{\Sigma}_2 \mathbf{Q}_2 \gamma \\
&= \xi^2 z_2^2 \gamma^\top \boldsymbol{\Pi}_2 \mathcal{S}_2(\boldsymbol{\Sigma}_2) \boldsymbol{\Pi}_2 \gamma + o_{a.s.}(1), \\
h_{22} &= \xi^2 z_1^2 \beta_1^\top \mathbf{Q}_1 \boldsymbol{\Sigma}_2 \mathbf{Q}_1 \beta_1 \\
&= \xi^2 z_1^2 \beta_1^\top \boldsymbol{\Pi}_1 \mathcal{S}_1(\boldsymbol{\Sigma}_2) \boldsymbol{\Pi}_1 \beta_1 + o_{a.s.}(1), \\
h_{33} &= z_2^2 \beta_2^\top \mathbf{Q}_2 \boldsymbol{\Sigma}_2 \mathbf{Q}_2 \beta_2 \\
&= z_2^2 \beta_2^\top \boldsymbol{\Pi}_2 \mathcal{S}_2(\boldsymbol{\Sigma}_2) \boldsymbol{\Pi}_2 \beta_2 + o_{a.s.}(1), \\
h_{44} &= \xi^2 z_1^2 z_2^2 \beta_1^\top \mathbf{Q}_1 \mathbf{Q}_2 \boldsymbol{\Sigma}_2 \mathbf{Q}_2 \mathbf{Q}_1 \beta_1 \\
&= \xi^2 z_1^2 z_2^2 \mathbb{E} \beta_1^\top \mathbf{Q}_1 \boldsymbol{\Pi}_2 \mathcal{S}_2(\boldsymbol{\Sigma}_2) \boldsymbol{\Pi}_2 \mathbf{Q}_1 \beta_1 + o_{a.s.}(1) \\
&= \xi^2 z_1^2 z_2^2 \beta_1^\top \boldsymbol{\Pi}_1 \mathcal{S}_1(\boldsymbol{\Pi}_2 \mathcal{S}_2(\boldsymbol{\Sigma}_2) \boldsymbol{\Pi}_2) \boldsymbol{\Pi}_1 \beta_1 + o_{a.s.}(1).
\end{aligned}$$

Similarly, one can obtain the limits of the remaining terms in  $h_{ij}$ ,  $1 < i, j < 6$ :

$$\begin{aligned}
h_{12} &= h_{21} = \xi^2 z_1 z_2 \boldsymbol{\gamma}^\top \mathbf{Q}_2 \boldsymbol{\Sigma}_2 \mathbf{Q}_1 \boldsymbol{\beta}_1 \\
&= \xi^2 z_1 z_2 \boldsymbol{\gamma}^\top \boldsymbol{\Pi}_2 \boldsymbol{\Sigma}_2 \boldsymbol{\Pi}_1 \boldsymbol{\beta}_1 + o_{a.s.}(1), \\
h_{13} &= h_{31} = \xi z_2^2 \boldsymbol{\beta}_2^\top \mathbf{Q}_2 \boldsymbol{\Sigma}_2 \mathbf{Q}_2 \boldsymbol{\gamma} \\
&= \xi z_2^2 \boldsymbol{\beta}_2^\top \boldsymbol{\Pi}_2 \mathcal{S}_2(\boldsymbol{\Sigma}_2) \boldsymbol{\Pi}_2 \boldsymbol{\gamma} + o_{a.s.}(1), \\
h_{14} &= h_{41} = \xi^2 z_1 z_2^2 \boldsymbol{\gamma}^\top \mathbf{Q}_2 \boldsymbol{\Sigma}_2 \mathbf{Q}_2 \mathbf{Q}_1 \boldsymbol{\beta}_1 \\
&= \xi^2 z_1 z_2^2 \boldsymbol{\gamma}^\top \boldsymbol{\Pi}_2 \mathcal{S}_2(\boldsymbol{\Sigma}_2) \boldsymbol{\Pi}_2 \boldsymbol{\Pi}_1 \boldsymbol{\beta}_1 + o_{a.s.}(1), \\
h_{23} &= h_{32} = \xi z_1 z_2 \boldsymbol{\beta}_1^\top \mathbf{Q}_1 \boldsymbol{\Sigma}_2 \mathbf{Q}_2 \boldsymbol{\beta}_2 \\
&= \xi z_1 z_2 \boldsymbol{\beta}_1^\top \boldsymbol{\Pi}_1 \boldsymbol{\Sigma}_2 \boldsymbol{\Pi}_2 \boldsymbol{\beta}_2 + o_{a.s.}(1), \\
h_{24} &= h_{42} = \xi^2 z_1^2 z_2 \boldsymbol{\beta}_1^\top \mathbf{Q}_1 \boldsymbol{\Sigma}_2 \mathbf{Q}_2 \mathbf{Q}_1 \boldsymbol{\beta}_1 \\
&= \xi^2 z_1^2 z_2 \boldsymbol{\beta}_1^\top \mathbf{Q}_1 \boldsymbol{\Sigma}_2 \boldsymbol{\Pi}_2 \mathbf{Q}_1 \boldsymbol{\beta}_1 + o_{a.s.}(1) \\
&= \xi^2 z_1^2 z_2 \boldsymbol{\beta}_1^\top \boldsymbol{\Pi}_1 \mathcal{S}_1(\boldsymbol{\Sigma}_2 \boldsymbol{\Pi}_2) \boldsymbol{\Pi}_1 \boldsymbol{\beta}_1 + o_{a.s.}(1), \\
h_{34} &= h_{43} = \xi z_1 z_2^2 \boldsymbol{\beta}_2^\top \mathbf{Q}_2 \boldsymbol{\Sigma}_2 \mathbf{Q}_2 \mathbf{Q}_1 \boldsymbol{\beta}_1 \\
&= \xi z_1 z_2^2 \boldsymbol{\beta}_2^\top \boldsymbol{\Pi}_2 \mathcal{S}_2(\boldsymbol{\Sigma}_2) \boldsymbol{\Pi}_2 \boldsymbol{\Pi}_1 \boldsymbol{\beta}_1 + o_{a.s.}(1).
\end{aligned}$$

Combining the above estimates, we conclude the proof of Theorem 1.

### B.3 PROOF OF THEOREM 2

We use the same notation as in Appendix B.2. Note that  $\gamma = \beta_1 - \beta_2 = 0$ . Denoting

$$H = \xi z_1 Q_1 + z_2 Q_2 + \xi z_1 z_2 Q_2 Q_1, \quad (25)$$

by equation 18 we have

$$\beta_2 = \beta \equiv H\beta + a_5 + a_6$$

972 Hence, the excess risk becomes  
 973

$$\begin{aligned} \mathbf{ER}(\beta_s) &= \|\Sigma_2^{1/2}(\beta_s - \beta)\|^2 \\ &= \beta^\top \mathbf{H}^\top \Sigma_2 \mathbf{H} \beta + 2 \sum_{i=5,6} \beta^\top \mathbf{H}^\top \Sigma_2 \mathbf{a}_i + \sum_{i=5,6} h_{ii}. \end{aligned}$$

978 Using Lemma 5, by Assumption 3 we have  
 979

$$\beta^\top \mathbf{H}^\top \Sigma_2 \mathbf{H} \beta - \frac{\tilde{\sigma}^2}{M} \text{Tr} \mathbf{H}^\top \Sigma_2 \mathbf{H} \prec \frac{\tilde{\sigma}^2}{M} \|\mathbf{H}^\top \Sigma_2 \mathbf{H}\|_F \lesssim \frac{1}{\sqrt{M}}.$$

982 By equation 25, we have  
 983

$$\begin{aligned} \frac{1}{M} \text{Tr} \Sigma_2 \mathbf{H} \mathbf{H}^\top &= \frac{1}{M} \left[ \xi^2 z_1^2 \text{Tr} \Sigma_2 \mathbf{Q}_1^2 + \xi z_1 z_2 \text{Tr} \Sigma_2 [\mathbf{Q}_1 \mathbf{Q}_2 + \mathbf{Q}_2 \mathbf{Q}_1] \right. \\ &\quad + z_2^2 \text{Tr} \Sigma_2 \mathbf{Q}_2^2 + \xi^2 z_1^2 z_2 \text{Tr} \Sigma_2 [\mathbf{Q}_2 \mathbf{Q}_1^2 + \mathbf{Q}_1^2 \mathbf{Q}_2] \\ &\quad \left. + 2\xi z_1 z_2 \text{Tr} \mathbf{Q}_2 \Sigma_2 \mathbf{Q}_2 \mathbf{Q}_1 + \xi^2 z_1^2 z_2^2 \text{Tr} \mathbf{Q}_2 \Sigma_2 \mathbf{Q}_2 \mathbf{Q}_1^2 \right] \\ &= \sum_{i=1}^6 t_i, \end{aligned} \tag{26}$$

992 where  
 993

$$\begin{aligned} t_1 &= \frac{1}{M} \xi^2 z_1^2 \text{Tr} \Sigma_2 \mathbf{Q}_1^2, t_2 = \frac{2}{M} \xi z_1 z_2 \text{Tr} \mathbf{Q}_1 \mathbf{Q}_2, t_3 = \frac{z_2^2}{M} \text{Tr} \Sigma_2 \mathbf{Q}_2^2, \\ t_4 &= 2 \frac{\xi^2}{M} z_1^2 z_2 \text{Tr} \Sigma_2 \mathbf{Q}_2 \mathbf{Q}_1^2, t_5 = \frac{2\xi z_1 z_2^2}{M} \text{Tr} \mathbf{Q}_2 \Sigma_2 \mathbf{Q}_2 \mathbf{Q}_1, t_6 = \frac{\xi^2 z_1^2 z_2^2}{M} \text{Tr} \mathbf{Q}_2 \Sigma_2 \mathbf{Q}_2 \mathbf{Q}_1^2. \end{aligned}$$

998 We next consider the terms  $t_i, i = 1, \dots, 6$ . In the subsequent proof, we shall make use of Lemma 6,  
 999 Lemma 7 and the property that  $\Sigma_2 \Pi_2 = \Pi_2 \Sigma_2$ .  
 1000

By equation 22, we have  
 1001

$$t_2 = \frac{2\xi z_1 z_2}{M} \text{Tr} \Pi_1 \Pi_2 \Sigma_2 + o_{a.s.}(1).$$

$$t_1 = \xi^2 z_1^2 \frac{1}{M} \text{Tr} \Sigma_2 \Pi_1' + o_{a.s.}(1).$$

1007 The limits of  $t_3, t_4, t_5, t_6$  can be derived by equation 23:  
 1008

$$t_3 = \frac{z_2^2}{M} \text{Tr} \Sigma_2 \Pi_2' + o_{a.s.}(1),$$

$$t_4 = 2 \frac{\xi^2 z_1^2 z_2}{M} \text{Tr} \Sigma_2 \Pi_2 \Pi_1' + o_{a.s.}(1),$$

$$t_5 = \frac{2\xi z_1 z_2^2}{M} \text{Tr} \Pi_2 \mathcal{S}_2(\Sigma_2) \Pi_2 \Pi_1 + o_{a.s.}(1),$$

1016 and  
 1017

$$t_6 = \frac{\xi^2 z_1^2 z_2^2}{M} \text{Tr} \Pi_2 \mathcal{S}_2(\Sigma_2) \Pi_2 \Pi_1' + o_{a.s.}(1).$$

1019 Using Lemma 5, we find  
 1020

$$|\beta^\top \mathbf{H}^\top \Sigma_2 \mathbf{a}_5| \prec \frac{1}{M^{3/2}} \|\mathbf{H}^\top \Sigma_2 (\mathbf{I}_M + z_2 \mathbf{Q}_2) \mathbf{Q}_1 \mathbf{X}_1\|_F \lesssim \frac{1}{\sqrt{M}},$$

$$|\beta^\top \mathbf{H}^\top \Sigma_2 \mathbf{a}_6| \prec \frac{1}{M^{3/2}} \|\mathbf{Q}_2 \mathbf{X}_2\|_F \lesssim \frac{1}{\sqrt{M}}.$$

1025 Therefore, the terms  $\beta^\top \mathbf{H}^\top \Sigma_2 \mathbf{a}_i, i = 5, 6$  are ignorable. The proof is now complete.

1026 **B.4 PROOF OF COROLLARY 3**

1027 Letting  $\lambda_t = \lambda_s = 0$ , by equation 3, we obtain

$$\begin{aligned} \beta_s &= \xi \beta_1^{\text{OLS}} + (1 - \xi) \beta_2^{\text{OLS}} \\ &= \xi (\mathbf{X}_1 \mathbf{X}_1^T)^{-1} \mathbf{X}_1 (\mathbf{X}_1^T \beta_1 + \boldsymbol{\varepsilon}_1) + (1 - \xi) (\mathbf{X}_2 \mathbf{X}_2^T)^{-1} \mathbf{X}_2 (\mathbf{X}_2^T \beta_2 + \boldsymbol{\varepsilon}_2) \\ &= \xi \beta_1 + (1 - \xi) \beta_2 + \xi (\mathbf{X}_1 \mathbf{X}_1^T)^{-1} \mathbf{X}_1 \boldsymbol{\varepsilon}_1 + (1 - \xi) (\mathbf{X}_2 \mathbf{X}_2^T)^{-1} \mathbf{X}_2 \boldsymbol{\varepsilon}_2. \end{aligned}$$

1033 Plugging this into  $\mathbf{ER}(\beta_s)$ , one may obtain that

$$\begin{aligned} \mathbf{ER}(\beta_s) &= \|\Sigma_2^{1/2}(\beta_2 - \beta_s)\|^2 \\ &= \|\Sigma_2^{1/2}[\xi \boldsymbol{\gamma} + \xi (\mathbf{X}_1 \mathbf{X}_1^T)^{-1} \mathbf{X}_1 \boldsymbol{\varepsilon}_1 + (1 - \xi) (\mathbf{X}_2 \mathbf{X}_2^T)^{-1} \mathbf{X}_2 \boldsymbol{\varepsilon}_2]\|^2 \\ &= \widehat{\mathbf{Bias}} + h_1 + h_2 + 2h_3 + 2h_4 + 2h_5, \end{aligned}$$

1039 where

$$\begin{aligned} h_1 &= \xi^2 \boldsymbol{\varepsilon}_1^T \mathbf{X}_1^T (\mathbf{X}_1 \mathbf{X}_1^T)^{-1} \Sigma_2 (\mathbf{X}_1 \mathbf{X}_1^T)^{-1} \mathbf{X}_1 \boldsymbol{\varepsilon}_1, \\ h_2 &= (1 - \xi)^2 \boldsymbol{\varepsilon}_2^T \mathbf{X}_2^T (\mathbf{X}_2 \mathbf{X}_2^T)^{-1} \Sigma_2 (\mathbf{X}_2 \mathbf{X}_2^T)^{-1} \mathbf{X}_2 \boldsymbol{\varepsilon}_2, \\ h_3 &= \xi^2 \boldsymbol{\gamma}^T \Sigma_2 (\mathbf{X}_1 \mathbf{X}_1^T)^{-1} \mathbf{X}_1 \boldsymbol{\varepsilon}_1, \\ h_4 &= \xi(1 - \xi) \boldsymbol{\gamma}^T \Sigma_2 (\mathbf{X}_2 \mathbf{X}_2^T)^{-1} \mathbf{X}_2 \boldsymbol{\varepsilon}_2, \\ h_5 &= \xi(1 - \xi) \boldsymbol{\varepsilon}_1^T \mathbf{X}_1^T (\mathbf{X}_1 \mathbf{X}_1^T)^{-1} \Sigma_2 (\mathbf{X}_2 \mathbf{X}_2^T)^{-1} \mathbf{X}_2 \boldsymbol{\varepsilon}_2. \end{aligned}$$

1047 By Lemmas 4-5, we have with high probability,

$$\begin{aligned} |h_2 - (1 - \xi)^2 \sigma^2 \text{Tr}(\mathbf{X}_2 \mathbf{X}_2^T)^{-1} \Sigma_2| &= |h_2 - (1 - \xi)^2 \sigma^2 \text{Tr}(\mathbf{Z}_2 \mathbf{Z}_2^T)^T| \\ &\prec (1 - \xi)^2 \sigma^2 \|(\mathbf{Z}_2 \mathbf{Z}_2^T)^{-1}\|_{\text{F}} \\ &= (1 - \xi)^2 \sigma^2 \sqrt{\sum_{i=1}^M \lambda_i^{-2} (\mathbf{Z}_2 \mathbf{Z}_2^T)} \\ &\lesssim (1 - \xi)^2 \sigma^2 \frac{1}{\sqrt{M}} \text{Tr}(\mathbf{Z}_2 \mathbf{Z}_2^T)^{-1} \\ &\lesssim \frac{1}{\sqrt{M}}. \end{aligned} \tag{27}$$

1058 Lemma 6 implies that with high probability,

$$\text{Tr}(\mathbf{Z}_2 \mathbf{Z}_2^T)^{-1} = \frac{M}{N_2 - M} + o_{a.s.}(1).$$

1062 Combining this with equation 27, we obtain with high probability,

$$h_2 = (1 - \xi)^2 \sigma^2 \frac{M}{N_2 - M} (1 + o_{a.s.}(1)).$$

1065 Similarly, one may derive with high probability,

$$\begin{aligned} |h_5| &\prec \sigma^2 \|\mathbf{X}_1^T (\mathbf{X}_1 \mathbf{X}_1^T)^{-1} \Sigma_2 (\mathbf{X}_2 \mathbf{X}_2^T)^{-1} \mathbf{X}_2\|_{\text{F}} \\ &= \sigma^2 \sqrt{\text{Tr}(\mathbf{X}_1 \mathbf{X}_1^T)^{-1} \Sigma_2 (\mathbf{X}_1 \mathbf{X}_1^T)^{-1}} \lesssim \frac{1}{\sqrt{M}} \widehat{\text{Var}}. \end{aligned}$$

1070 Using Lemmas 4-5, the following estimate holds with high probability,

$$\begin{aligned} |h_1 - \xi^2 \sigma^2 \text{Tr}(\mathbf{X}_1 \mathbf{X}_1^T)^{-1} \Sigma_2| &\prec \|\Sigma_2 (\mathbf{X}_1 \mathbf{X}_1^T)^{-1}\|_{\text{F}} \\ &\lesssim \sqrt{M \|\Sigma_2\|^2 \|\Sigma_1\|^{-2} \|(\mathbf{Z}_1 \mathbf{Z}_1^T)^{-1}\|^2} \lesssim \frac{1}{\sqrt{M}}. \end{aligned}$$

1075 Then by Lemma 4 and Lemma 6, one has with high probability,

$$\begin{aligned} \text{Tr}(\mathbf{X}_1 \mathbf{X}_1^T)^{-1} \Sigma_2 &= \frac{1}{N_1} \text{Tr} \left( \frac{1}{N_1} \mathbf{Z}_1 \mathbf{Z}_1^T \right)^{-1} \Sigma_1^{-1/2} \Sigma_2 \Sigma_1^{-1/2} \\ &= \frac{1}{N_1} \frac{N_1}{N_1 - M} \text{Tr} \Sigma_2 \Sigma_1^{-1} + o_{a.s.}(1). \end{aligned}$$

1080 Therefore, for  $\xi \neq 0$ , we get with high probability,  
 1081

$$1082 \frac{1}{\xi^2 \sigma^2} h_1 = \text{Tr}(\mathbf{X}_1 \mathbf{X}_1^\top)^{-1} \Sigma_2 + o_{a.s.}(1) = \frac{1}{N_1 - M} \text{Tr} \Sigma_2 \Sigma_1^{-1} + o_{a.s.}(1).$$

1084 We note that  
 1085

$$\|\gamma\|^2 \gtrsim \widehat{\text{Bias}} = \xi^2 \|\Sigma_2^{1/2} \gamma\|^2 \gtrsim \lambda_{\min}(\Sigma_2) \|\gamma\|^2 \gtrsim \|\gamma\|^2.$$

1086 Since  $\sigma_M(\Sigma_1) \lesssim 1$ ,  $\sigma_1(\Sigma_2) \lesssim 1$ , it is easy to see  $\widehat{\text{Var}} \sim 1$ . Using Lemmas 4-5, we get with high  
 1087 probability  
 1088

$$\begin{aligned} 1089 |h_3| &\prec \xi^2 \sigma \|\gamma^\top \Sigma_2 (\mathbf{X}_1 \mathbf{X}_1^\top)^{-1} \mathbf{X}_1\| \\ 1090 &\leq \xi^2 \sigma \sqrt{\widehat{\text{Bias}}} \|\Sigma_2^{1/2}\| \|(\mathbf{X}_1 \mathbf{X}_1^\top)^{-1} \mathbf{X}_1\| \\ 1091 &\lesssim \frac{\sqrt{\widehat{\text{Bias}}}}{M^{1/4}} \frac{1}{M^{1/4}} \leq \frac{\widehat{\text{Bias}}}{\sqrt{M}} + \frac{1}{\sqrt{M}} \\ 1092 &\lesssim \frac{1}{\sqrt{M}} (\widehat{\text{Bias}} + \widehat{\text{Var}}). \\ 1093 \end{aligned}$$

1094 Similarly, we can estimate with high probability  
 1095

$$1096 |h_4| \prec \frac{1}{\sqrt{M}} (\widehat{\text{Bias}} + \widehat{\text{Var}}).$$

1097 Combining the above estimates on  $h_i, i = 1, 2, 3, 4, 5$ , the proof of Corollary 3 is completed.  
 1098

## 1102 B.5 PROOF OF THEOREM 3

1103 For simplicity, we present the proof only for deterministic  $\beta_1$  and  $\beta_2$ ; the extension to the random  
 1104 case follows by similar reasoning and is therefore omitted. Denote  $\mathbf{P}_{X_1}$  and  $\mathbf{P}_{X_2}$  by  
 1105

$$1106 \mathbf{P}_{X_1} = (\mathbf{X}_1 \mathbf{X}_1^\top)^+ \mathbf{X}_1 \mathbf{X}_1^\top, \quad \mathbf{P}_{X_2} = (\mathbf{X}_2 \mathbf{X}_2^\top)^+ \mathbf{X}_2 \mathbf{X}_2^\top.$$

1108 Note that for any rectangular matrix  $\mathbf{A}$  and compatible  $\mathbf{B}$ ,

$$1109 (\mathbf{A} \mathbf{A}^\top)^+ \mathbf{B} = \lim_{\lambda \rightarrow 0^+} (\mathbf{A} \mathbf{A}^\top + \lambda \mathbf{I}_M)^{-1} \mathbf{B}.$$

1110 We can apply this to  $\mathbf{A}_1 = \frac{1}{\sqrt{N_i}} \mathbf{X}_i$  for  $i = 1, 2$  and rewrite the bias as  
 1111

$$1112 \text{Bias} = \lim_{\lambda \rightarrow 0^+} f_M(\lambda),$$

1113 where  
 1114

$$1115 f_M(\lambda) = \xi^2 \gamma^\top \Sigma_2 \gamma + 2 \sum_{i=1}^6 b_i + \sum_{i=1}^4 h_{ii} + 2 \sum_{1 \leq i \neq j \leq 4} h_{ij},$$

1116 and all terms on the right-hand side are given in Section B.2, under the setting  $\lambda_t = \lambda_s = \lambda$ . It is  
 1117 straightforward to see that  $|f_M(\lambda)| \lesssim 1$ . Now we consider  $f'_M(\lambda)$ . Let  $\lambda_{\min}^+(\cdot)$  denote the smallest  
 1118 positive eigenvalue. Lemma 4 implies that for  $i = 1, 2$ ,  
 1119

$$1120 \frac{1}{N_i} \lambda_{\max}(\mathbf{X}_i \mathbf{X}_i^\top) \leq 2\sigma_1^i \left(1 + \sqrt{\frac{M}{N_i}}\right)^2, \quad \frac{1}{N_i} \lambda_{\min}^+(\mathbf{X}_i \mathbf{X}_i^\top) \geq \frac{1}{2} \sigma_M^i \left(1 - \sqrt{\frac{M}{N_i}}\right)^2, \text{ a.s.}$$

1121 Recall that  $\|\beta_1\|, \|\beta_2\| \leq c$ . Then, by equation 18, we have with high probability  
 1122

$$\begin{aligned} 1123 \left| \frac{d}{d\lambda} h_{22} \right| &= \left| \frac{d}{d\lambda} \xi^2 \lambda^2 \beta_1^\top \mathbf{Q}_1 \Sigma_2 \mathbf{Q}_1 \beta_1 \right| \\ 1124 &= 2\xi^2 \left| \lambda \beta_1^\top \mathbf{Q}_1^2 \frac{1}{N_1} \mathbf{X}_1 \mathbf{X}_1^\top \Sigma_2 \mathbf{Q}_1 \beta_1 \right| \\ 1125 &\leq 2\xi^2 \|\beta_1\|^2 \|\lambda \mathbf{Q}_1\| \|\Sigma_2\| \left\| \mathbf{Q}_1^2 \frac{1}{N_1} \mathbf{X}_1 \mathbf{X}_1^\top \right\| \\ 1126 &\leq C_\xi \frac{\lambda_{\max}(\mathbf{X}_1 \mathbf{X}_1^\top / N_1)}{(\lambda_{\min}(\mathbf{X}_1 \mathbf{X}_1^\top / N_1) + \lambda)^2} \lesssim 1. \end{aligned}$$

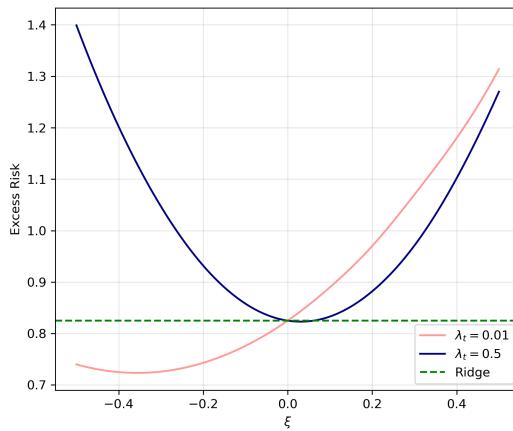


Figure 3: Theoretical excess risk for different  $\lambda_t$ . Settings:  $(M, N_1, N_2) = (200, 200, 600)$ ,  $\Sigma_1 = \Sigma_2 = \mathbf{I}_M$ ,  $\lambda_s = 0.5$ , SNR=4,  $\beta_1 = \beta_2 \sim \mathcal{N}(0, \frac{4}{M})$ ,  $\sigma^2 = 1$ .

The remaining terms in  $f'_M(\lambda)$  can be bounded in a similar manner, and hence  $|f'_M(\lambda)| \lesssim 1$  almost surely. Therefore,  $\{f_M(x)\}$  is equicontinuous almost surely. By the Arzela-Ascoli theorem,  $f_M$  converges uniformly to its limit  $f$ , almost surely. By the Moore-Osgood theorem, we can exchange limits and get,

$$\lim_{M \rightarrow \infty} \lim_{\lambda \rightarrow 0^+} f_M(\lambda) = \lim_{\lambda \rightarrow 0^+} \lim_{M \rightarrow \infty} f_M(\lambda) = f(0^+), \quad a.s.$$

Similarly, letting  $g_M(\lambda) = h_{55} + h_{66}$  with  $h_{55}$  and  $h_{66}$  as defined in Section B.2 under the setting  $\lambda_t = \lambda_s = \lambda$ , we get  $g_M$  converges uniformly to its limit,  $g$ , and

$$\lim_{M \rightarrow \infty} \lim_{\lambda \rightarrow 0^+} g_M(\lambda) = \lim_{\lambda \rightarrow 0^+} \lim_{M \rightarrow \infty} g_M(\lambda) = g(0^+), \quad a.s.$$

By Theorem 1,  $f = \widehat{\text{Bias}}$  and  $g = \widehat{\text{Var}}$  under the setting  $\lambda = \lambda_t = \lambda_s$ . It is easy to verify that the remaining terms appearing in 19 are asymptotically negligible, and that  $f, g$  are right-continuous. The proof is completed.

## B.6 PROOF OF COROLLARY 2

The Stieltjes transform of the Marchenko-Pastur distribution is given by

$$\underline{m}_i(z) = \int \frac{d\varrho_{MP,i}(x)}{x - z} = \frac{\left(1 - \frac{M}{N_i}\right) - z - \sqrt{\left(z - 1 - \frac{M}{N_i}\right)^2 - 4\frac{M}{N_i}}}{2\frac{M}{N_i}z}. \quad (28)$$

We take the derivative of  $\mathbf{ER}(\beta_s)$  with respect to  $\xi$ , and evaluate it at  $\xi = 0$ :

$$\frac{\partial}{\partial \xi} \mathbf{ER}(\beta_s) \Big|_{\xi=0} = \sigma^2 \left( \lambda_t \lambda_s \underline{m}_1 \text{SNR} - \frac{M}{N_2} \right) \frac{d}{dz} (z \underline{m}_2(z)) \Big|_{z=-\lambda_s} + o_{a.s.}(1).$$

Since

$$z \underline{m}_2(z) = \int \frac{z}{x - z} d\varrho_{MP,2}(x) = -1 + \int \frac{x}{x - z} d\varrho_{MP,2}(x),$$

we hence get that

$$\frac{d}{dz} (z \underline{m}_2(z)) \Big|_{z=-\lambda_s} > 0. \quad (29)$$

Therefore,  $\frac{\partial}{\partial \xi} \mathbf{ER}(\beta_s) \Big|_{\xi=0}$  and  $\lambda_t \lambda_s \text{SNR} - \frac{M}{N_2}$  share the same sign almost surely. That is, the limiting optimal value of  $\xi$  is negative when equation 10 holds, which establishes Corollary 2.

We provide an example in Figure 3 to illustrate this corollary. Specifically, when  $\lambda_t = 0.5$ , the limiting optimal value of  $\xi$  is positive, whereas when  $\lambda_t = 0.01$ , it becomes negative.

1188 B.7 DETAILS FOR PROPOSITION 1 AND REMARK 2  
11891190 Recall that  $\mathbf{ER}_0$  denotes the excess risk of the ridge regression model trained solely on the target  
1191 domain data.1192 **Assumption A.1.** When  $\beta_1, \beta_2$  are deterministic, we assume that :

1193  
1194 
$$\left| \beta_1^\top \lambda_t \lambda_s [\mathbf{\Pi}_1 \Sigma_2 \mathbf{\Pi}_2 - \lambda_s \mathbf{\Pi}_1 \mathbf{\Pi}_2 \mathcal{S}_2(\Sigma_2) \mathbf{\Pi}_2] \beta_2 - \frac{\sigma^2}{N_2} \text{Tr}[\Sigma_2(\mathbf{\Pi}_2 - \lambda_s \mathbf{\Pi}'_2)] \right. \\ 1195 \left. + \beta_2^\top [\lambda_s^2 \mathbf{\Pi}_2 \mathcal{S}_2(\Sigma_2) \mathbf{\Pi}_2 - \lambda_s \mathbf{\Pi}_2 \Sigma_2] \gamma \right| > c,$$
  
1196  
1197

1198 where  $c$  is a positive constant.1199 **Assumption A.2.** If  $\beta = \beta_1 = \beta_2$  is random, we assume that

1200  
1201 
$$\left| \frac{\tilde{\sigma}^2}{M} \left[ \lambda_t \lambda_s \text{Tr}[\mathbf{\Pi}_1 \mathbf{\Pi}_2 \Sigma_2] - \lambda_t \lambda_s^2 \text{Tr}[\mathbf{\Pi}_2 \mathcal{S}_2(\Sigma_2) \mathbf{\Pi}_2 \mathbf{\Pi}_1] \right] - \frac{\sigma^2}{N_2} \text{Tr}[\Sigma_2(\mathbf{\Pi}_2 - \lambda_s \mathbf{\Pi}'_2)] \right| > c,$$
  
1202

1203 where  $c$  is a positive constant.1204 **Proof of Proposition 1:** (i) Suppose the conditions in Theorem 1 hold. Note that

1205  
1206 
$$\left| \frac{\partial}{\partial \xi} \mathbf{ER}(\beta_s) \right|_{\xi=0} = 2\beta_1^\top [\lambda_t \lambda_s \mathbf{\Pi}_1 \Sigma_2 \mathbf{\Pi}_2 - \lambda_t \lambda_s^2 \mathbf{\Pi}_1 \mathbf{\Pi}_2 \mathcal{S}_2(\Sigma_2) \mathbf{\Pi}_2] \beta_2 \\ 1207 + 2\beta_2^\top [\lambda_s^2 \mathbf{\Pi}_2 \mathcal{S}_2(\Sigma_2) \mathbf{\Pi}_2 - \lambda_s \mathbf{\Pi}_2 \Sigma_2] \gamma - \frac{2\sigma^2}{N_2} \text{Tr}[\Sigma_2(\mathbf{\Pi}_2 - \lambda_s \mathbf{\Pi}'_2)] + o_{a.s.}(1).$$
  
1208

1209 Under the conditions in Theorem 1 and Assumption A.1, the asymptotic excess risk is a quadratic  
1210 function whose minimizer is bounded away from 0. Therefore,  $\min_\xi \mathbf{ER}(\beta_s)$  is strictly less than  
1211  $\mathbf{ER}_0$  almost surely.1212 (ii) Similarly, suppose Assumption A.2, under the conditions of Theorem 2, the inequality equation  
1213 13 holds by noticing that

1214  
1215 
$$\left| \frac{\partial}{\partial \xi} \mathbf{ER}(\beta_s) \right|_{\xi=0} = \frac{2\tilde{\sigma}^2}{M} \left[ \lambda_t \lambda_s \text{Tr}[\mathbf{\Pi}_1 \mathbf{\Pi}_2 \Sigma_2] - \lambda_t \lambda_s^2 \text{Tr}[\mathbf{\Pi}_2 \mathcal{S}_2(\Sigma_2) \mathbf{\Pi}_2 \mathbf{\Pi}_1] \right] - \frac{2\sigma^2}{N_2} \text{Tr}[\Sigma_2(\mathbf{\Pi}_2 - \lambda_s \mathbf{\Pi}'_2)] \\ 1216 + o_{a.s.}(1).$$
  
1217

1218 **Further discussion on Remark 2:** To clarify the dependence of Assumption A.1 on the geometry  
1219 of  $\Sigma_1, \Sigma_2, \beta_1, \beta_2$  and the noise strength  $\sigma^2$ , we consider a simple example in which  $\Sigma_2 = \mathbf{I}_M$ .  
1220 Then we have

1221  
1222 
$$\left| \frac{\partial}{\partial \xi} \mathbf{ER}(\beta_s) \right|_{\xi=0} = \lambda_s \lambda_t (\underline{m}_2 - \lambda_s \underline{m}'_2) \beta_1^\top \mathbf{\Pi}_1 \beta_2 - \frac{\sigma^2 M}{N_2} (\underline{m}_2 - \lambda_s \underline{m}'_2) - \lambda_s (\underline{m}_2 - \lambda_s \underline{m}'_2) \beta_2^\top \gamma \\ 1223 + o_{a.s.}(1) \\ 1224 = \underbrace{\left( \lambda_s \lambda_t \beta_1^\top \mathbf{\Pi}_1 \beta_2 - \frac{\sigma^2 M}{N_2} - \lambda_s \beta_2^\top \gamma \right)}_e \frac{d}{dz} (\underline{m}_2(z)) \Big|_{z=-\lambda_s} + o_{a.s.}(1),$$
  
1225  
1226  
1227  
1228  
1229  
1230  
1231  
1232  
1233  
1234  
1235  
1236  
1237  
1238  
1239  
1240  
1241

(30)

where  $\underline{m}_2(z)$  is defined in equation 28. Recalling equation 29, we have  $\left| \frac{\partial}{\partial \xi} \mathbf{ER}(\beta_s) \right|_{\xi=0} > c$  if  
1233  $|e| > C$  for some constant  $C$ . Below, we discuss two cases, when  $\Sigma_1 = \mathbf{I}_M$  and when  $\Sigma_1 \neq \mathbf{I}_M$ :1234 •  $\Sigma_1 = \mathbf{I}_M$ . The term  $e$  becomes

1235  
1236 
$$e = \lambda_s \lambda_t \underline{m}_1 \beta_1^\top \beta_2 - \frac{\sigma^2 M}{N_2} - \lambda_s \beta_2^\top (\beta_1 - \beta_2).$$
  
1237

1238 Recall that the limiting ridge risk is minimized at  $\lambda_s^* = \frac{\sigma^2 M}{N_2 \|\beta_2\|^2}$ , with asymptotic excess  
1239 risk  $\sigma^2 \frac{M}{N_2} \underline{m}_2(-\lambda_s^*)$  (Hastie et al., 2022). Taking  $\lambda_s = \lambda_s^*$ , we have  
1240

1241 
$$e = \lambda_s^* (\lambda_t \underline{m}_1 - 1) \beta_1^\top \beta_2.$$

1242 Note that

1243 
$$\lambda_t \underline{m}_1 - 1 = - \int \frac{x}{x + \lambda_t} d\varrho_{\text{MP},1}(x) < 0.$$
 1244

1245 Therefore, in a small neighborhood containing 0,  $\mathbf{ER}(\beta_s)$  is monotonic in  $\xi$ , indicating  
1246 that the teacher's supervision is helpful – even outperforming the optimal ridge regression  
1247 – provided that  $\beta_1$  and  $\beta_2$  are not asymptotically orthogonal.1248 •  $\Sigma_1 \neq \mathbf{I}_M$ . By taking  $\lambda_s = \lambda_s^*$ ,  $e$  becomes

1249 
$$1250 e = \lambda_s^* (\lambda_t \beta_1^\top \Pi_1 \beta_2 - \beta_2^\top \beta_1) = -\lambda_s^* \sum_{i=1}^M \frac{m_1 \sigma_i}{1 + m_1 \sigma_i} \beta_1^\top \mathbf{u}_i \mathbf{u}_i^\top \beta_2, \quad (31)$$
 1251

1252 where  $m_1$  is determined by Lemma 2 and  $\Sigma_1 = \sum_{i=1}^M \sigma_i \mathbf{u}_i \mathbf{u}_i^\top$  represents the spectral  
1253 decomposition of  $\Sigma_1$ . By equation 31, the alignment of  $\beta_i$  ( $i = 1, 2$ ) with the eigenvectors  
1254 of  $\Sigma_1$  determines whether Assumption A.1 is satisfied. Therefore, given  $\lambda_s = \lambda_s^*$ , under  
1255 the “help” of covariate shift, even if  $\beta_1^\top \beta_2 = 0$ , it may still be possible to find a  $\xi$  such  
1256 that  $\mathbf{ER}(\beta_s) < \mathbf{ER}_0$ , a.s. By comparing with the case where  $\Sigma_1 = \mathbf{I}_M$ , we find that the  
1257 presence of covariate shift can, in some cases, be beneficial.1258  
1259 Note that, up to asymptotically negligible terms,  $\mathbf{ER}$  can be expressed as a quadratic function of  
1260  $\xi : \mathbf{ER}(\beta_s, \xi) = A\xi^2 + B\xi + C$ . Below we provide closed-form expressions for the asymptotic  
1261 optimal  $\xi^* = -\frac{B}{2A}$  under several common settings.1262  
1263 (1) When  $\gamma = \beta_1 - \beta_2$ ,  $\frac{M}{N_1}, \frac{M}{N_2} < (1 + \tau)^{-1}$ ,

1264 
$$\xi^* = \left( \gamma^\top \Sigma_2 \gamma + \sigma^2 \frac{M}{N_2 - M} + \frac{\sigma^2}{N_1 - M} \text{Tr} \Sigma_2 \Sigma_1^{-1} \right)^{-1} \frac{\sigma^2 M}{N_2 - M} \in (0, 1).$$
 1265

1266  
1267 (2) When  $\beta = \beta_1 = \beta_2$  is random, and  $\Sigma_1 = \Sigma_2 = \mathbf{I}_M$ ,

1268 
$$\xi^* = \frac{\frac{M}{N_2} (\underline{m}_2 - \lambda_s \underline{m}'_2) - \frac{\tilde{\sigma}^2}{\sigma^2} \lambda_t \lambda_s (\underline{m}_1 \underline{m}_2 - \lambda_s \underline{m}'_2 \underline{m}_1)}{A_1 + A_2 + A_3},$$
 1269

1270 where

1271 
$$A_1 = \frac{\tilde{\sigma}^2}{\sigma^2} (\lambda_t^2 \underline{m}'_1 - 2\lambda_t^2 \lambda_s \underline{m}_2 \underline{m}'_1 + \lambda_t^2 \lambda_s^2 \underline{m}'_1 \underline{m}'_2),$$
 1272

1273 
$$A_2 = \frac{M}{N_1} (\underline{m}_1 - 2\lambda_s \underline{m}_1 \underline{m}_2 + \lambda_s^2 \underline{m}_1 \underline{m}'_2 - \lambda_t \underline{m}'_1 + 2\lambda_t \lambda_s \underline{m}_2 \underline{m}'_1 - \lambda_t \lambda_s^2 \underline{m}'_1 \underline{m}'_2),$$
 1274

1275 
$$A_3 = \frac{M}{N_2} (\underline{m}_2 - \lambda_s \underline{m}'_2).$$
 1276

1277  
1278 (3) When  $\beta = \beta_1 = \beta_2$ ,  $\Sigma_2 = \mathbf{I}_M$ ,

1279 
$$\xi^* = \frac{\beta^\top [\lambda_t \lambda_s^2 \underline{m}'_2 \Pi_1 - \lambda_t \lambda_s \underline{m}_2 \Pi_1] \beta + \frac{\sigma^2 M}{N_2} (\underline{m}'_2 - \lambda_s \underline{m}'_2)}{\beta^\top \lambda_t^2 [1 + \lambda_s^2 \underline{m}'_2 - 2\lambda_s \underline{m}_2] \Pi'_1 \beta + \frac{\sigma^2 (1 - 2\lambda_s \underline{m}_2 + \lambda_s^2 \underline{m}'_2)}{N_1} \text{Tr} [\Pi_1 - \lambda_t \Pi'_1] + \frac{\sigma^2 M}{N_2} (\underline{m}_2 - \lambda_s \underline{m}'_2)}.$$
 1280

1281  
1282 B.8 PROOF OF LEMMA 71283 The following result, which is an immediate consequence of Lemma 2, will be used in the proof  
1284 below:

1285 
$$-zm = \left( 1 + \frac{1}{N} \text{Tr} \Sigma \Pi(z) \right)^{-1}. \quad (32)$$
 1286

1287 We abuse notation by writing  $z_1$  and  $z_2$  for  $\tilde{z}_1$  and  $\tilde{z}_2$ , respectively, whenever there is no risk of  
1288 ambiguity. Without loss of generality, we assume  $\|\mathbf{u}\| = \|\mathbf{v}\| = 1$  and  $z_1, z_2$  lie on the negative real  
1289 axis, as the other cases follow by analogous arguments.

Using standard techniques of martingale decomposition (see, e.g., [Bai & Silverstein \(2010\)](#)), we can prove the almost sure convergence of the random part:

$$\mathbf{u}^\top \mathbf{Q}(z_1) \mathbf{A} \mathbf{Q}(z_2) \mathbf{v} = \mathbf{u}^\top \mathbb{E} \mathbf{Q}(z_1) \mathbf{A} \mathbf{Q}(z_2) \mathbf{v} + o_{a.s.}(1). \quad (33)$$

Therefore, it suffices to consider the term  $\mathbf{u}^\top \mathbb{E} \mathbf{Q}(z_1) \mathbf{A} \mathbf{Q}(z_2) \mathbf{v}$ . Let  $\sigma_1 \geq \dots \geq \sigma_M$  denote the eigenvalues of  $\Sigma$ . For the sequence of deterministic matrices, we denote  $\mathbf{A}_M = o(1)$  if  $\|\mathbf{A}_M\| \rightarrow 0$ . Since

$$\mathbf{Q}(z_1) \mathbf{A} \mathbf{Q}(z_2) = \mathbf{Q}(z_1) \mathbf{A} \Pi(z_2) + \mathbf{Q}(z_1) \mathbf{A} (\mathbf{Q}(z_2) - \Pi(z_2)), \quad (34)$$

we obtain by Lemma 6 that

$$\begin{aligned} \mathbf{u}^\top \mathbb{E} \mathbf{Q}(z_1) \mathbf{A} \mathbf{Q}(z_2) \mathbf{v} &= \mathbf{u}^\top \mathbb{E} \mathbf{Q}(z_1) \mathbf{A} \Pi(z_2) \mathbf{v} + \mathbf{u}^\top \mathbb{E} \mathbf{Q}(z_1) \mathbf{A} (\mathbf{Q}(z_2) - \Pi(z_2)) \mathbf{v} \\ &= \mathbf{u}^\top \Pi(z_1) \mathbf{A} \Pi(z_2) \mathbf{v} + \mathbf{u}^\top \mathbb{E} \mathbf{Q}(z_1) \mathbf{A} (\mathbf{Q}(z_2) - \Pi(z_2)) \mathbf{v} + o(1), \end{aligned} \quad (35)$$

where the second identity follows from Lemma 6, the Dominated Convergence Theorem and the fact that

$$\|\Pi(z_1)\| = \max_i |z_1 + z_1 m(z_1) \sigma_i|^{-1} \leq |z_1|^{-1}, \quad \|\mathbf{A} \Pi(z_2) \mathbf{v}\| \leq \|\mathbf{A}\| \|\Pi(z_2)\| \leq |z_2|^{-1} \|\mathbf{A}\|.$$

Therefore, our task reduces to finding the deterministic equivalent of

$$\mathbb{E} \mathbf{Q}(z_1) \mathbf{A} (\mathbf{Q}(z_2) - \Pi(z_2)).$$

Denote

$$\mathbf{X}_{-k} = \mathbf{X} - \mathbf{x}_k \mathbf{e}_k^\top, \quad \mathbf{Q}_{-k}(z) = \left( \frac{\mathbf{X}_{-k} \mathbf{X}_{-k}^\top}{N} - z \mathbf{I}_M \right)^{-1}.$$

By Sherman-Morrison formula, one may easily check that

$$\begin{aligned} \mathbf{Q}(z) &= \mathbf{Q}_{-k}(z) - \frac{\frac{1}{N} \mathbf{Q}_{-k}(z) \mathbf{x}_k \mathbf{x}_k^\top \mathbf{Q}_{-k}(z)}{1 + \frac{1}{N} \mathbf{x}_k^\top \mathbf{Q}_{-k}(z) \mathbf{x}_k}, \\ \mathbf{Q}(z) \mathbf{x}_k &= \frac{\mathbf{Q}_{-k}(z) \mathbf{x}_k}{1 + \frac{1}{N} \mathbf{x}_k^\top \mathbf{Q}_{-k}(z) \mathbf{x}_k}. \end{aligned} \quad (36)$$

We show here the following result for future use:

$$\frac{1}{N} \mathbb{E} \text{Tr} \mathbf{C} \mathbf{Q}_{-1}(z_1) \mathbf{A} \mathbf{Q}_{-1}(z_2) = \frac{1}{N} \mathbb{E} \text{Tr} \mathbf{C} \mathbf{Q}(z_1) \mathbf{A} \mathbf{Q}(z_2) + o(1), \quad (37)$$

where  $\mathbf{C} \in \mathbb{R}^{M \times M}$  is a deterministic matrix with  $\|\mathbf{C}\| \leq C$  for some constant  $C$ . We decompose

$$\begin{aligned} &\mathbf{Q}(z_1) \mathbf{A} \mathbf{Q}(z_2) - \mathbf{Q}_{-1}(z_1) \mathbf{A} \mathbf{Q}_{-1}(z_2) \\ &= [\mathbf{Q}(z_1) - \mathbf{Q}_{-1}(z_1)] \mathbf{A} \mathbf{Q}(z_2) + \mathbf{Q}_{-1}(z_1) \mathbf{A} [\mathbf{Q}_{-1}(z_2) - \mathbf{Q}_{-1}(z_1)]. \end{aligned}$$

Applying the identity

$$\mathbf{A}^{-1} - \mathbf{B}^{-1} = \mathbf{B}^{-1} (\mathbf{B} - \mathbf{A}) \mathbf{A}^{-1}, \quad (38)$$

we have for  $i = 1, 2$ , and  $\tilde{\mathbf{C}} \in \mathbb{R}^{M \times M}$  with finite spectral norm (where  $\tilde{\mathbf{C}}$  may be a deterministic matrix, or a random matrix that is either dependent on or independent of  $\mathbf{X}$ ),

$$\frac{1}{N} |\text{Tr}[\mathbf{Q}(z_i) - \mathbf{Q}_{-1}(z_i)] \tilde{\mathbf{C}}| = \frac{1}{N^2} |\mathbf{x}_i^\top \mathbf{Q}(z_i) \tilde{\mathbf{C}} \mathbf{Q}_{-1}(z_1) \mathbf{x}_1| \leq \frac{C}{N^2} \|\mathbf{x}_1\|^2 = o_{a.s.}(1).$$

We denote  $d = \min\{\text{dist}(z_1, \mathbb{R}^+), \text{dist}(z_2, \mathbb{R}^+)\}$ . One may easily check that  $d \sim 1$ . Then by

$$\frac{1}{N} |\text{Tr}[\mathbf{Q}(z_i) - \mathbf{Q}_{-1}(z_i)] \tilde{\mathbf{C}}| \leq \frac{M}{N} (\|\mathbf{Q}(z_i) \tilde{\mathbf{C}}\| + \|\mathbf{Q}_{-1}(z_i) \tilde{\mathbf{C}}\|) \leq \frac{2M}{dN}, \quad \text{for } i = 1, 2,$$

and the Dominated Convergence Theorem, we obtain equation 37. By similar arguments, we get for any deterministic unit vectors  $\mathbf{u}, \mathbf{v}$ ,

$$\begin{aligned} \mathbf{u}^\top \mathbb{E} \tilde{\mathbf{C}} \mathbf{Q}(z_i) \mathbf{C} \mathbf{v} &= \mathbf{u}^\top \mathbb{E} \tilde{\mathbf{C}} \mathbf{Q}_{-k}(z_i) \mathbf{C} \mathbf{v} + o(1) \\ &= \mathbf{u}^\top \tilde{\mathbf{C}} \Pi(z_i) \mathbf{C} \mathbf{v} + o(1), \quad i = 1, 2, \end{aligned} \quad (39)$$

$$\mathbf{u}^\top \mathbb{E} \tilde{\mathbf{C}} \mathbf{Q}(z_1) \mathbf{A} \mathbf{Q}(z_2) \mathbf{C} \mathbf{v} = \mathbf{u}^\top \mathbb{E} \tilde{\mathbf{C}} \mathbf{Q}_{-k}(z_1) \mathbf{A} \mathbf{Q}_{-k}(z_2) \mathbf{C} \mathbf{v} + o(1),$$

1350 where  $\tilde{\mathbf{C}}$  and  $\mathbf{C}$  are deterministic  $M \times M$  matrices with finite spectral norms.  
1351

1352 We denote

$$1353 \quad b_k = \frac{1}{N} \mathbf{x}_k^\top \mathbf{Q}_{-k}(z_2) \mathbf{x}_k, \quad \tilde{b} = \frac{1}{N} \mathbb{E} \mathbf{x}_k^\top \mathbf{Q}_{-k}(z_2) \mathbf{x}_k, \\ 1354 \\ 1355 \quad \mathbf{b}_k = \frac{1}{N} \mathbf{x}_k^\top \mathbf{Q}_{-k}(z_1) \mathbf{x}_k, \quad \tilde{\mathbf{b}} = \frac{1}{N} \mathbb{E} \mathbf{x}_k^\top \mathbf{Q}_{-k}(z_1) \mathbf{x}_k.$$

1356 It follows directly from the proof of equation 37 that  
1357

$$1358 \quad \tilde{b} = \frac{1}{N} \mathbb{E} \text{Tr} \mathbf{\Sigma} \mathbf{Q}_{-1}(z_2) = \frac{1}{N} \text{Tr} \mathbf{\Sigma} \mathbf{\Pi}(z_2) + o(1), \\ 1359 \\ 1360 \quad \tilde{\mathbf{b}} = \frac{1}{N} \mathbb{E} \text{Tr} \mathbf{\Sigma} \mathbf{Q}_{-1}(z_2) = \frac{1}{N} \text{Tr} \mathbf{\Sigma} \mathbf{\Pi}(z_2) + o(1). \quad (40)$$

1362 Recalling equation 32, we rewrite  $\mathbb{E} \mathbf{Q}(z_1) \mathbf{A}(\mathbf{Q}(z_2) - \mathbf{\Pi}(z_2))$  as  
1363

$$1364 \quad \mathbb{E} \mathbf{Q}(z_1) \mathbf{A}(\mathbf{Q}(z_2) - \mathbf{\Pi}(z_2)) = \mathbb{E} \mathbf{Q}(z_1) \mathbf{A} \mathbf{Q}(z_2) (\mathbf{I}_M - \mathbf{Q}^{-1}(z_2) \mathbf{\Pi}(z_2)) \\ 1365 \\ 1366 = \mathbb{E} \mathbf{Q}(z_1) \mathbf{A} \mathbf{Q}(z_2) (\mathbf{\Pi}^{-1}(z_2) - \mathbf{Q}^{-1}(z_2)) \mathbf{\Pi}(z_2) \\ 1367 \\ 1368 = \mathbb{E} \mathbf{Q}(z_1) \mathbf{A} \mathbf{Q}(z_2) \left( -\frac{1}{N} \mathbf{X} \mathbf{X}^\top - z_2 m \mathbf{\Sigma} \right) \mathbf{\Pi}(z_2) \\ 1369 \\ 1370 = \mathbb{E} \mathbf{Q}(z_1) \mathbf{A} \mathbf{Q}(z_2) \frac{\mathbf{\Sigma} \mathbf{\Pi}(z_2)}{1 + \frac{1}{N} \text{Tr} \mathbf{\Sigma} \mathbf{\Pi}(z_2)} - \frac{1}{N} \mathbb{E} \mathbf{Q}(z_1) \mathbf{A} \mathbf{Q}(z_2) \mathbf{X} \mathbf{X}^\top \mathbf{\Pi}(z_2). \quad (41)$$

1372 An application of equation 36 yields that  
1373

$$1374 \quad \frac{1}{N} \sum_{k=1}^N \mathbb{E} \mathbf{Q}(z_1) \mathbf{A} \mathbf{Q}(z_2) \mathbf{x}_k \mathbf{x}_k^\top = \frac{1}{N} \sum_{k=1}^N \mathbb{E} \mathbf{Q}(z_1) \mathbf{A} \frac{\mathbf{Q}_{-k}(z_2) \mathbf{x}_k \mathbf{x}_k^\top}{1 + b_k} \\ 1375 \\ 1376 = \frac{1}{N(1 + \tilde{b})} \sum_{k=1}^N \mathbb{E} \mathbf{Q}(z_1) \mathbf{A} \mathbf{Q}_{-k}(z_2) \mathbf{x}_k \mathbf{x}_k^\top \left[ 1 + \frac{\tilde{b} - b_k}{(1 + b_k)} \right] \\ 1377 \\ 1378 = \frac{1}{N(1 + \tilde{b})} \left[ \sum_{k=1}^N \mathbb{E} \mathbf{Q}(z_1) \mathbf{A} \mathbf{Q}_{-k}(z_2) \mathbf{x}_k \mathbf{x}_k^\top + \mathbb{E} \mathbf{Q}(z_1) \mathbf{A} \mathbf{Q}(z_2) \mathbf{X} \mathbf{X}^\top \right] \\ 1379 \\ 1380 = \frac{1}{1 + \tilde{b}} (\mathbb{E} \mathbf{F}_1 + \mathbb{E} \mathbf{F}_2), \quad (42)$$

1385 where  $\mathbf{B} = \text{diag}(\tilde{b} - b_1, \dots, \tilde{b} - b_N)$ , and  
1386

$$1387 \quad \mathbf{F}_1 = \frac{1}{N} \sum_{k=1}^N \mathbf{Q}(z_1) \mathbf{A} \mathbf{Q}_{-k}(z_2) \mathbf{x}_k \mathbf{x}_k^\top, \quad \mathbf{F}_2 = \frac{1}{N} \mathbf{Q}(z_1) \mathbf{A} \mathbf{Q}(z_2) \mathbf{X} \mathbf{X}^\top.$$

1389 We now bound the spectral norm of  $\mathbf{F}_2$ . Define the event  
1390

$$1392 \quad \mathcal{E} = \left\{ \frac{1}{N} \|\mathbf{Z} \mathbf{Z}^\top\| \leq 2 \left( 1 + \sqrt{\frac{M}{N}} \right)^2 \right\}.$$

1394 We then have  
1395

$$1396 \quad \|\mathbb{E} \mathbf{F}_2\| \leq \mathbb{E} \|\mathbf{F}_2\| \leq \frac{1}{N} \frac{\|\mathbf{A}\|}{d^2} \mathbb{E} \|\mathbf{X} \mathbf{B} \mathbf{X}^\top\| \\ 1397 \\ 1398 \leq \frac{\|\mathbf{A}\| \|\mathbf{\Sigma}\|}{d^2} [4(1 + \sqrt{\phi})^2 \mathbb{E} \|\mathbf{B}\| \delta(\mathcal{E}) + \frac{1}{N} \mathbb{E} \|\mathbf{B}\| \|\mathbf{Z} \mathbf{Z}^\top\| \delta(\mathcal{E}^C)] \\ 1399 \\ 1400 \leq C \mathbb{E} \max_k |\tilde{b} - b_k| + \frac{1}{N} \sqrt{\mathbb{E} \max_k |\tilde{b} - b_k|^2 \mathbb{E} \|\mathbf{Z} \mathbf{Z}^\top\|^2 \delta(\mathcal{E}^C)}.$$

1402 By using the inequality that (see e.g. [Bai & Silverstein \(2010\)](#))  
1403

$$\mathbb{P}(\mathcal{E}^C) \leq N^{-\ell} \text{ for any } \ell > 0,$$

1404 we have

$$\begin{aligned} \mathbb{E}\|\mathbf{Z}\mathbf{Z}^\top\|^2\delta(\mathcal{E}^C) &\leq \mathbb{E}\|\mathbf{Z}\mathbf{Z}^\top\|_{\mathbb{F}}^2\delta(\mathcal{E}^C) \leq \sqrt{\mathbb{E}\|\mathbf{Z}\mathbf{Z}^\top\|_{\mathbb{F}}^4\mathbb{P}(\mathcal{E}^C)} \\ &\leq N^{100}o(N^{-101}) = o(N^{-1}). \end{aligned} \quad (43)$$

1408 It can be shown by Lemma 3 that for  $\ell \geq 1$ ,

$$\begin{aligned} \mathbb{P}(|\tilde{b} - b_k| > t) &\leq \frac{\mathbb{E}|\mathbf{z}_k^\top \Sigma^{1/2} \mathbf{Q}_{-k} \Sigma^{1/2} \mathbf{z}_k - \text{Tr} \Sigma \mathbf{Q}_{-k}(z)|^\ell}{(Nt)^\ell} \\ &= t^{-\ell} \frac{\mathbb{E}[\mathbb{E}_{-k}|\mathbf{z}_k^\top \Sigma^{1/2} \mathbf{Q}_{-k} \Sigma^{1/2} \mathbf{z}_k - \text{Tr} \Sigma \mathbf{Q}_{-k}(z)|^\ell]}{N^\ell} \\ &\leq t^{-\ell} C \frac{\mathbb{E}[(\text{Tr} \mathbf{Q}_{-k}^2)^{\ell/2} + \text{Tr}(\mathbf{Q}_{-k})^\ell]}{N^\ell} \\ &\leq Ct^{-\ell} N^{-\ell/2}, \end{aligned}$$

1418 where we use the fact that

$$\text{Tr}(\mathbf{Q}_{-k}(z_2))^\ell \leq M\|\mathbf{Q}_{-k}(z_2)\|^\ell \leq \frac{M}{d^\ell}.$$

1421 By taking a large enough  $\ell$ , we have

$$\begin{aligned} \mathbb{E} \max_k |\tilde{b} - b_k| &= \left( \int_{t \leq N^{-1/4}} + \int_{t > N^{-1/4}} \right) \mathbb{P}(\max_k |\tilde{b} - b_k| > t) dt \\ &\leq N^{-1/4} + \int_{t > N^{-1/4}} \sum_{k=1}^N \mathbb{P}(|\tilde{b} - b_k| > t) dt \\ &\leq 2N^{-1/4}. \end{aligned} \quad (44)$$

1430 Similarly, one may obtain

$$\mathbb{E} \max_k |\tilde{b} - b_k|^2 = o(1). \quad (45)$$

1432 This, along with equation 43 and equation 44, implies that

$$\|\mathbb{E}\mathbf{F}_2\| = o(1).$$

1435 By using equation 36, we rewrite

$$\begin{aligned} \mathbb{E}\mathbf{F}_1 &= \frac{1}{N} \sum_{k=1}^N \mathbb{E} \mathbf{Q}(z_1) \mathbf{A} \mathbf{Q}_{-k}(z_2) \mathbf{x}_k \mathbf{x}_k^\top \\ &= \frac{1}{N} \sum_{k=1}^N \mathbb{E} [\mathbf{Q}_{-k}(z_1) - \frac{1}{N} \frac{\mathbf{Q}_{-k}(z_1) \mathbf{x}_k \mathbf{x}_k^\top \mathbf{Q}_{-k}(z_1)}{1 + \mathbf{b}_k}] \mathbf{A} \mathbf{Q}_{-k}(z_2) \mathbf{x}_k \mathbf{x}_k^\top \\ &= \frac{1}{N} \sum_{k=1}^N \mathbb{E} \mathbf{Q}_{-k}(z_1) \mathbf{A} \mathbf{Q}_{-k}(z_2) \mathbf{x}_k \mathbf{x}_k^\top - \frac{1}{N} \sum_{k=1}^N \mathbb{E} \frac{\frac{1}{N} \mathbf{Q}_{-k}(z_1) \mathbf{x}_k \mathbf{x}_k^\top \mathbf{Q}_{-k}(z_1)}{1 + \mathbf{b}_k} \mathbf{A} \mathbf{Q}_{-k}(z_2) \mathbf{x}_k \mathbf{x}_k^\top \\ &= \mathbb{E} \mathbf{Q}_{-1}(z_1) \mathbf{A} \mathbf{Q}_{-1}(z_2) \Sigma - \frac{1}{(1 + \tilde{\mathbf{b}})N} \sum_{k=1}^N \mathbb{E} \frac{1}{N} \mathbf{Q}_{-k}(z_1) \mathbf{x}_k \mathbf{x}_k^\top \mathbf{Q}_{-k}(z_1) \mathbf{A} \mathbf{Q}_{-k}(z_2) \mathbf{x}_k \mathbf{x}_k^\top \\ &\quad - \frac{1}{(1 + \tilde{\mathbf{b}})N^2} \sum_{k=1}^N \mathbb{E} \mathbf{Q}_{-k}(z_1) \mathbf{x}_k \mathbf{x}_k^\top \mathbf{Q}_{-k}(z_1) \mathbf{A} \mathbf{Q}_{-k}(z_2) \mathbf{x}_k \mathbf{x}_k^\top \frac{(\mathbf{b}_k - \tilde{\mathbf{b}})}{1 + \mathbf{b}_k} \\ &= \mathbb{E} \mathbf{Q}_{-1}(z_1) \mathbf{A} \mathbf{Q}_{-1}(z_2) \Sigma - \frac{1}{1 + \tilde{\mathbf{b}}} (\mathbb{E}\mathbf{F}_1 + \mathbb{E}\mathbf{F}_2), \end{aligned} \quad (46)$$

1454 where

$$\begin{aligned} \mathbf{F}_1 &= \frac{1}{N} \mathbf{Q}_{-1}(z_1) \mathbf{x}_1 \mathbf{x}_1^\top \mathbf{Q}_{-1}(z_2) \mathbf{A} \mathbf{Q}_{-1}(z_2) \mathbf{x}_1 \mathbf{x}_1^\top, \\ \mathbf{F}_2 &= \frac{1}{N} \mathbf{Q}_{-1}(z_1) \mathbf{x}_1 \mathbf{x}_1^\top \mathbf{Q}_{-1}(z_2) \mathbf{A} \mathbf{Q}_{-1}(z_2) \mathbf{x}_1 \mathbf{x}_1^\top \frac{(\mathbf{b}_1 - \tilde{\mathbf{b}})}{1 + \mathbf{b}_1}. \end{aligned}$$

1458 We first consider  $\mathbb{E}\mathcal{F}_2$ . Let  $\tilde{\mathbf{u}}, \tilde{\mathbf{v}}$  denote a pair of unit vectors satisfying  
 1459

$$1460 \tilde{\mathbf{u}}, \tilde{\mathbf{v}} = \arg \max_{\|\tilde{\mathbf{u}}\|=\|\tilde{\mathbf{v}}\|=1} |\tilde{\mathbf{u}}^\top \mathbb{E}\mathcal{F}_2 \tilde{\mathbf{v}}|,$$

1462 and let  $\mathbf{y} = \mathbf{Q}_{-1}(z_1)\tilde{\mathbf{u}} = (y_1, \dots, y_M)^\top$ . Using the Burkholder's inequality (Burkholder, 1973), we  
 1463 have

$$1464 \mathbb{E}|\mathbf{y}^\top \mathbf{x}_1|^4 = \mathbb{E} \left| \sum_{i=1}^M y_i x_{i1} \right|^4 \leq c \mathbb{E} \left| \sum_{i=1}^M y_i^2 \right|^2 + c \mathbb{E} \sum_{i=1}^M |y_i x_{i1}|^4 \\ 1465 \leq C \mathbb{E} \|\mathbf{y}\|^4 + C \mathbb{E} \sum_{i=1}^M y_i^4 \lesssim 1,$$

1470 where we use the inequality

$$1471 \sum_{i=1}^M y_i^4 \leq \left( \sum_{i=1}^M y_i^2 \right)^2 \leq \|\mathbf{y}\|^4.$$

1474 Likewise, we have  $\mathbb{E}|\mathbf{x}_1^\top \tilde{\mathbf{v}}|^4 \lesssim 1$ . It follows from Lemma 3 that  
 1475

$$1476 \mathbb{E}|\mathbf{b}_1 - \tilde{\mathbf{b}}|^\ell \leq \frac{c}{N^\ell} [(\text{Tr} \mathbf{Q}_{-1}^2(z_1))^{\ell/2} + \text{Tr} \mathbf{Q}_{-1}^\ell(z_1)] \leq \frac{C}{N^{\ell/2}},$$

1478 and

$$1480 \mathbb{E}|\mathbf{x}_1^\top \mathbf{Q}_{-1}(z_1)\mathbf{x}_1|^\ell \leq C \mathbb{E}|\mathbf{x}_1^\top \mathbf{Q}_{-1}(z_1)\mathbf{x}_1 - \mathbb{E}\text{Tr} \Sigma \mathbf{Q}_{-1}(z_1)|^\ell + C|\mathbb{E}\text{Tr} \Sigma \mathbf{Q}_{-1}(z_1)|^\ell \lesssim N^\ell.$$

1482 Since  $\mathbf{b}_k > 1$ , we can bound the spectral norm of  $\mathbb{E}\mathcal{F}_2$  as

$$1483 \|\mathbb{E}\mathcal{F}_2\| = |\tilde{\mathbf{u}}^\top \mathbb{E}\mathcal{F}_2 \tilde{\mathbf{v}}| \leq \mathbb{E}|\tilde{\mathbf{u}}^\top \mathcal{F}_2 \tilde{\mathbf{v}}| \\ 1484 \leq \frac{1}{N} \mathbb{E}|\tilde{\mathbf{u}}^\top \mathbf{Q}_{-1}(z_1)\mathbf{x}_1 \mathbf{x}_1^\top \mathbf{Q}_{-1}(z_1) \mathbf{A} \mathbf{Q}_{-1}(z_2) \mathbf{x}_1 \mathbf{x}_1^\top \tilde{\mathbf{v}}| |\mathbf{b}_1 - \tilde{\mathbf{b}}| \\ 1485 \leq \frac{1}{N} \sqrt{\mathbb{E}|\mathbf{y}^\top \mathbf{x}_1 \mathbf{x}_1^\top \tilde{\mathbf{v}}|^2 \mathbb{E}|\mathbf{x}_1^\top \mathbf{Q}_{-1}(z_1) \mathbf{A} \mathbf{Q}_{-1}(z_2) \mathbf{x}_1 (\mathbf{b}_1 - \tilde{\mathbf{b}})|^2} \\ 1486 \leq \frac{1}{N} \sqrt{\sqrt{\mathbb{E}|\mathbf{y}^\top \mathbf{x}_1|^4 \mathbb{E}|\mathbf{x}_1^\top \tilde{\mathbf{v}}|^4} \sqrt{\mathbb{E}|\mathbf{x}_1^\top \mathbf{Q}_{-1}(z_1) \mathbf{A} \mathbf{Q}_{-1}(z_2) \mathbf{x}_1|^4} \mathbb{E}|\mathbf{b}_1 - \tilde{\mathbf{b}}|^4} \\ 1487 \leq C \frac{1}{N} o(N) = o(1).$$

1494 Therefore, it suffices to find the deterministic equivalent of  $\mathbb{E}\mathcal{F}_1$ . We recall the definition above  
 1495 equation 34 that  $\mathbf{A}_M = o(1)$  if  $\|\mathbf{A}_M\| = o(1)$ . Let  $\mathbb{E}_{-1}(\cdot) = \mathbb{E}[\cdot | \mathbf{x}_2, \dots, \mathbf{x}_N]$ . We have  
 1496

$$1497 \mathbb{E}\mathcal{F}_1 = \frac{1}{N} \mathbb{E} \mathbf{Q}_{-1}(z_1) \mathbf{x}_1 \mathbf{x}_1^\top \mathbf{Q}_{-1}(z_1) \mathbf{A} \mathbf{Q}_{-1}(z_2) \mathbf{x}_1 \mathbf{x}_1^\top \\ 1498 = \frac{1}{N} \mathbb{E} \mathbf{Q}_{-1}(z_1) [\mathbb{E}_{-1} \mathbf{x}_1 \mathbf{x}_1^\top \mathbf{Q}_{-1}(z_1) \mathbf{A} \mathbf{Q}_{-1}(z_2) \mathbf{x}_1 \mathbf{x}_1^\top] \\ 1499 = \frac{1}{N} \mathbb{E} \mathbf{Q}_{-1}(z_1) \Sigma^{1/2} \mathbb{E}_{-1} [\mathbf{z}_1 \mathbf{z}_1^\top \Sigma^{1/2} \mathbf{Q}_{-1}(z_1) \mathbf{A} \mathbf{Q}_{-1}(z_2) \Sigma^{1/2} \mathbf{z}_1 \mathbf{z}_1^\top] \Sigma^{1/2} \\ 1500 = \frac{1}{N} \mathbb{E} \mathbf{Q}_{-1}(z_1) [\text{Tr} \Sigma \mathbf{Q}_{-1}(z_1) \mathbf{A} \mathbf{Q}_{-1}(z_2)] \Sigma \\ 1501 + \frac{1}{N} \mathbb{E} \mathbf{Q}_{-1}(z_1) \Sigma [\mathbf{Q}_{-1}(z_1) \mathbf{A} \mathbf{Q}_{-1}(z_2) + \mathbf{Q}_{-1}(z_2) \mathbf{A} \mathbf{Q}_{-1}(z_1)] \Sigma \\ 1502 + \frac{1}{N} (\mathbb{E} z_{11}^4 - 3) \mathbb{E} \mathbf{Q}_{-1}(z_1) \Sigma^{1/2} \text{diag}(\Sigma^{1/2} \mathbf{Q}_{-1}(z_1) \mathbf{A} \mathbf{Q}_{-1}(z_2) \Sigma^{1/2}) \Sigma^{1/2} \\ 1503 = \frac{1}{N} \mathbb{E} [\text{Tr} \Sigma \mathbf{Q}_{-1}(z_1) \mathbf{A} \mathbf{Q}_{-1}(z_2)] \mathbf{Q}_{-1}(z_1) \Sigma + o(1) \\ 1504 + \frac{1}{N} [\mathbb{E} \text{Tr} \Sigma \mathbf{Q}(z_1) \mathbf{A} \mathbf{Q}(z_2)] \Pi(z_1) \Sigma + o(1), \\ 1505 \\ 1506 \\ 1507 \\ 1508 \\ 1509 \\ 1510 \\ 1511$$

1512 where the last identity is due to equation 37, equation 39 and  
 1513

$$\begin{aligned}
 1514 & \frac{1}{N} \mathbb{E}[\text{Tr} \Sigma \mathbf{Q}_{-1}(z_1) \mathbf{A} \mathbf{Q}_{-1}(z_2)] \mathbf{Q}_{-1}(z_1) \Sigma \\
 1515 & = \frac{1}{N} \mathbb{E}[\text{Tr} \Sigma \mathbf{Q}_{-1}(z_1) \mathbf{A} \mathbf{Q}_{-1}(z_2) - \mathbb{E} \text{Tr} \Sigma \mathbf{Q}_{-1}(z_1) \mathbf{A} \mathbf{Q}_{-1}(z_2)] \mathbf{Q}_{-1}(z_1) \Sigma \\
 1516 & + \frac{1}{N} [\mathbb{E} \text{Tr} \Sigma \mathbf{Q}_{-1}(z_1) \mathbf{A} \mathbf{Q}_{-1}(z_2)] \mathbf{Q}_{-1}(z_1) \Sigma \\
 1517 & = \frac{1}{N} \mathbb{E}[\text{Tr} \Sigma \mathbf{Q}(z_1) \mathbf{A} \mathbf{Q}(z_2)] \Pi(z_1) \Sigma + o(1).
 \end{aligned}$$

1522 By equation 40, equation 42, equation 46 and equation 47 and the fact that  $\|\Pi(z_2)\|$  is bounded, we  
 1523 have

$$\begin{aligned}
 1524 & \frac{1}{N} \mathbb{E} \mathbf{Q}(z_1) \mathbf{A} \mathbf{Q}(z_2) \mathbf{X} \mathbf{X}^\top \Pi(z_2) \\
 1525 & = \frac{1}{(1 + \tilde{b})} \mathbb{E} \mathbf{F}_1 \Pi(z_2) + o(1) \\
 1526 & = \frac{1}{1 + \tilde{b}} \left[ \mathbb{E} \mathbf{Q}_{-1}(z_1) \mathbf{A} \mathbf{Q}_{-1}(z_2) \Sigma \Pi(z_2) - \frac{1}{1 + \tilde{b}} \mathbb{E} \mathbf{F}_1 \Pi(z_2) \right] + o(1) \\
 1527 & = \frac{\mathbb{E} \mathbf{Q}(z_1) \mathbf{A} \mathbf{Q}(z_2) \Sigma \Pi(z_2)}{1 + \frac{1}{N} \text{Tr} \Sigma \Pi(z_2)} - \frac{\frac{1}{N} [\mathbb{E} \text{Tr} \Sigma \mathbf{Q}(z_1) \mathbf{A} \mathbf{Q}(z_2)] \Pi(z_1) \Sigma \Pi(z_2)}{(1 + \frac{1}{N} \text{Tr} \Sigma \Pi(z_2))(1 + \frac{1}{N} \text{Tr} \Sigma \Pi(z_1))} + o(1).
 \end{aligned}$$

1534 This, along with equation 34, equation 41, leads to

$$\begin{aligned}
 1535 & \mathbb{E} \mathbf{Q}(z_1) \mathbf{A} \mathbf{Q}(z_2) \\
 1536 & = \Pi(z_1) \mathbf{A} \Pi(z_2) + \frac{\frac{1}{N} [\mathbb{E} \text{Tr} \Sigma \mathbf{Q}(z_1) \mathbf{A} \mathbf{Q}(z_2)] \Pi(z_1) \Sigma \Pi(z_2)}{(1 + \frac{1}{N} \text{Tr} \Sigma \Pi(z_2))(1 + \frac{1}{N} \text{Tr} \Sigma \Pi(z_1))} + o(1).
 \end{aligned} \tag{48}$$

1539 Multiplying both sides of the above equation on the left by  $\Sigma$ , and taking the trace, we obtain  
 1540

$$\begin{aligned}
 1541 & \frac{1}{N} \mathbb{E} \text{Tr} \Sigma \mathbf{Q}(z_1) \mathbf{A} \mathbf{Q}(z_2) \\
 1542 & = \frac{1}{N} \text{Tr} \Sigma \Pi(z_1) \mathbf{A} \Pi(z_2) + \frac{\frac{1}{N} [\mathbb{E} \text{Tr} \Sigma \mathbf{Q}(z_1) \mathbf{A} \mathbf{Q}(z_2)] \frac{1}{N} \text{Tr} \Sigma \Pi(z_1) \Sigma \Pi(z_2)}{(1 + \frac{1}{N} \text{Tr} \Sigma \Pi(z_2))(1 + \frac{1}{N} \text{Tr} \Sigma \Pi(z_1))} + o(1).
 \end{aligned}$$

1546 It follows that

$$\begin{aligned}
 1547 & \frac{1}{N} \mathbb{E} \text{Tr} \Sigma \mathbf{Q}(z_1) \mathbf{A} \mathbf{Q}(z_2) \\
 1548 & = \left( 1 - \frac{\frac{1}{N} \text{Tr} \Sigma \Pi(z_1) \Sigma \Pi(z_2)}{(1 + \frac{1}{N} \text{Tr} \Sigma \Pi(z_2))(1 + \frac{1}{N} \text{Tr} \Sigma \Pi(z_1))} \right)^{-1} \frac{1}{N} \text{Tr} \Sigma \Pi(z_1) \mathbf{A} \Pi(z_2) + o(1).
 \end{aligned} \tag{49}$$

1552 Plugging equation 49 into equation 48, we get

$$\begin{aligned}
 1553 & \mathbb{E} \mathbf{Q}(z_1) \mathbf{A} \mathbf{Q}(z_2) = \Pi(z_1) \mathbf{A} \Pi(z_2) \\
 1554 & + \frac{\frac{1}{N} \text{Tr} \Sigma \Pi(z_1) \mathbf{A} \Pi(z_2)}{(1 + \frac{1}{N} \text{Tr} \Sigma \Pi(z_2))(1 + \frac{1}{N} \text{Tr} \Sigma \Pi(z_1)) - \frac{1}{N} \text{Tr} \Sigma \Pi(z_1) \Sigma \Pi(z_2)} \Pi(z_1) \Sigma \Pi(z_2) + o(1).
 \end{aligned} \tag{50}$$

1558 The result equation 15 follows by combining the equation 50 with equation 33. Now we prove  
 1559 equation 17. Using a proof analogous to that of equation 33, we can obtain that

$$\frac{1}{M} \text{Tr} \mathbf{C} [\mathbf{Q}(z_1) \mathbf{A} \mathbf{Q}_2(z_2) - \mathbb{E} \mathbf{Q}(z_1) \mathbf{A} \mathbf{Q}(z_2)] = o_{a.s.}(1). \tag{51}$$

1562 We denote the spectral decomposition of  $\mathbf{C}$  by

$$\mathbf{C} = \sum_{i=1}^M \lambda_i \mathbf{u}_i \mathbf{v}_i^\top.$$

1566 By equation 50, we have  
 1567

$$\begin{aligned}
 \frac{1}{M} \text{Tr} \mathbf{C} \mathbb{E} \mathbf{Q}(z_1) \mathbf{A} \mathbf{Q}(z_2) &= \frac{1}{M} \text{Tr} \sum_{i=1}^M \lambda_i \mathbf{u}_i \mathbf{v}_i^\top \mathbb{E} \mathbf{Q}(z_1) \mathbf{A} \mathbf{Q}(z_2) \\
 &= \frac{1}{M} \sum_{i=1}^M \lambda_i \mathbf{v}_i^\top \mathbb{E} \mathbf{Q}(z_1) \mathbf{A} \mathbf{Q}(z_2) \mathbf{u}_i \\
 &= \frac{1}{M} \sum_{i=1}^M \lambda_i \mathbf{v}_i^\top \mathbf{\Pi}(z_1) \mathcal{S}(\mathbf{A}) \mathbf{\Pi}(z_2) \mathbf{u}_i + o(1) \\
 &= \frac{1}{M} \text{Tr} \mathbf{C} \mathbf{\Pi}(z_1) \mathcal{S}(\mathbf{A}) \mathbf{\Pi}(z_2) + o(1).
 \end{aligned}$$

1579 This, along with equation 51, establishes equation 17.  
 1580

### 1581 B.9 DETAILS OF SECTION 5.2

1582 In Case 1 of Section 5.2, where  $\mathbf{X}_1 = \alpha \mathbf{X}_2 + \tilde{\mathbf{X}}_1$ , the following result holds.  
 1583

1584 **Proposition 2.** Suppose that  $\tilde{\mathbf{X}}_1, \mathbf{X}_2, \varepsilon_1$  and  $\varepsilon_2$  satisfy Assumptions 1-2. Then Theorem 1 continues  
 1585 to hold. Moreover, if we additionally impose Assumption 3, then Theorem 2 remains valid.  
 1586

1587 **Proof:** We recall that  $z_1 = -\lambda_t, z_2 = -\lambda_s$ . We only consider  $h_{55}$  in equation 19 here and the  
 1588 remaining terms can be handled analogously. By equation 21, it suffices to estimate

$$\frac{1}{N} \text{Tr}(\mathbf{I}_M + z_2 \mathbf{Q}_2) \mathbf{\Sigma}_2 (\mathbf{I}_M + z_2 \mathbf{Q}_2) (\mathbf{Q}_1 + z_1 \mathbf{Q}_1^2).$$

1591 Since  
 1592

$$\mathbf{Q}_1 = \left( \frac{1}{N} \mathbf{X}_1 \mathbf{X}_1^\top - z_1 \mathbf{I}_M \right)^{-1} = \left( \frac{1}{N} \tilde{\mathbf{X}}_1 \tilde{\mathbf{X}}_1^\top - z_1 \mathbf{I}_M + \mathbf{\Delta} \right)^{-1}, \quad \mathbf{\Delta} = \frac{\alpha}{N} (\tilde{\mathbf{X}}_1 \mathbf{X}_2^\top + \mathbf{X}_2 \tilde{\mathbf{X}}_1^\top + \alpha \mathbf{X}_2 \mathbf{X}_2^\top).$$

1593 We denote  $\tilde{\mathbf{Q}}_1 = \left( \frac{1}{N} \tilde{\mathbf{X}}_1 \tilde{\mathbf{X}}_1^\top - z_1 \mathbf{I}_M \right)^{-1}$ . Applying equation 38 and Lemma 4, we have with high  
 1594 probability,  
 1595

$$\|\mathbf{E}\| = \|\tilde{\mathbf{Q}}_1 - \mathbf{Q}_1\| = \|\tilde{\mathbf{Q}}_1 \mathbf{\Delta} \mathbf{Q}_1\| \leq \frac{1}{|z_1|^2} \|\mathbf{\Delta}\| \lesssim \alpha = o(1). \quad (52)$$

1600 Then we obtain  
 1601

$$\begin{aligned}
 \frac{z_2}{N} \text{Tr} \mathbf{Q}_2 \mathbf{\Sigma}_2 \mathbf{Q}_1^2 &= \frac{z_2}{N} \text{Tr} \mathbf{Q}_2 \mathbf{\Sigma}_2 (\tilde{\mathbf{Q}}_1^2 + \underbrace{\mathbf{E}^2 + \mathbf{E} \tilde{\mathbf{Q}}_1 + \tilde{\mathbf{Q}}_1 \mathbf{E}}_{\hat{\mathbf{E}}}) \\
 &= \frac{z_2}{N} \text{Tr} \mathbf{Q}_2 \mathbf{\Sigma}_2 \tilde{\mathbf{Q}}_1^2 + o_{a.s.}(1),
 \end{aligned}$$

1602 where we use the fact that  
 1603

$$\frac{1}{N} \text{Tr} \mathbf{Q}_2 \mathbf{\Sigma}_2 \hat{\mathbf{E}} \leq \frac{M}{N} \|\mathbf{Q}_2 \mathbf{\Sigma}_2 \hat{\mathbf{E}}\| \lesssim \|\hat{\mathbf{E}}\| = o_{a.s.}(1).$$

1604 By similar argument, we have  
 1605

$$h_{55} = \frac{\xi^2 \sigma^2}{N^2} \text{Tr}(\mathbf{I}_M + z_2 \mathbf{Q}_2) \mathbf{\Sigma}_2 (\mathbf{I}_M + z_2 \mathbf{Q}_2) (\tilde{\mathbf{Q}}_1 + z_1 \tilde{\mathbf{Q}}_1^2) + o_{a.s.}(1).$$

1613 The proof is completed.  
 1614

1615 For Case 2 in Section 5.2, we have the following proposition, which also covers the setting of self-  
 1616 distillation.  
 1617

1618 **Proposition 3.** Suppose  $\mathbf{X}_1 = \mathbf{X}_2 + \mathbf{A}$  is a signal-plus-noise data matrix, with  $\|\mathbf{A}\| = o(\sqrt{M})$ .  
 1619 The regression parameter vector  $\beta = \beta_1 = \beta_2$  satisfies Assumption 3. When  $\lambda_s \neq \lambda_t$ , we have

$$\widehat{\text{Bias}} = \frac{a}{M} \text{Tr} \mathbf{\Sigma}_2 [\mathbf{\Pi}_2(-\lambda_t) - \mathbf{\Pi}_2(-\lambda_s)] + \frac{b}{M} \text{Tr} \mathbf{\Sigma}_2 \mathbf{\Pi}_2'(-\lambda_t) + \frac{c}{M} \text{Tr} \mathbf{\Sigma}_2 \mathbf{\Pi}_2'(-\lambda_s),$$

1620

and

1621

$$\widehat{\text{Var}} = \frac{\xi^2 \sigma^2}{N_1} \left( d \text{Tr} \Sigma_2 [\Pi_2(-\lambda_t) - \Pi_2(-\lambda_s)] + e \text{Tr} \Sigma_2 \Pi'_2(-\lambda_t) + f \text{Tr} \Sigma_2 \Pi'_2(-\lambda_s) \right) \\ + (1 - \xi)^2 \sigma^2 \frac{1}{N_2} \text{Tr} \Sigma_2 [\Pi_2(-\lambda_s) - \lambda_s \Pi'_2(-\lambda_s)],$$

1622

where

1623

$$a = \frac{2\xi \lambda_1 \lambda_2}{\lambda_s - \lambda_t} + \frac{2\xi \lambda_t \lambda_s (\xi \lambda_t - \lambda_s)}{(\lambda_s - \lambda_t)^2} - \frac{2\xi \lambda_s^2 \lambda_t^2}{(\lambda_s - \lambda_t)^3}, \quad b = \xi^2 \lambda_t^2 - \frac{2\xi^2 \lambda_t^2 \lambda_s}{\lambda_t - \lambda_s}, \\ c = \lambda_s^2 - \frac{2\xi \lambda_t \lambda_s^2}{\lambda_t - \lambda_s}, \quad d = \frac{2\lambda_s}{\lambda_t - \lambda_s} + \frac{\lambda_s^2}{(\lambda_t - \lambda_s)^2} + \frac{\lambda_t \lambda_s^2}{(\lambda_s - \lambda_t)^3} + \frac{2\lambda_s \lambda_t}{(\lambda_t - \lambda_s)^2}, \\ e = -\lambda_t + \frac{2\lambda_s \lambda_t}{(\lambda_t - \lambda_s)^2} - \frac{\lambda_t \lambda_s^2}{(\lambda_s - \lambda_t)^2}, \quad f = \frac{\lambda_s^2}{\lambda_s - \lambda_t} - \frac{\lambda_t \lambda_s^2}{(\lambda_s - \lambda_t)^2}.$$

1624

When  $\lambda = \lambda_s = \lambda_t$ ,  $\widehat{\text{Bias}}$  is given in equation 54 and

1625

$$\widehat{\text{Var}} = (1 - \xi)^2 \sigma^2 \frac{1}{N_2} \text{Tr} \Sigma_2 (\Pi_2 + \lambda \Pi'_2) + \frac{\xi^2 \sigma^2}{N_1} \text{Tr} \Sigma_2 [\Pi_2 - 3\lambda \Pi'_2 + 3\lambda^2 \Pi_2^{(2)} - \lambda^3 \Pi_2^{(3)}],$$

1626

$$\text{with } \Pi_2^{(k)} = \frac{d^k \Pi_2(z)}{dz^k} \Big|_{z=-\lambda}.$$

1627

**Proof:** By an argument analogous to that used for equation 52, one may readily verify that

1628

$$\|\mathbf{Q}_1(z) - \mathbf{Q}_2(z)\| = o(1).$$

1629

Then equation 26 becomes

1630

$$\frac{1}{M} \text{Tr} \Sigma_2 \mathbf{H} \mathbf{H}^\top = \frac{1}{M} \left[ \xi^2 z_1^2 \text{Tr} \Sigma_2 \mathbf{Q}_2^2(z_1) + 2\xi z_1 z_2 \underbrace{\text{Tr} \Sigma_2 \mathbf{Q}_2(z_1) \mathbf{Q}_2(z_2)}_{t_1} + 2\xi^2 z_1^2 z_2 \underbrace{\text{Tr} \Sigma_2 \mathbf{Q}_2(z_2) \mathbf{Q}_2^2(z_1)}_{t_2} \right. \\ + z_2^2 \text{Tr} \Sigma_2 \mathbf{Q}_2^2(z_2) + 2\xi z_1 z_2^2 \underbrace{\text{Tr} [\mathbf{Q}_2(z_2) \Sigma_2 \mathbf{Q}_2(z_2) \mathbf{Q}_2(z_1)]}_{t_3} \\ \left. + \xi^2 z_1^2 z_2^2 \underbrace{\text{Tr} [\mathbf{Q}_2(z_2) \Sigma_2 \mathbf{Q}_2(z_2) \mathbf{Q}_2^2(z_1)]}_{t_4} \right] + o(1).$$

1631

We note that when  $z_1 \neq z_2$ ,

1632

$$\mathbf{Q}_2(z_1) - \mathbf{Q}_2(z_2) = (z_1 - z_2) \mathbf{Q}_2(z_1) \mathbf{Q}(z_2), \quad \mathbf{Q}_2(z_1) \mathbf{Q}_2(z_2) = \mathbf{Q}_2(z_2) \mathbf{Q}_2(z_1).$$

1633

Then we have

1634

$$\frac{1}{M} t_1 = \frac{2\xi z_1 z_2}{M} \text{Tr} \Sigma_2 \frac{\mathbf{Q}_2(z_1) - \mathbf{Q}_2(z_2)}{z_1 - z_2} = \frac{2\xi z_1 z_2}{M(z_1 - z_2)} \text{Tr} \Sigma_2 [\Pi_2(z_1) - \Pi_2(z_2)] + o_{a.s.}(1),$$

1635

1636

1637

1638

1639

1640

1641

1642

1643

1644

1645

1646

1647

1648

1649

1650

1651

1652

1653

1654

1655

1656

1657

1658

1659

1660

1661

1662

1663

1664

1665

1666

1667

1668

1669

1670

1671

1672

1673

$$\frac{1}{M} t_2 = \frac{2\xi^2 z_1^2 z_2}{M(z_1 - z_2)} \text{Tr} \Sigma_2 [\mathbf{Q}_2(z_1) - \mathbf{Q}_2(z_2)] \mathbf{Q}_2(z_1)$$

$$= -\frac{2\xi^2 z_1^2 z_2}{M(z_1 - z_2)} \text{Tr} \Sigma_2 \Pi'_2(z_1) - \frac{2\xi^2 z_1^2 z_2}{M(z_1 - z_2)^2} \text{Tr} \Sigma_2 [\Pi_2(z_1) - \Pi_2(z_2)] + o_{a.s.}(1),$$

$$\frac{1}{M} t_3 = \frac{2\xi z_1 z_2^2}{M(z_1 - z_2)} \text{Tr} \Sigma_2 [\mathbf{Q}_2(z_1) - \mathbf{Q}_2(z_2)] \mathbf{Q}_2(z_2)$$

$$= -\frac{2\xi z_1 z_2^2}{M(z_1 - z_2)} \text{Tr} \Sigma_2 \Pi'_2(z_2) + \frac{2\xi z_1 z_2^2}{M(z_1 - z_2)^2} \text{Tr} \Sigma_2 [\Pi_2(z_1) - \Pi_2(z_2)] + o_{a.s.}(1),$$

$$\frac{1}{M} t_4 = \frac{\xi^2 z_1^2 z_2^2}{M(z_1 - z_2)^2} \text{Tr} \Sigma_2 [\mathbf{Q}_2(z_1) - \mathbf{Q}_2(z_2)]^2$$

$$= \frac{\xi z_1^2 z_2^2}{M(z_1 - z_2)^2} \text{Tr} \Sigma_2 \left[ \mathbf{Q}_2^2(z_2) + \mathbf{Q}_2^2(z_1) - 2 \frac{\mathbf{Q}_2(z_1) - \mathbf{Q}_2(z_2)}{z_1 - z_2} \right]$$

$$= \frac{\xi^2 z_1^2 z_2^2}{M(z_1 - z_2)^2} \text{Tr} \Sigma_2 [\Pi'_2(z_1) + \Pi'_2(z_2)] - \frac{2\xi z_1^2 z_2^2}{M(z_1 - z_2)^3} \text{Tr} \Sigma_2 [\Pi_2(z_1) - \Pi_2(z_2)] + o_{a.s.}(1).$$

Based on above results, we have  $\frac{1}{M} \text{Tr} \Sigma_2 \mathbf{H} \mathbf{H}^\top = \widehat{\text{Bias}} + o_{a.s.}(1)$ . As for the variance, one may check that

$$\begin{aligned}
h_{55} = & \frac{\xi^2 \sigma^2}{N_1} \text{Tr} \left[ \Sigma_2 [\Pi_2(z_1) + z_1 \Pi'_2(z_1)] + 2z_2(z_2 - z_1)^{-1} \Sigma_2 [\Pi_2(z_2) - \Pi_2(z_1)] \right. \\
& + z_2^2(z_2 - z_1)^{-1} \Sigma_2 \Pi'_2(z_2) + z_2^2(z_1 - z_2)^{-2} \Sigma_2 [\Pi_2(z_1) - \Pi_2(z_2)] - 2z_1 z_2(z_2 - z_1)^{-2} \text{Tr} \Sigma_2 \Pi'_2(z_1) \\
& + 2z_1 z_2(z_1 - z_2)^{-2} \text{Tr} \Sigma_2 [\Pi_2(z_2) - \Pi_2(z_1)] \\
& \left. + z_1 z_2^2(z_1 - z_2)^{-2} \text{Tr} \Sigma_2 [\Pi'_2(z_1) + \Pi'_2(z_2)] - z_1 z_2^2(z_1 - z_2)^3 \text{Tr} [\Pi_2(z_1) - \Pi_2(z_2)] \right] + o_{a.s.}(1), \tag{53}
\end{aligned}$$

and the limit of  $h_{66}$  is the same as that in equation 24, where  $h_{55}, h_{66}$  are given in Appendix B.2.

When  $\lambda = \lambda_t = \lambda_s$ , denoting  $\mathbf{Q}_2 = \mathbf{Q}_2(-\lambda)$ , we have

$$\begin{aligned} \frac{1}{M} \text{Tr} \Sigma_2 \mathbf{H} \mathbf{H}^\top &= \frac{1}{M} \left[ (1 + \xi)^2 \lambda^2 \text{Tr} \Sigma_2 \mathbf{Q}_2^2 - (2\xi + \xi^2) \lambda^3 \text{Tr} \Sigma_2 \mathbf{Q}_2^3 + \xi^2 \lambda^4 \text{Tr} \Sigma_2 \mathbf{Q}_2^4 \right] \\ &= \underbrace{\frac{1}{M} \left[ (1 + \xi)^2 \lambda^2 \text{Tr} \Sigma_2 \mathbf{\Pi}'_2 - (2\xi + \xi^2) \lambda^3 \text{Tr} \Sigma_2 \mathbf{\Pi}_2^{(2)} + \xi^4 \lambda^4 \text{Tr} \Sigma_2 \mathbf{\Pi}_2^{(3)} \right]}_{\text{Bias}} + o_{a.s.}(1), \end{aligned} \quad (54)$$

where we use Vitali's convergence theorem. Similarly, we have

$$h_{55} = \frac{\xi^2 \sigma^2}{N_1} \text{Tr} \Sigma_2 [\boldsymbol{\Pi}_2 + 3z\boldsymbol{\Pi}'_2 + 3z^2\boldsymbol{\Pi}^{(2)}_2 + z^3\boldsymbol{\Pi}^{(3)}_2] + o_{a.s.}(1),$$

and the limit of  $h_{66}$  coincides with the one given in equation 24. The proof is completed.

## C ADDITIONAL EXPERIMENTAL DETAILS

## C.1 NONLINEAR MODELS

When considering more complex models, we fix the student model and let the teacher model be a deeper fully connected neural network:

$$f_{\text{NN}}^t = \mathbf{a}_t^\top \sigma(\widetilde{\mathbf{W}}_3 \sigma(\widetilde{\mathbf{W}}_2 \sigma(\widetilde{\mathbf{W}}_1 \mathbf{x}))),$$

where

$$\mathbf{a}_t = \arg \min_{\mathbf{a}} \frac{1}{N_1} \|\mathbf{y}_1 - [\sigma(\widetilde{\mathbf{W}}_3 \sigma(\widetilde{\mathbf{W}}_2 \sigma(\widetilde{\mathbf{W}}_1 \mathbf{X}_1)))^\top \mathbf{a}]^2 + \lambda_t \|\mathbf{a}\|^2.$$

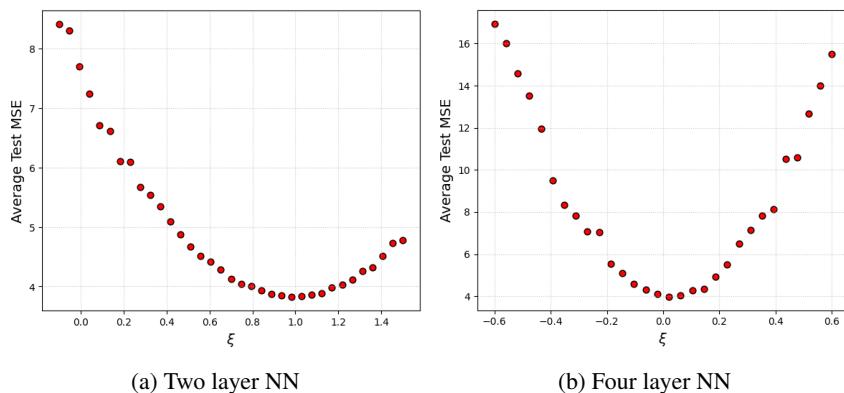


Figure 4: Excess risk estimated over 30 trials. We set  $(M, N_1, N_2) = (50, 200, 100)$ . (a) Settings:  $(\lambda_t, \lambda_s) = (0.5, 0.2)$ ,  $\sigma(x) = x^3$ . The weight matrices  $\tilde{\mathbf{W}} \in \mathbb{R}^{n_1 \times M}$  and  $\mathbf{W} \in \mathbb{R}^{n \times M}$  have i.i.d. centered Gaussian entries with variance  $M^{-1}$ , where  $(n, n_1) = (100, 200)$ . (b) Settings:  $\lambda_t = \lambda_s = 0.2$ ,  $(n_0, n_1, n_2, n_3) = (M, 600, 400, 200)$ . The weight matrices  $\tilde{\mathbf{W}}_i \in \mathbb{R}^{n_i \times n_{i-1}}$  have i.i.d. centered Gaussian random variables with variance  $n_{i-1}^{-1}$ . We use the Leaky ReLU activation:  $\sigma(x) = 0.01x\delta(x \leq 0) + x\delta(x > 0)$ .

## C.2 DEMONSTRATION OF COROLLARY 3

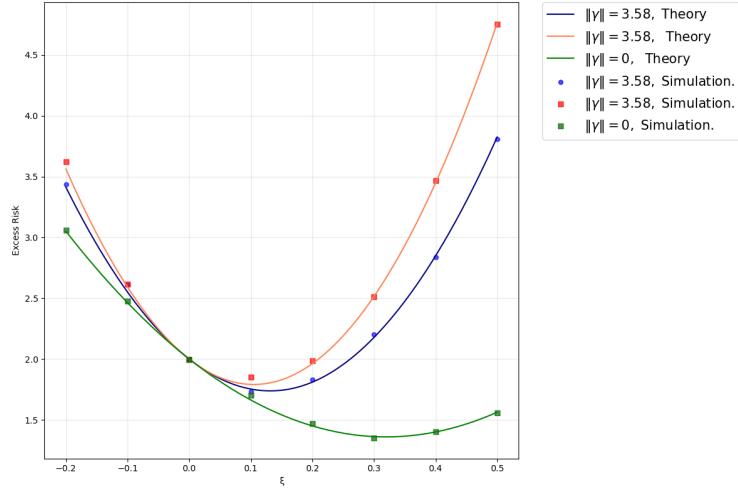


Figure 5: Theoretical predictions (solid curves) versus simulation results (scatter points, averaged over 100 independent trials) for ridgeless regression. We set  $(M, N_1, N_2) = (400, 600, 600)$ ,  $\beta_2 = \frac{4}{\sqrt{M}}(1, \dots, 1)^\top$ ,  $\sigma^2 = 1$  and  $\Sigma_2 = \mathbf{I}_M$ . We label the case  $\|\gamma\| = 3.58$  as  $\gamma = \frac{-8}{\sqrt{M}}(1, \dots, 1, 0, \dots, 0)^\top$  with the first  $M/5$  entries equal to 1. The orange and green curves correspond to the setting where  $\Sigma_1 = \text{diag}(4, \dots, 4, \frac{1}{4}, \dots, \frac{1}{4})$ , with the first half of the diagonal entries equal to 4 and the second half equal to  $\frac{1}{4}$ . The dark blue curve corresponds to the setting where  $\Sigma_1 = 4\mathbf{I}_M$ .

Figure 5 presents empirical results that support 3. The gap between the orange and green curves quantifies the impact of model shift on the excess risk. Furthermore, the gap between the dark blue and orange curves reflects the role of the term  $\text{Tr} \Sigma_1^{-1} \Sigma_2$  as characterized in Corollary 3.

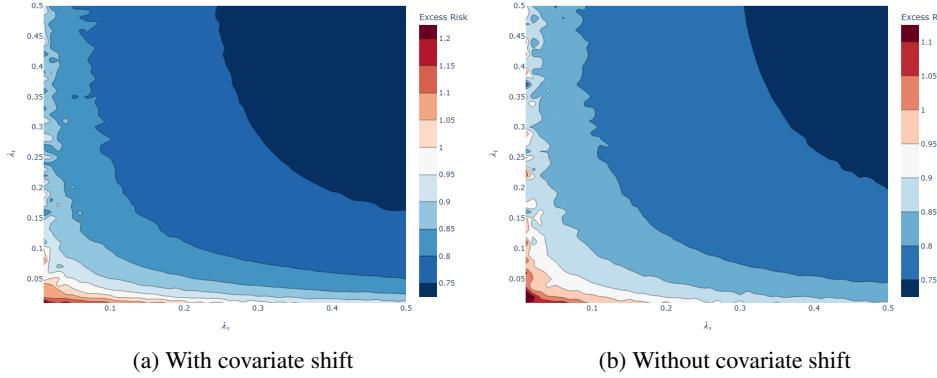
### C.3 IMPACT OF REGULARIZATION PARAMETERS

To examine the impact of the regularization parameters  $\lambda_t, \lambda_s$ , we plot the empirical excess risk of the student model for  $(\lambda_t, \lambda_s) \in [0.01, 0.5]^2$  in Figures 6-8 (averaged over 5 trials), correspond-

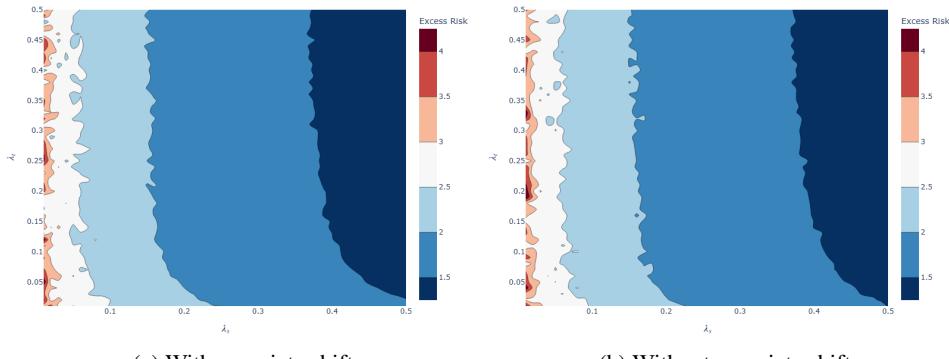
1782 ing to  $\xi = 0.5, -0.5$  and  $1.5$ , respectively. We set  $\beta_1 = \beta_2 \sim \mathcal{N}(0, \frac{1}{M} \mathbf{I}_M)$ ,  $(M, N_1, N_2) =$   
 1783  $(400, 300, 200)$ ,  $\sigma^2 = 1$ . We set  $\Sigma_2 = \mathbf{I}_M$  in the absence of covariate shift. Under covariate shift,  
 1784 we set  $\Sigma_1 = \text{diag}(d_1, \dots, d_M)$ , where  
 1785

$$d_i = 0.64\delta(i \leq M/2) + 0.25\delta(M/2 < i \leq M).$$

1786 From these figures, we observe that when  $\xi > 1$ , the influence of  $\lambda_t$  becomes large. In contrast, in  
 1787 the case  $\xi = -0.5$ ,  $\lambda_s$  almost dominates the variation of the excess risk, reflecting a weaker impact  
 1788 of the teacher's guidance (anti-learning against the teacher's supervision).  
 1789



1790 Figure 6: Excess risk when  $\xi = 0.5$ .  
 1791



1792 Figure 7: Excess risk when  $\xi = -0.5$ .  
 1793

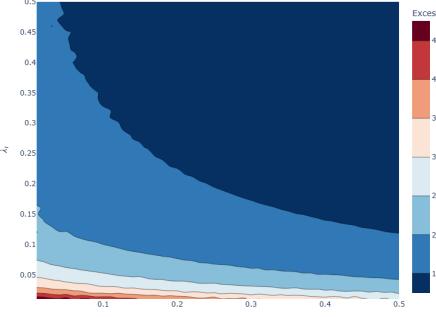
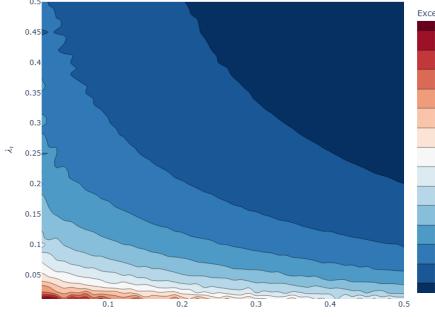
1836  
 1837  
 1838  
 1839  
 1840  
 1841  
 1842  
 1843  
 1844  
 1845  
 1846  
 1847  
 1848  
 1849  
 1850  
 1851  
 1852  
 1853  
 1854  
 1855  
 1856  
 1857        
 1858  
 1859  
 1860  
 1861  
 1862  
 1863  
 1864  
 1865  
 1866  
 1867  
 1868      (a) With covariate shift  
 1869        
 1870      (b) Without covariate shift  
 1871  
 1872  
 1873  
 1874  
 1875  
 1876  
 1877  
 1878  
 1879  
 1880  
 1881  
 1882  
 1883  
 1884  
 1885  
 1886  
 1887  
 1888  
 1889

Figure 8: Excess risk when  $\xi = 1.5$ .