High-dimensional Asymptotics of Generalization Performance in Continual Ridge Regression

Yihan Zhao ZHAO-YH23@MAILS.TSINGHUA.EDU.CN
Department of Mathematical Sciences
Tsinghua University
Beijing, 100084, China

Wenqing Su SUWENQING@SNNU.EDU.CN
School of Mathematics and Statistics
Shaanxi Normal University
Xi'an, 710119, China

Ying Yang YANGYING@TSINGHUA.EDU.CN
Department of Statistics and Data Science
Tsinghua University
Beijing, 100084, China

Abstract

Continual learning is motivated by the need to adapt to real-world dynamics in tasks and data distribution while mitigating catastrophic forgetting. Despite significant advances in continual learning techniques, the theoretical understanding of their generalization performance lags behind. This paper examines the theoretical properties of continual ridge regression in high-dimensional linear models, where the dimension is proportional to the sample size in each task. Using random matrix theory, we derive exact expressions of the asymptotic prediction risk, thereby enabling the characterization of three evaluation metrics of generalization performance in continual learning: average risk, backward transfer, and forward transfer. Furthermore, we present the theoretical risk curves to illustrate the trends in these evaluation metrics throughout the continual learning process. Our analysis reveals several intriguing phenomena in the risk curves, demonstrating how model specifications influence the generalization performance. Simulation studies are conducted to validate our theoretical findings.

Keywords: Continual learning, continual ridge regression, high-dimensional asymptotics, generalization performance, risk curves

1 Introduction

Continual learning, also termed incremental learning or lifelong learning, trains models on sequential tasks with evolving data distributions. The ideal learner is obtained using both previous data and current data to ensure that the current model is adapted to the current task while maintaining the performance on previous tasks. However, due to memory limitations, the previous data may not be available when the new task arrives, which results in performance reduction on previous tasks. This phenomenon is referred to as catastrophic forgetting (McCloskey and Cohen, 1989; McClelland et al., 1995). To mitigate the influence

of catastrophic forgetting, substantial progress has been made in continual learning by designing various techniques, leading to impressive success in practical applications (De Lange et al., 2022; Wang et al., 2024). Despite their methodological innovations, theoretical understandings of continual learning methods remain scarce even in relatively simple models. When a provably effective continual learning method is applied to a statistical model, how can we characterize its generalization performance (the ability to maintain stable predictions across previous tasks while adapting to new prediction tasks)? Specifically, how do model specifications, such as model complexity and task similarity, influence the generalization performance under different patterns of task dynamics? Addressing this question comprehensively relies on an insightful theory to capture the stepwise performance in continual learning, which is underdeveloped in most existing work.

To bridge this gap, we begin with the linear regression framework, since it provides a principled starting point that balances analytical tractability and practical utility. We focus on the theoretical properties of continual ridge regression (Li et al., 2023) in high-dimensional random-designed linear models. As a standard continual learning method in regression models, continual ridge regression adds l_2-regularization terms to the loss functions to constrain changes in the regression coefficients. Motivated by the question mentioned above, our primary concern is to establish a rigorous characterization of the stepwise generalization performance in continual ridge regression. Building upon this characterization, we aim to interpret how the generalization performance is influenced by key model specifications. Besides, the severity of catastrophic forgetting can be modulated to some extent by the choice of regularization parameters. This allows us to quantitatively investigate how catastrophic forgetting affects the generalization performance in continual ridge regression. Our contributions are summarized as follows.

- Using asymptotic random matrix theory, we derive the exact expression of asymptotic prediction risk on a single prediction task in a high-dimensional regime where the parameter dimension grows proportionally with the training sample size in each task. This result explicitly characterizes how the asymptotic prediction risk depends on model complexity and task similarity, where the model complexity is quantified by the ratio of parameter dimension to sample size, and task similarity is characterized by the joint empirical spectral distribution of task-specific covariance matrices.

- We establish the asymptotic behavior of three evaluation metrics for continual ridge regression—average risk, backward transfer, and forward transfer—and characterize their precise dependence on model complexity and task similarity in the high-dimensional setting. These metrics capture the key capabilities of interest in continual learning: average risk measures the overall performance across tasks, backward transfer evaluates how learning new tasks influences performance on previously learned tasks, and forward transfer quantifies how learning a current task improves performance on future tasks.

- We demonstrate our theoretical findings through three representative examples with different structures of task-specific covariance matrices. In each example, the asymptotic risk curves are derived to visualize the stepwise generalization performance in the procedure of continual ridge estimation. We find that the asymptotic risk curves

behave differently according to the dynamics of task-specific covariance matrices and the choice of regularization parameters. In addition, simulations are also conducted to verify our theoretical risk curves.

1.1 Related Works

Over the past few years, researchers have developed a massive number of continual learning methods. Roughly speaking, these methods can be categorized into three groups, including regularization-based approach (Kirkpatrick et al., 2017; Li and Hoiem, 2018; Ritter et al., 2018; Jung et al., 2020), replay-based approach (Lopez-Paz and Ranzato, 2017; Rebuffi et al., 2017; Riemer et al., 2019; Buzzega et al., 2021; Jin et al., 2021) and architecture-based approach (Ramesh and Chaudhari, 2022; Gurbuz and Dovrolis, 2022; Fini et al., 2022). Regularization-based approaches introduce regularization terms into loss functions to balance the information from old tasks and new tasks. Kirkpatrick et al. (2017) used weighted regularization methods to overcome catastrophic forgetting in neural networks. Li and Hoiem (2018) applied function regularization methods to enable learning without forgetting in convolutional neural networks. Replay-based approaches store a few previous data in a memory buffer and replay them in future tasks. Lopez-Paz and Ranzato (2017) proposed the Gradient Episodic Memory (GEM) model, which alleviated forgetting while allowing beneficial transfer of knowledge to previous tasks. Jin et al. (2021) proposed the Gradient based Memory EDiting (GMED) framework, which edited the examples stored in the replay memory. Architecture-based approaches construct task-specific parameters to prevent catastrophic forgetting. Ramesh and Chaudhari (2022) grouped the associated tasks and split the learning capacity across sets of synergistic tasks. Gurbuz and Dovrolis (2022) employed connection rewriting in sparse neural networks to create new plastic paths that reused existing knowledge on novel tasks.

In recent years, several theoretical studies have been conducted on continual learning. For example, Lee et al. (2021) studied continual learning in the teacher-student setup to find out the theoretical reasons for interference between tasks. Li et al. (2022) established sample complexity and generalization error bounds for new tasks in continual representation learning problems. Yang et al. (2023) proposed a theoretical analysis of a SPCA-based continual learning algorithm using high-dimensional statistics. Wen et al. (2024) analyzed the theoretical properties of contrastive continual learning methods. The results in the above studies are not as explicit as ours since they focus on model-free prediction problems. By contrast, some theoretical works have focused on results within concrete models, especially regression models. Evron et al. (2022) theoretically characterized the worst-case catastrophic forgetting in over-parameterized linear regression models. Different from our work, they assumed that the collection of tasks has cyclic task orderings, so that the mechanism of forgetting is quite different. Li et al. (2023) presented a fixed-design analysis of the l_2-regularized continual learning algorithm for linear regression models. However, they only consider a two-task setting and provide risk bounds, which cannot adequately explain the generalization performance of continual ridge regression. Very recently, Zhao et al. (2024) initiated a statistical analysis of a family of generalized l_2-regularized continual learning algorithms for linear regression models. Although they derived precise results of estimation error, their analysis did not provide a comprehensive evaluation of the generalization

ability of estimators in continual learning. Goldfarb and Hand (2025) demonstrated that overparameterization can mitigate forgetting by considering a two-task latent space regression model. While their work revealed the influence of model complexity on generalization performance, it did not account for the impact of task similarities.

1.2 Organizations and Notations

The rest of this paper is organized as follows. In Section 2, we introduce the settings of continual risk regression, formally define the continual ridge estimator and compute its prediction risk. In Section 3, we present our main result of asymptotic prediction risk and provide three concrete examples of covariance structures to illustrate our main theorem. In Section 4, we derive asymptotic risk curves to evaluate the performance of continual ridge regression, and conduct experiments to verify our results. We also discuss the procedure of choosing regularization parameters in continual learning frameworks. In Section 5, we conclude the paper and provide possible extensions. Technical proofs are included in the Appendix.

Throughout this article, $\mathbb{C}^+ = \{z \in \mathbb{C}, \mathrm{Im}(z) > 0\}$, and $\lfloor x \rfloor$ denotes the greatest integer less than or equal to x. Let $\|\cdot\| := \|\cdot\|_2$ denote the Euclidean norm of a real vector, and $\langle \cdot, \cdot \rangle$ denote the inner product induced by the Euclidean norm. For any symmetric real matrix $A \in \mathbb{R}^{p\times p}$, $\lambda_{max}(A)$ and $\lambda_{min}(A)$ denote the largest and smallest eigenvalues of A respectively. Let $\{\lambda_t\}_{t\geq 1} \in \mathbb{R}^p$ be a sequence of positive numbers, and $\{A_t\}_{t\geq 1} \in \mathbb{R}^{p\times p}$ be a sequence of positive definite matrices. To handle boundary cases in recursive expressions uniformly (e.g., in Lemma 1 and Theorem 3), we define the empty scalar product as $\prod_{t=k}^{l} \lambda_t = 1$ if $k > l$, and the empty matrix product as $A_l A_{l-1} \cdots A_k = I_p$ if $k > l$.

2 Continual Ridge Regression

Consider a continual learning problem with a sequence of T tasks. In the t-th task ($t = 1, \cdots, T$), we observe a dataset $\mathcal{D}_t = \{(x_{t,i}, y_{t,i}) \in \mathbb{R}^p \times \mathbb{R}\}_{i=1}^{n_t}$, where $x_{t,i}$ is a random feature vector, $y_{t,i}$ is the corresponding response, and n_t denotes the sample size of $\mathcal{D}_t$. We assume that the feature vectors in different tasks are independent, and a shared linear regression model governs all tasks, such that

$$y_{i,t} = x_{i,t}^\top \beta + \epsilon_{i,t}, \quad t = 1, \cdots, T, \quad i = 1, \cdots, n_t,$$

where the feature vectors in the t-th task $\{x_{t,i}\}_{i=1}^{n_t}$ are i.i.d. with $\mathbb{E}(x_{t,i}) = 0, \mathrm{Cov}(x_{t,i}) = \Sigma_t$, and the noise terms $\{\epsilon_{t,i}\}_{t=1,\cdots,T, i=1,\cdots,n_t}$, across all tasks are i.i.d. with $\mathbb{E}(\epsilon_{t,i}) = 0, \mathrm{Var}(\epsilon_{t,i}) = \sigma^2$. For analytical convenience, we rewrite the linear model for each task t as the matrix form

$$y_t = X_t \beta + \epsilon_t, \quad t = 1, \cdots, T, \tag{1}$$

where $X_t = (x_{1,t}, \cdots, x_{n_t,t})^\top \in \mathbb{R}^{n_t \times p}$, $y_t = (y_{1,t}, \cdots, y_{n_t,t})^\top \in \mathbb{R}^{n_t}$ and $\epsilon_t = (\epsilon_{1,t}, \cdots, \epsilon_{n_t,t})^\top \in \mathbb{R}^{n_t}$.

In continual learning frameworks, the estimation of β is updated upon the arrival of each new task. Let $\hat{\beta}_t$ denote the estimator after the arrival of task t, then $\hat{\beta}_t$ is computed using the current dataset $\mathcal{D}_t$ while retaining no direct access to previous datasets $\mathcal{D}_1, \cdots, \mathcal{D}_{t-1}$ due

to memory constraints. Instead of previous datasets, historical information is incorporated through the previous estimator $\hat{\beta}_{t-1}$. Continual ridge regression updates the estimator by fitting the current dataset with ridge regularization to constrain the change from previous estimator $\hat{\beta}_{t-1}$ to the new estimator $\hat{\beta}_t$. Specifically, the updating rule is

$$\hat{\beta}_t = \arg\min_{\beta}\{\frac{1}{n_t}\|X_t\beta - y_t\|^2 + \lambda_t\|\beta - \hat{\beta}_{t-1}\|^2\}, \quad t = 1, \cdots, T, \tag{2}$$

where $\lambda_t > 0$ is the ridge tuning parameter at step t. The solution of continual ridge estimator is

$$\hat{\beta}_t = \hat{\beta}_t^{\text{ridge}} + \lambda_t(\hat{\Sigma}_t + \lambda_t I_p)^{-1}\hat{\beta}_{t-1}, \quad t = 1, \cdots, T,$$

where $\hat{\beta}_t^{\text{ridge}} = (\hat{\Sigma}_t + \lambda_t I_p)^{-1}(\frac{1}{n_t}X_t^\top y_t)$ is the ridge estimator using dataset $\mathcal{D}_t$, and $\hat{\Sigma}_t = \frac{1}{n_t}X_t^\top X_t$ denotes the sample covariance matrix of X_t. If we initialize $\hat{\beta}_0 = 0$, then the explicit form of the continual ridge estimator is given by

$$\hat{\beta}_t = \hat{\beta}_t^{\text{ridge}} + A_t\hat{\beta}_{t-1}^{\text{ridge}} + A_tA_{t-1}\hat{\beta}_{t-2}^{\text{ridge}} + \cdots + A_tA_{t-1}\cdots A_2\hat{\beta}_1^{\text{ridge}}, \quad t = 1, \cdots, T, \tag{3}$$

where $A_t = \lambda_t(\hat{\Sigma}_t + \lambda_t I_p)^{-1}$.

To evaluate the generalization performance of continual learning methods in linear regression model, we first define the out-of-sample excess risk conditional on training data $X = (X_1^\top, \cdots, X_T^\top)^\top$. If the input test data x_0 satisfies $\mathbb{E}(x_0) = 0, \text{Cov}(x_0) = \Sigma_0$, and the response variable is denoted as y_0, then the corresponding prediction risk of an estimator $\hat{\beta}$ is

$$R_X(\hat{\beta}; \beta, \Sigma_0) = \mathbb{E}[(x_0^\top\hat{\beta} - y_0)^2|X] - \mathbb{E}[(x_0^\top\beta - y_0)^2|X] = \mathbb{E}[(x_0^\top\hat{\beta} - x_0^\top\beta)^2|X].$$

Our primary interest is to evaluate the generalization performance of continual ridge estimator given by (3). Advised by Wang et al. (2024), we consider the following three aspects of evaluation rules:

- **Overall performance**, which is evaluated by (weighted) average risk

$$\bar{R}_X(\hat{\beta}_T; \beta) = \sum_{t=1}^{T}\omega_t R_X(\hat{\beta}_T; \beta, \Sigma_t),$$

  where $\omega_t > 0$ for all t and $\sum_{t=1}^{T}\omega_t = 1$. The weights are chosen according to the prior knowledge of the prediction task. By default, the test data at step T comes from one of the T training tasks with probability proportional to the sample size of each training dataset. Under this assumption, we may choose $\omega_t = n_t/\sum_{k=1}^{T} n_k$.

- **Memory stability**, which is evaluated by backward transfer

$$BWT_X(\hat{\beta}_T; \beta) = \sum_{t=1}^{T-1}\tilde{\omega}_t\big(R_X(\hat{\beta}_T; \beta, \Sigma_t) - R_X(\hat{\beta}_t; \beta, \Sigma_t)\big),$$

  where $\tilde{\omega}_t = n_t/\sum_{k=1}^{T-1} n_k$. Backward transfer on a sequence of tasks quantifies the average influence of learning new tasks on old tasks. If the backward transfer is less than 0, learning new tasks are beneficial for predicting on old tasks. Otherwise, learning new tasks in this task sequence produce performance reduction on old tasks.

- **Learning plasticity**, which is evaluated by forward transfer

$$FWT_X(\hat{\beta}_T;\beta)=\sum_{t=2}^{T}\bar{\omega}_t\big(R_X(\hat{\beta}_t;\beta,\Sigma_t)-R_X(\hat{\beta}_t^{\text{ridge}};\beta,\Sigma_t)\big),$$

  where $\bar{\omega}_t = n_t/\sum_{k=2}^{T} n_k$. Forward transfer on a sequence of tasks quantifies the average influence of historical information on learning current tasks. Note that the risk of continual ridge estimator measures the performance when learning with historical information, while the risk of ridge estimator at current step measures the performance when learning without historical information. If the forward transfer is less than 0, using historical information is beneficial for learning current tasks. Otherwise, a continual learning method is worse than simply learning current tasks.

From the preceding definitions, we observe that these evaluation metrics can be expressed as linear functions of the prediction risk associated with specific continual ridge estimators (the standard ridge estimator corresponds to the special case where the task number $T = 1$). Therefore, it remains to analyze the prediction risk of continual ridge estimators.

Note that the prediction risk has a bias-variance decomposition (Hastie et al., 2022),

$$R_X(\hat{\beta};\beta,\Sigma_0)=B_X(\hat{\beta};\beta,\Sigma_0)+V_X(\hat{\beta};\beta,\Sigma_0),$$

where

$$B_X(\hat{\beta};\beta,\Sigma_0)=\big(\mathbb{E}(\hat{\beta}|X)-\beta\big)^\top\Sigma_0\big(\mathbb{E}(\hat{\beta}|X)-\beta\big), \tag{4}$$

$$V_X(\hat{\beta};\beta,\Sigma_0)=\mathrm{Tr}\big(\mathrm{Cov}(\hat{\beta}|X)\Sigma_0\big). \tag{5}$$

To end this section, we derive the expressions of bias and variance terms for continual ridge estimators.

Lemma 1 *Under model (1), the bias and variance terms of continual ridge estimator (3) are respectively*

$$B_X(\hat{\beta}_T;\beta,\Sigma_0)=\beta^\top A_1A_2\cdots A_T\Sigma_0A_T\cdots A_2A_1\beta,$$

$$V_X(\hat{\beta}_T;\beta,\Sigma_0)=\sigma^2\sum_{t=1}^{T}\frac{1}{\lambda_t n_t}\mathrm{Tr}\big[A_TA_{T-1}\cdots A_{t+1}(A_t-A_t^2)A_{t+1}\cdots A_{T-1}A_T\Sigma_0\big].$$

The details of calculations above can be found in Appendix B.1.

3 Asymptotics of Prediction Risk

In this section, we investigate the asymptotic behaviors of the prediction risk for continual ridge estimators. The computation of high-dimensional asymptotic risk is intimately connected to the limiting spectral properties of sample covariance matrices. Random matrix theory (RMT) is a powerful tool to characterize these properties (Bai and Silverstein, 2010; Yao et al., 2015). To establish our theory, we begin by reviewing key concepts from RMT.

Let $A \in \mathbb{R}^{p\times p}$ be a symmetric matrix, then the empirical spectral distribution (ESD) of A is defined as $F_A(x) = p^{-1}\sum_{i=1}^{p}\mathbf{1}\{\lambda_i(A) \le x\}$. For any distribution F supported on $[0,\infty)$, its Stieltjes transform is defined as

$$m_F(z) = \int \frac{1}{x-z} dF(x), z \in \mathbb{C}\backslash[0,\infty).$$

If A is positive semi-definite, the Stieltjes transform of F_A is given by

$$m_A(z) = \int \frac{1}{x-z} dF_A(x) = \frac{1}{p}\mathrm{Tr}(A - zI_p)^{-1}, z \in \mathbb{C}\backslash[0,\infty).$$

Assume now that the data matrix is generated as $X = Z\Sigma^{1/2}$, where $\Sigma \in \mathbb{R}^{p\times p}$ is a deterministic positive semidefinite matrix, and $Z \in \mathbb{R}^{n\times p}$ has i.i.d. entries with zero mean and unit variance. The well-known Marčenko-Pastur theorem (Marčenko and Pastur, 1967; Silverstein and Bai, 1995) states that, if $p, n \to \infty$ such that $p/n \to \gamma \in (0,\infty)$, and F_Σ converges weakly to some limit distribution H, then the ESD of sample covariance matrix $\hat{\Sigma} = \frac{1}{n}X^\top X$ converges weakly to a limiting distribution $F_{H,\gamma}$. The characterization of $F_{H,\gamma}$ relies on its Stieltjes transform. Consider the corresponding matrix $\tilde{\Sigma} = \frac{1}{n}XX^\top$. By the definition of Stieltjes transform, $m_{\hat{\Sigma}}(z)$ and $m_{\tilde{\Sigma}}(z)$ are linked by

$$m_{\hat{\Sigma}}(z) = \frac{1}{\gamma} m_{\tilde{\Sigma}}(z) + \frac{1-\gamma}{\gamma z}.$$

Note that, the Marčenko-Pastur theorem also ensures that $F_{\tilde{\Sigma}}$ has a limiting distribution, denoted as $\tilde{F}_{H,\gamma}$. Let $m_{H,\gamma}(z)$ and $\tilde{m}_{H,\gamma}(z)$ be the Stieltjes transform of $F_{H,\gamma}$ and $\tilde{F}_{H,\gamma}$ respectively, then they are liked by

$$m_{H,\gamma}(z) = \frac{1}{\gamma}\tilde{m}_{H,\gamma}(z) + \frac{1-\gamma}{\gamma z}. \tag{6}$$

From the Marčenko-Pastur Theorem, $\tilde{m}_{H,\gamma}(z)$ is given by the unique solution of

$$\tilde{m}_{H,\gamma}(z) = \Big(-z + \gamma\int \frac{t}{1+\tilde{m}_{H,\gamma}(z)t} dH(t)\Big)^{-1}, (z, \tilde{m}_{H,\gamma}(z)) \in \mathbb{C}^+ \times \mathbb{C}^+. \tag{7}$$

In the special case where $\Sigma = I_p$ for all n, p, the Stieltjes transform of limiting distribution $m_\gamma(z)$ has an explicit form

$$m_\gamma(z) = \frac{-(1-\gamma-z) + \sqrt{(1-\gamma-z)^2 - 4\gamma z}}{-2\gamma z}. \tag{8}$$

Remark 2 *The Marčenko-Pastur Theorem only provides the expression of Stieltjes transform $m_{H,\gamma}(z)$ on $z \in \mathbb{C}^+$. To complete the definition of Stieltjes transform $m_{H,\gamma}(z)$ on $\mathbb{C}\backslash[0,\infty)$, we need to notice that (6) and (7) can be extended to $z \in \mathbb{C}\backslash[0,\infty)$ by continuity. The restriction $z \in \mathbb{C}^+$ makes sure that the equation (7) has unique solution (see, e.g., Yao et al. (2015)).*

3.1 The Main Theorem

To ensure the existence of the limiting risk in general settings, some technical assumptions are needed.

Assumption 1 *In each dataset $\mathcal{D}_t$, the sample matrix $X_t = Z_t\Sigma_t^{1/2}$, where Z_t has i.i.d. entries with zero mean, unit variance, and finite 16-th moment. Besides, the test data $x_0 = \Sigma_0^{1/2}z_0$, and z_0 has zero mean and unit variance.*

Assumption 2 *The task number T is a constant. In each dataset $\mathcal{D}_t$, $n_t, p \to \infty$ and $p/n_t \to \gamma_t \in (0, \infty)$.*

Assumption 3 *The signal $\|\beta\|^2 \to r^2 \in (0, \infty)$, as $p \to \infty$.*

Assumption 1 describes the data generation setting, Assumption 2 defines the asymptotic regime in each dataset, and Assumption 3 fixes the asymptotic scale of regression coefficients. These assumptions are classical in analysis of high-dimensional regression models. The last two assumptions establish the relationship between different tasks.

Assumption 4 *The covariance matrices $\{\Sigma_t, t = 1, \cdots, T\}$ and Σ_0 are commutable, and there exists two constants $c < C$ such that $0 < c \leq \lambda_{min}(\Sigma_t) \leq \lambda_{max}(\Sigma_t) \leq C < \infty$ for $t = 1, \cdots, T$ and $t = 0$.*

The commutability of covariance matrices in Assumption 4 ensures that, there exists an orthogonal matrix $U \in \mathbb{R}^{p\times p}$ such that $\Sigma_t = UD_tU^\top$ for $t = 1, \cdots, T$ and $\Sigma_0 = UD_0U^\top$, where $\{D_t\}_{t=1,\cdots,T}$ and D_0 are diagonal matrices. Let $\Sigma_t = \sum_{i=1}^p d_{t,i}u_iu_i^\top$ be the simultaneous eigenvalue decomposition of Σ_t for $t = 1, \cdots, T$ and $t = 0$, then we define the joint ESD as

$$H_n(x_1, \cdots, x_T, x_0) = \frac{1}{p}\sum_{i=1}^{p} \mathbf{1}\{d_{1,i} \leq x_1, \cdots, d_{T,i} \leq x_T, d_{0,i} \leq x_0\}, \tag{9}$$

and a weighted joint ESD as

$$G_n(x_1, \cdots, x_T, x_0) = \frac{1}{\|\beta\|^2}\sum_{i=1}^{p} \langle\beta, u_i\rangle^2 \mathbf{1}\{d_{1,i} \leq x_1, \cdots, d_{T,i} \leq x_T, d_{0,i} \leq x_0\}. \tag{10}$$

Here we note that, although $\{\Sigma_t, t = 1, \cdots, T\}$ and Σ_0 can be diagonalized simultaneously by different orthogonal matrix, the values of joint ESD (9) and weighted joint ESD (10) are unique regardless of the choice of orthogonal matrix U.

The next assumption describes the relationship between covariance matrices and the true coefficients in the limiting form.

Assumption 5 *The joint distributions H_n and G_n converge weakly to limit joint distributions H and G, respectively.*

By Assumption 5, the marginal ESDs F_{Σ_t} and F_{Σ_0} converges weakly to H_t and H_0, respectively, where H_t and H_0 are the marginal distributions of H with respect to x_t and x_0.

For $t = 1, \cdots, T$, define $\tilde{m}_t = \tilde{m}_{H_t,\gamma_t}(-\lambda_t)$, and

$$\mu_t = \Big[\Big(1 + \int \frac{\gamma_t s}{\lambda_t(1+\tilde{m}_t s)} dH_t(s)\Big)^2 - \int \frac{\gamma_t s^2}{\lambda_t^2(1+\tilde{m}_t s)^2} dH_t(s)\Big]^{-1},$$
$$a_t = \int \frac{s_t s_0}{\prod_{j=t}^T \lambda_j^2(1+\tilde{m}_j s_j)^2} dH(\mathbf{s}), \quad b_t = \int \frac{\lambda_t(1+\tilde{m}_t s_t)\cdot s_0}{\prod_{j=t}^T \lambda_j^2(1+\tilde{m}_j s_j)^2} dH(\mathbf{s}),$$
$$c_t = \int \frac{s_0}{\prod_{j=t}^T \lambda_j^2(1+\tilde{m}_j s_j)^2} dH(\mathbf{s}), \quad g_t = \int \frac{s_0}{\prod_{j=t}^T \lambda_j^2(1+\tilde{m}_j s_j)^2} dG(\mathbf{s}).$$

where $\mathbf{s} = (s_1, \cdots, s_T, s_0)$. For $1 \le t \le l \le T$, define

$$a_{t,l} = \int \frac{s_t s_l}{\prod_{j=t}^l \lambda_j^2(1+\tilde{m}_j s_j)^2} dH(\mathbf{s}), \quad b_{t,l} = \int \frac{\lambda_t(1+\tilde{m}_t s_t)\cdot s_l}{\prod_{j=t}^l \lambda_j^2(1+\tilde{m}_j s_j)^2} dH(\mathbf{s}),$$
$$c_{t,l} = \int \frac{s_l}{\prod_{j=t}^l \lambda_j^2(1+\tilde{m}_j s_j)^2} dH(\mathbf{s}), \quad g_{t,l} = \int \frac{s_l}{\prod_{j=t}^l \lambda_j^2(1+\tilde{m}_j s_j)^2} dG(\mathbf{s}).$$

Next, we define $\{\rho_t, 1 \le t \le T\}$ recursively as $\rho_T = a_T$, and

$$\rho_t = a_t + \sum_{j=t+1}^T \gamma_j \mu_j a_{t,j} \rho_j, \quad 1 \le t \le T-1.$$

Define $\{\rho_{s,t}^{(1)}, 1 \le s \le t \le T\}$ as $\rho_{s,s}^{(1)} = 0$ and

$$\rho_{s,t}^{(1)} = b_{s,t} + \sum_{j=s}^{t-1} \gamma_j \mu_j a_{j,t} \rho_{s,j}^{(1)}, \quad 1 \le s < t \le T.$$

Define $\{\rho_{s,t}^{(2)}, 1 \le s \le t \le T\}$ as $\rho_{s,s}^{(2)} = c_{s,s}$ and

$$\rho_{s,t}^{(2)} = c_{s,t} + \sum_{j=s}^{t-1} \gamma_j \mu_j a_{j,t} \rho_{s,j}^{(2)}, \quad 1 \le s < t \le T.$$

We are now ready to state our main theorem on the asymptotic prediction risk of continual ridge estimators in general settings.

Theorem 3 *Under Assumption 1-5, it holds almost surely that*

$$B_X(\hat{\beta}_T; \beta, \Sigma_0) \to \tilde{B}_T(r, \boldsymbol{\gamma}, \boldsymbol{\lambda}, G, H) := r^2\Big(\prod_{j=1}^T \lambda_j^2\Big)\Big(g_1 + \sum_{t=1}^T \gamma_t \mu_t \rho_t g_{1,t}\Big),$$
$$V_X(\hat{\beta}_T; \beta, \Sigma_0) \to \tilde{V}_T(\boldsymbol{\gamma}, \boldsymbol{\lambda}, H) := \sigma^2 \sum_{t=1}^T \gamma_t \Big(\prod_{s=t+1}^T \lambda_s^2\Big)(L_{1,t} - \lambda_t L_{2,t}),$$

and

$$R_X(\hat{\beta}_T;\beta,\Sigma_0)\to\tilde{R}_T(r,\boldsymbol{\gamma},\boldsymbol{\lambda},G,H):=\tilde{B}_T(r,\boldsymbol{\gamma},\boldsymbol{\lambda},G,H)+\tilde{V}_T(\boldsymbol{\gamma},\boldsymbol{\lambda},H),$$

where

$$L_{1,t}=b_t+\sum_{j=t}^{T}\gamma_j\mu_j a_j\rho_{t,j}^{(1)},\quad L_{2,t}=c_t+\sum_{j=t}^{T}\gamma_j\mu_j a_j\rho_{t,j}^{(2)}.$$

The proof of Theorem 3 is provided in Appendix B.3. In Theorem 3, the regularization parameters $\boldsymbol{\lambda}=(\lambda_1\cdots,\lambda_T)^\top$ are treated as constants independent on n and p. Note that the asymptotic bias term is related to the coefficients β and the asymptotic variance term is related to the variance of noise σ^2. As $n\to\infty$, the variance of noise is assumed to remain constant, while the dimension of regression coefficients is growing. The variation of β in this limiting process is characterized by the signal strength r^2 and the joint distribution G through g_1 and $g_{1,t}$. In addition, $\boldsymbol{\gamma}=(\gamma_1,\cdots,\gamma_T)^\top$ is intuitively interpreted as model complexity parameters, and the joint distribution H characterizes the relationship of covariance matrices in different tasks. These factors influence both the asymptotic bias term and the asymptotic variance term.

3.2 Examples

To get a better understanding of our main theorem, we focus on some concrete examples by considering different structures of covariance matrices in each task. Note that Assumption 4 guarantees that $\Sigma_1,\cdots,\Sigma_T$ and Σ_0 are simultaneously diagonalizable. Suppose $\Sigma_t=UD_tU^\top$ for $t=1,\cdots,T$ and $\Sigma_0=UD_0U^\top$, where $\{D_t\}_{t=1,\cdots,T}$ and D_0 are diagonal matrices, and U is an orthogonal matrix. Let $\mathring{Z}_t=Z_tU$, by rotational invariance of Z_t, the entries of $\mathring{Z}_t$ and Z_t have the same joint distribution, and $\mathring{Z}_1,\cdots,\mathring{Z}_T$ are independent. If we define $\mathring{X}_t=\mathring{Z}_tD_t^{1/2},\mathring{\Sigma}_t=\frac{1}{n_t}\mathring{X}_t^\top\mathring{X}_t$, and $\mathring{A}_t=\lambda_t(\mathring{\Sigma}_t+\lambda_tI_p)^{-1}$, we have

$$\hat{\Sigma}_t=\frac{1}{n_t}UD_t^{1/2}(Z_tU)^\top(Z_tU)D_t^{1/2}U^\top=U\mathring{\Sigma}_tU^\top,$$

and $A_t=U^\top\mathring{A}_tU$. Thus by Lemma 1,

$$\begin{aligned}B_X(\hat{\beta};\beta,\Sigma_0)&=(U\beta)^\top\mathring{A}_1\mathring{A}_2\cdots\mathring{A}_TD_0\mathring{A}_T\cdots\mathring{A}_2\mathring{A}_1(U\beta),\\V_X(\hat{\beta};\beta,\Sigma_0)&=\sigma^2\sum_{t=1}^{T}\frac{1}{\lambda_tn_t}\mathrm{Tr}\big[\mathring{A}_T\mathring{A}_{T-1}\cdots\mathring{A}_{t+1}(\mathring{A}_t-\mathring{A}_t^2)\mathring{A}_{t+1}\cdots\mathring{A}_{T-1}\mathring{A}_TD_0\big].\end{aligned}$$

From the above derivation, we observe that the prediction risk can be expressed in terms of diagonal covariance matrices and the transformed coefficients $\mathring{\beta}=U\beta$. Therefore, in the following cases, we assume that all the covariance matrices are diagonal matrices.

3.2.1 Identity Covariance Matrices

We first consider a trivial case where $\Sigma_1 = \cdots = \Sigma_T = \Sigma_0 = I_p$. In scenarios where task distributions are identical, the continual learning framework effectively reduces to an online learning framework, as no task-specific adaptation or catastrophic forgetting mitigation is required. Nevertheless, it remains important to investigate the properties of continual ridge estimation in this setting, since this algorithm is closely related to online learning methods. Taking the derivative in the updating rule (2), we have

$$\hat{\beta}_t = \hat{\beta}_{t-1} + \lambda^{-1} n_t^{-1} X_t^\top (X_t \hat{\beta}_t - y_t).$$

To enhance the updating efficiency, one may substitute $\hat{\beta}_{t-1}$ for $\hat{\beta}_t$ in the right-hand side of the equation, yielding

$$\hat{\beta}_t = \hat{\beta}_{t-1} - \eta n_t^{-1} X_t^\top (y_t - X_t \hat{\beta}_{t-1}),$$

which is the updating rule for online (batch) gradient descent method with the learning rate $\eta = \lambda^{-1}$ (see, e.g., Hoi et al. (2021)). Assuming computational efficiency is not a constraint, the updating rule of continual ridge regression seems to be a reasonable choice for online learning.

The following result shows the asymptotic risk in this setting, which has been largely simplified compared to the main theorem.

Theorem 4 *If $\Sigma_1 = \cdots = \Sigma_T = \Sigma_0 = I_p$, under Assumption 1-3, it holds almost surely that*

$$B_X(\hat{\beta}_T; \beta, \Sigma_0) \to \tilde{B}_T(r, \boldsymbol{\gamma}, \boldsymbol{\lambda}) := r^2 \prod_{t=1}^{T} \left[\lambda_t^2 m'_{\gamma_t}(-\lambda_t)\right],$$

$$V_X(\hat{\beta}_T; \beta, \Sigma_0) \to \tilde{V}_T(\boldsymbol{\gamma}, \boldsymbol{\lambda}) := \sigma^2 \sum_{t=1}^{T} \gamma_t \upsilon_t \prod_{s=t+1}^{T} \left[\lambda_s^2 m'_{\gamma_s}(-\lambda_s)\right],$$

and

$$R_X(\hat{\beta}_T; \beta, \Sigma_0) \to \tilde{R}_T(r, \boldsymbol{\gamma}, \boldsymbol{\lambda}) := \tilde{B}_T(r, \boldsymbol{\gamma}, \boldsymbol{\lambda}) + \tilde{V}_T(\boldsymbol{\gamma}, \boldsymbol{\lambda}),$$

where

$$\upsilon_t = m_{\gamma_t}(-\lambda_t) - \gamma_t m'_{\gamma_t}(-\lambda_t),$$

and m_{γ_t} has an explicit form as (8).

The proof of Theorem 4 is provided in Appendix B.2. The main difference between Theorem 4 and Theorem 3 is that, the asymptotic prediction risk of continual ridge estimation depends on β only through the limiting signal strength r^2 in the bias term.

Remark 5 *The asymptotic result of continual ridge regression coincides with that of classical high-dimensional ridge regression. When $T = 1$, the continual learning framework reduces to the ridge regularization. To describe the impact of the true parameter on the*

asymptotic risk, we define the limiting signal-to-noise ratio by $SNR = r^2/\sigma^2$. *According to our analysis, the asymptotic prediction risk is given by*

$$\tilde{R}(r,\gamma,\lambda) = \sigma^2\Big\{\lambda^2 m'_\gamma(-\lambda)SNR + \gamma[m_\gamma(-\lambda) - \gamma m'_\gamma(-\lambda)]\Big\},$$

which is consistent with existing results. (see, e.g., Dobriban and Wager (2018), Hastie et al. (2022)).

3.2.2 Isotropic Covariance Matrices with Different Scales

If the covariance matrices are different in each task, it may be hard to find a simplified consequence of limiting risk. Moreover, even the numeral calculation of the results in Theorem 3 is not a trivial task, unless the covariance matrices have simple structures. A natural idea is to assume that all the covariance matrices are isotropic, and the difference of tasks embody in the variation of scales. Specifically, we assume that $\Sigma_t = \delta_t I_p, t = 1, \cdots, T$ and $\Sigma_0 = \delta_0 I_p$, where $\{\delta_t\}_{t=1,\cdots,T}$ and δ_0 are different positive constants. In this setting, we have an explicit form of Stieltjes transform

$$m_t = \frac{-[\delta_t(1-\gamma_t)+\lambda_t] + \sqrt{[\delta_t(1-\gamma_t)+\lambda_t]^2 + 4\delta_t\gamma_t\lambda_t}}{2\gamma_t\delta_t\lambda_t},$$

and $\tilde{m}_t = m_t\gamma_t + \frac{1-\gamma_t}{\lambda_t}$. Furthermore, the elements of the expressions in Theorem 3 can be simplified as

$$\begin{aligned}
\mu_t &= \Big\{\big[1 + \frac{\gamma_t\delta_t}{\lambda_t(1+\tilde{m}_t\delta_t)}\big]^2 - \frac{\gamma_t\delta_t^2}{\lambda_t^2(1+\tilde{m}_t\delta_t)^2}\Big\}^{-1},\\
a_t &= \frac{\delta_t\delta_0}{\prod_{j=t}^T \lambda_j^2(1+\tilde{m}_j\delta_j)^2}, \quad a_{tl} = \frac{\delta_t\delta_l}{\prod_{j=t}^l \lambda_j^2(1+\tilde{m}_j\delta_j)^2},\\
b_t &= \frac{\lambda_t(1+\tilde{m}_t\delta_t)\delta_0}{\prod_{j=t}^T \lambda_j^2(1+\tilde{m}_j\delta_j)^2}, \quad b_{tl} = \frac{\lambda_t(1+\tilde{m}_t\delta_t)\delta_l}{\prod_{j=t}^l \lambda_j^2(1+\tilde{m}_j\delta_j)^2},\\
c_t &= \frac{\delta_0}{\prod_{j=t}^T \lambda_j^2(1+\tilde{m}_j\delta_j)^2}, \quad c_{tl} = \frac{\delta_l}{\prod_{j=t}^l \lambda_j^2(1+\tilde{m}_j\delta_j)^2}.
\end{aligned}$$

Since all covariance matrices are isotropic, any β with $\|\beta\|^2 = r^2$ satisfies Assumption 3 and Assumption 5, and it holds that $g_1 = c_1$ and $g_{1,t} = c_{1,t}$.

3.2.3 Covariance Matrices with Different Block Sizes

We finally focus on a fundamentally different setting where covariance matrices are not isotropic. For simplicity, we assume that each covariance matrix has two different eigenvalues, the first of which is δ and the second is 1. However, in different covariance matrices, the proportion of numbers of these two eigenvalues may be different. For $\Sigma_t \in \mathbb{R}^{p\times p}$, we suppose the first $\lfloor \pi_t p \rfloor$ eigenvalues are all δ ($\delta \neq 1$) and the others are 1, where $\pi_t \in (0,1)$ is a constant. Namely, $\Sigma_t = \mathrm{diag}\{\delta I_{\lfloor \pi_t p\rfloor}, I_{p-\lfloor \pi_t p\rfloor}\}, t = 1, \cdots, T$ and $t = 0$.

In this example, the joint ESD H_n converges weakly to a limiting distribution H. Let $\{q_1, \cdots, q_{T+1}\} = \{1, \cdots, T+1\}$ such that $\pi_{q_t} = \pi_{(t)}$, where $\pi_{(1)} \leq \pi_{(2)} \leq \cdots \leq \pi_{(T+1)}$ is

the ordered sequence of $\pi_1, \cdots, \pi_T, \pi_0$. Then the limiting distribution H satisfies

$$\int f(\mathbf{s}) dH(\mathbf{s}) = \sum_{t=0}^{T+1} (\pi_{(t+1)} - \pi_{(t)}) f(s_t), \tag{11}$$

where $\pi_{(0)} = 0, \pi_{(T+2)} = 1$, and

$$s_t = \delta(1, \cdots, 1)^\top + (1-\delta) \sum_{k=1}^{t} e_{q_k} \in \mathbb{R}^{T+1}, t = 0, 1, \cdots, T+1.$$

Here $e_1 = (1, 0, \cdots, 0)^\top, \cdots, e_{T+1} = (0, 0, \cdots, 1)^\top$.

By (11), we have

$$\mu_t = \Big\{ \big[1 + \frac{\pi_t \gamma_t \delta}{\lambda_t(1+\tilde{m}_t\delta)} + \frac{(1-\pi_t)\gamma_t}{\lambda_t(1+\tilde{m}_t)}\big]^2 - \big[\frac{\pi_t\gamma_t\delta^2}{\lambda_t^2(1+\tilde{m}_t\delta)^2} + \frac{(1-\pi_t)\gamma_t}{\lambda_t^2(1+\tilde{m}_t)^2}\big] \Big\}^{-1},$$

$$a_t = \sum_{k=0}^{T+1} (\pi_{(k+1)} - \pi_{(k)}) \frac{s_{k,t} s_{k,T+1}}{\prod_{j=t}^{T} \lambda_j^2 (1+\tilde{m}_j s_{k,j})^2}, a_{t,l} = \sum_{k=0}^{T+1} (\pi_{(k+1)} - \pi_{(k)}) \frac{s_{k,t} s_{k,l}}{\prod_{j=t}^{l} \lambda_j^2 (1+\tilde{m}_j s_{k,j})^2},$$

$$b_t = \sum_{k=0}^{T+1} (\pi_{(k+1)} - \pi_{(k)}) \frac{\lambda_t(1+\tilde{m}_t s_{k,t}) s_{k,T+1}}{\prod_{j=t}^{T} \lambda_j^2 (1+\tilde{m}_j s_{k,j})^2}, b_{t,l} = \sum_{k=0}^{T+1} (\pi_{(k+1)} - \pi_{(k)}) \frac{\lambda_t(1+\tilde{m}_t s_{k,t}) s_{k,l}}{\prod_{j=t}^{l} \lambda_j^2 (1+\tilde{m}_j s_{k,j})^2},$$

$$c_t = \sum_{k=0}^{T+1} (\pi_{(k+1)} - \pi_{(k)}) \frac{s_{k,T+1}}{\prod_{j=t}^{T} \lambda_j^2 (1+\tilde{m}_j s_{k,j})^2}, c_{t,l} = \sum_{k=0}^{T+1} (\pi_{(k+1)} - \pi_{(k)}) \frac{s_{k,l}}{\prod_{j=t}^{l} \lambda_j^2 (1+\tilde{m}_j s_{k,j})^2},$$

where $s_{k,t}$ is the t-th elements of $\vec{s_k}$. By (7), $\tilde{m}_t$ is determined by

$$\tilde{m}_t = \Big(\lambda_t + \frac{\pi_t\gamma_t\delta}{1+\tilde{m}_t\delta} + \frac{(1-\pi_t)\gamma_t}{1+\tilde{m}_t}\Big)^{-1}.$$

Unlike the isotropic case, it is tedious to derive the explicit expression of $\tilde{m}_t$. We may calculate $\tilde{m}_t$ using iterative methods.

The main difference between isotropic settings and the present case is that, the limiting risk may not exist for all β satisfying Assumption 3. The source of this problem is that G_n may not converge weakly to G for complex structures of covariance matrices. However, if the components of β are equally distributed along the eigenvectors of $\{\Sigma_t\}_{t=1,\cdots,T}$ (which are the same for different covariance matrices), it always holds that $G = H$, and the limiting risk exists regardless of the structure of covariance matrices. If β is fixed, a natural example is $\beta = \frac{r}{\sqrt{p}}(1, \cdots, 1)^\top$, since we assume that the covariance matrices are all diagonal. In more cases, β can be regarded as random regression coefficients (Dicker and Erdogdu, 2017; Dobriban and Sheng, 2020). If the prior distribution of β satisfies $\mathbb{E}\beta = 0$ and $\mathrm{Cov}(\beta) = \frac{r^2}{p} I_p$, then β is equally distributed and the limiting posterior risk exists. Anyhow, if β are equally distributed, we have $g_1 = c_1$ and $g_{1,t} = c_{1,t}$, and the limiting risk can be calculated by Theorem 3.

4 Experiments

In this section, we analyze the asymptotic risk curves for continual ridge estimators and verify our results by experiments. For convenience, we assume that each task has identical sample size n, so that $\gamma_1 = \cdots = \gamma_T = p/n$. Thereby, the asymptotics of our evaluation metrics for generalization performance are simplified to

$$\bar{R}_X(\hat{\beta};\beta) \to \frac{1}{T}\sum_{t=1}^{T}\tilde{R}_{T,t}, \quad a.s.,$$

$$BWT_X(\hat{\beta}_T;\beta) \to \frac{1}{T-1}\sum_{t=1}^{T-1}(\tilde{R}_{T,t}-\tilde{R}_{t,t}), \quad a.s.,$$

$$FWT_X(\hat{\beta}_T;\beta) \to \frac{1}{T-1}\sum_{t=2}^{T}(\tilde{R}_{t,t}-\tilde{R}_t^{\text{ridge}}), \quad a.s.,$$

where $\tilde{R}_{T,t}$ and $\tilde{R}_t^{\text{ridge}}$ denote the almost sure limits of $R_X(\hat{\beta}_T;\beta,\Sigma_t)$ and $R_X(\hat{\beta}_t^{\text{ridge}};\beta,\Sigma_t)$ from Theorem 3, respectively.

In each experiment, we plot the theoretical risk curves using Theorem 3. Meanwhile, we calculate the evaluation metrics above by simulated data and compare the results with the theoretical risk curves. We consider the data generated settings mentioned in section 3.2. In each setting, we are interested in the relationship between the evaluation metrics and the task number T under different model complexity (quantified by $\gamma = p/n$). Besides, we also consider the influence of the choice of regularization parameters $\boldsymbol{\lambda}$ on generalization performance.

4.1 Tuning Parameters Selection

Before turning to explicit cases, we focus on the choice of regularization parameters $\boldsymbol{\lambda} = (\lambda_1, \cdots, \lambda_T)^\top$. From (2), we know that the choice of λ_t is related to the balance between the knowledge from $\mathcal{D}_t$ and previous tasks. If λ_t is too large, the estimation is not adapted to the current task. Alternatively, if λ_t is too small, the catastrophic forgetting may happen on previous tasks. Intuitively, there exists an optimal choice, and the oracle optimal regularization parameter is the minima of the risk function. However, this is obviously not a reasonable choice in continual learning procedures, because the risk function contains information from all steps, which is unavailable when we choose regularization parameters at each step before future tasks arrive. In other words, a reasonable strategy of parameter selection may determine λ_t at step t. If we write $\tilde{R}_t = t^{-1}\sum_{s=1}^{t}\tilde{R}_{t,s}$ as the average risk function at time t, we may choose $\boldsymbol{\lambda} = (\lambda_1, \cdots, \lambda_T)^\top$ entrywise by

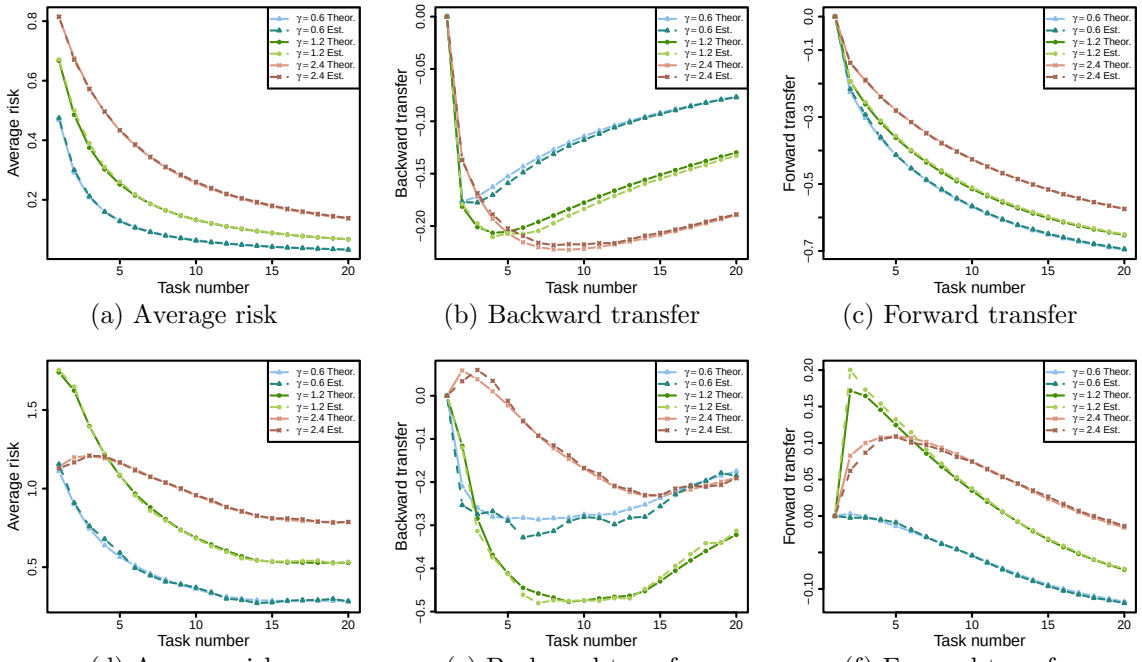
$$\lambda_t = \underset{\lambda>0}{\arg\min}\, \tilde{R}_t(\lambda_1, \cdots, \lambda_{t-1}, \lambda), \quad t = 1, \cdots, T. \tag{12}$$

It is a greedy selection strategy, since we choose the (oracle) presently optimal parameter at each step. Although there is no evidence to show that the present choice is optimal for the final average risk function, this strategy is somehow acceptable if there is no prior information about future tasks at present time.

Remark 6 *In practice, the regularization parameters may be chosen from a data-driven criterion, but standard techniques, such as cross-validation, may be invalid in continual learning frameworks. Similar as the determination of oracle parameters, it is important to notice that a reasonable criterion chooses λ_t using the information at time t. Specifically, only the dataset $\mathcal{D}_t$ and the estimator $\hat{\beta}_{t-1} = \hat{\beta}_{t-1}(\lambda_1, \cdots, \lambda_{t-1})$ are available, where $\lambda_1, \cdots, \lambda_{t-1}$ are determined at previous steps. Designing such a criterion is beyond the scope of this article.*

4.2 Experimental Results

In each of the following experiments, we set the maximum task number $T = 20$, sample size $n = 100$ and the parameter dimension $p = \lfloor n\gamma \rfloor$, where γ is the metric of model complexity. We consider three levels of model complexity: $\gamma = 0.6, 1.2, 2.4$. In each experiment, we assume that the noise is Gaussian with variance $\sigma^2 = 1$. The signal strength r^2 is set to be 1, and the true parameter β is generated as $\beta = \frac{r}{\sqrt{p}}(\beta_1, \cdots, \beta_p)^\top$, where β_k is randomly sampled by $\mathbf{P}(\beta_k = 1) = \mathbf{P}(\beta_k = -1) = 1/2$. For the choice of regularization parameters, the standard setting $\boldsymbol{\lambda}_{st}$ follows from (12), while a contrast setting is designed as $\boldsymbol{\lambda}_{st}/20$, which means there exists catastrophic forgetting in the learning procedure. Each experiment is repeated $B = 100$ times and we present the average results as estimation curves. The asymptotic results are presented as theoretical curves in each plot.



(a) Average risk (b) Backward transfer (c) Forward transfer

(d) Average risk (e) Backward transfer (f) Forward transfer

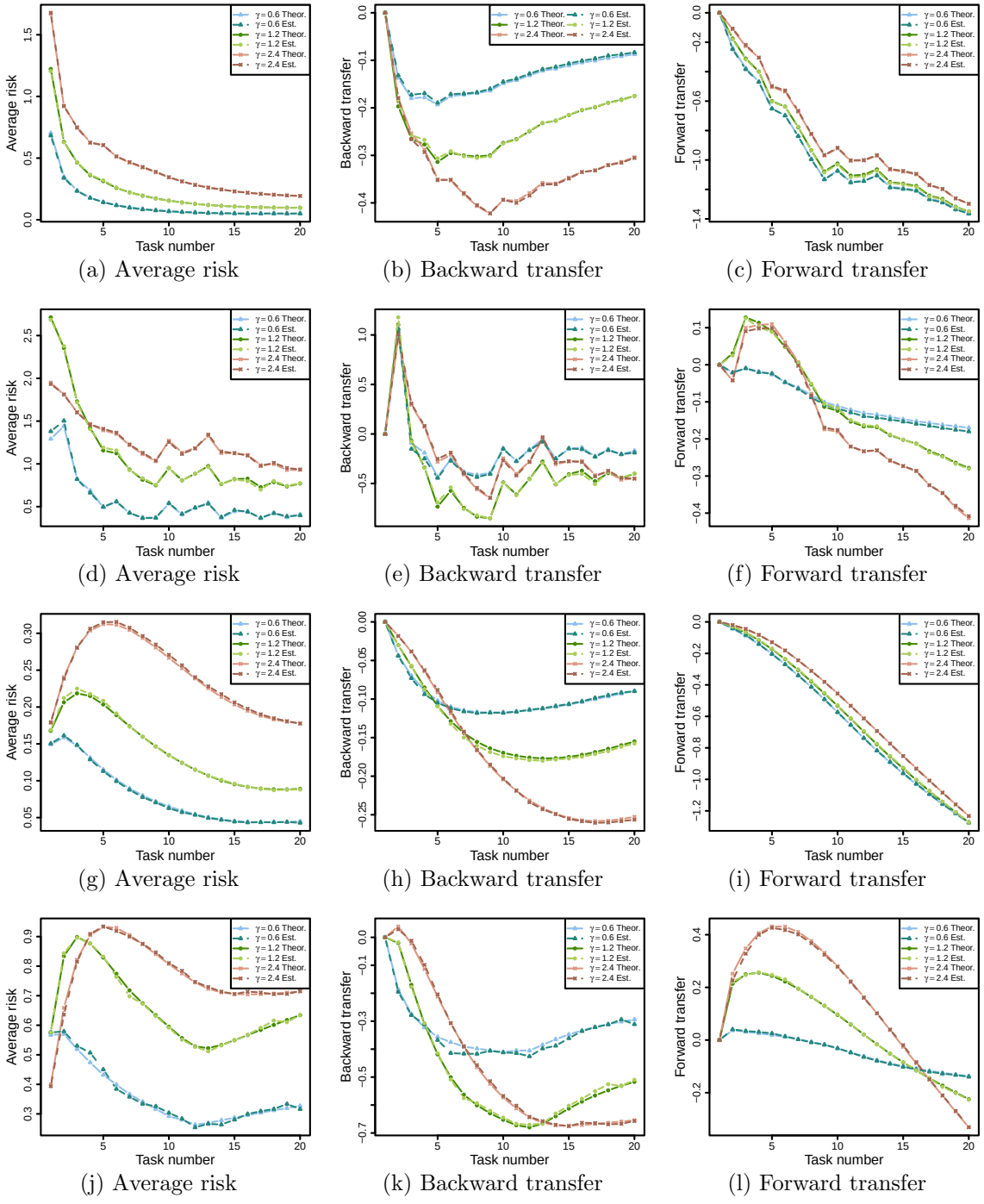
Figure 1: Risk curves with respect to the task number: The covariance matrices are all identity matrices. In the first row, the regularization parameters are $\boldsymbol{\lambda} = \boldsymbol{\lambda}_{st}$, while in the second row, $\boldsymbol{\lambda} = \boldsymbol{\lambda}_{st}/20$.

Identity covariance matrices In this example, we consider a setting where the covariance matrices are identical across all tasks. Under this assumption, increasing the number of tasks would be equivalent to increasing the sample size if catastrophic forgetting is disregarded. Therefore, one would expect the average risk to decrease monotonically as more tasks are observed. The simulation results, presented in Figure 1, confirm this behavior when the regularization parameters are appropriately tuned. Moreover, the results demonstrate that continual ridge estimation exhibits both forward transfer and backward transfer capabilities, and the dominant transfer shifts from backward transfer to forward transfer as the task number increases. However, when regularization parameters are set too small, continual ridge estimation has significant performance reduction, manifested on larger average risk and delayed emergence of forward transfer.

Isotropic covariance matrices with different scales In this example, we consider $\Sigma_t = \delta_t I_p, t = 1, \cdots T$, where $\{\delta_t\}_{t=1,\cdots T}$ are different positive constants. The parameters $\boldsymbol{\delta} = (\delta_1, \cdots, \delta_T)^\top$ describe the patterns of covariance shift. We investigate two distinct mechanisms: (1) random shift: $\delta_1, \cdots, \delta_T \sim i.i.d.U(0.5, 3.5)$; (2) increasing shift: $\delta_t = 4t/(T+1)$. The simulation results are presented in Figure 2. When the regularization parameters are appropriately tuned, we observe that the average risk decreases monotonically under the random shift mechanism, while rising to a peak and then decreasing under the increasing shift mechanism. This behavior suggests that the continual ridge estimation identifies the trend of covariance shift in the early steps, then leveraging the acquired knowledge to enhance overall performance. Besides, the performance of transfer capacities are similar to the first example. When regularization parameters are set too small, we explore that the performance of continual estimator is significantly disturbed by randomness of covariance shift, and the forward transfer behaves poorly since the trends of covariance shift are not identified effectively.

Covariance matrices with different block sizes In the last example, we consider $\Sigma_t = \mathrm{diag}\{\delta I_{\lfloor \pi_t p \rfloor}, I_{p-\lfloor \pi_t p \rfloor}\}, t = 1, \cdots T$, where $\{\pi_t\}_{t=1,\cdots T} \in [0, 1]$ and $\delta > 1$ are positive constants. The parameters $\boldsymbol{\pi} = (\pi_1, \cdots, \pi_T)^\top$ describe the variation of block size, and δ describes the scale of the main block. In this experiment we set $\delta = 5$. We also investigate two distinct mechanisms: (1) random block size: $\pi_1, \cdots, \pi_T \sim i.i.d.U(0, 1)$; (2) increasing block size: $\pi_t = t/T$. The simulation results are presented in Figure 3. When the regularization parameters are appropriately tuned, the average risk decreases nearly monotonically for both covariance block mechanisms, and the performance of transfer capacities are similar to the first two examples. In the under-regularized setting, there is also performance reduction on average risk and forward transfer due to catastrophic forgetting.

Our experimental results show that the performance of continual risk regression is highly sensitive to regularization parameters. With properly chosen regularization parameters, the average risk decreases nearly monotonically when the number of tasks is large enough, and the forward transfer becomes increasingly pronounced over time, providing that the task-specific covariance matrices exhibit reasonable structures. Conversely, inappropriate regularization parameters lead to performance reduction, resulting in an unstable relationship between the number of tasks and both average risk and transfer capacities.



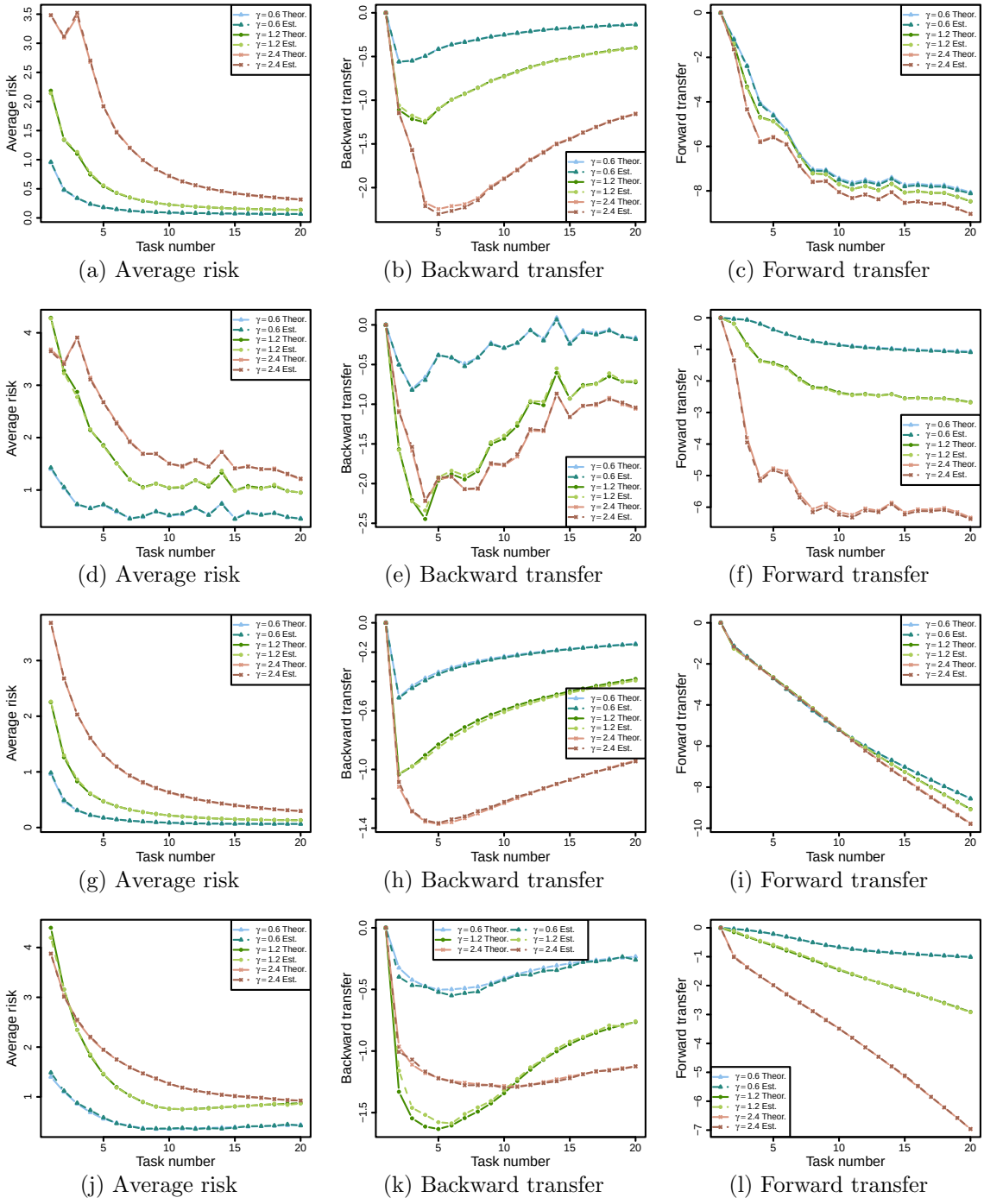
(a) Average risk (b) Backward transfer (c) Forward transfer

(d) Average risk (e) Backward transfer (f) Forward transfer

(g) Average risk (h) Backward transfer (i) Forward transfer

(j) Average risk (k) Backward transfer (l) Forward transfer

Figure 2: Risk curves with respect to the task number: In the first and second row, the covariance matrices satisfy random shift mechanism with $\boldsymbol{\lambda} = \boldsymbol{\lambda}_{st}$ and $\boldsymbol{\lambda} = \boldsymbol{\lambda}_{st}/20$ respectively. In the third and fourth row, the covariance matrices satisfy increasing shift mechanism with $\boldsymbol{\lambda} = \boldsymbol{\lambda}_{st}$ and $\boldsymbol{\lambda} = \boldsymbol{\lambda}_{st}/20$ respectively.



(a) Average risk (b) Backward transfer (c) Forward transfer

(d) Average risk (e) Backward transfer (f) Forward transfer

(g) Average risk (h) Backward transfer (i) Forward transfer

(j) Average risk (k) Backward transfer (l) Forward transfer

Figure 3: Risk curves with respect to the task number: In the first and second row, the covariance matrices have random block size with $\boldsymbol{\lambda} = \boldsymbol{\lambda}_{st}$ and $\boldsymbol{\lambda} = \boldsymbol{\lambda}_{st}/20$ respectively. In the third and fourth row, the covariance matrices have increasing block size with $\boldsymbol{\lambda} = \boldsymbol{\lambda}_{st}$ and $\boldsymbol{\lambda} = \boldsymbol{\lambda}_{st}/20$ respectively.

5 Conclusions

In this paper, we establish a rigorous theoretical framework for analyzing continual ridge regression in high-dimensional linear models with random designs. More specifically, we derive exact asymptotic expressions for prediction risk, explicitly characterizing its dependence on model complexity and task similarity. Average risk, backward transfer, and forward transfer are formalized to evaluate stepwise generalization performance, with their asymptotic behaviors related to asymptotic prediction risk. Through three representative examples, we demonstrate how the performance of risk curves vary with structures of task-specific covariance and choices of regularization parameters, supported by numerical simulations. Our results enrich theoretical analysis for continual learning methods, offering interpretable insights into the interplay of dimensionality, task relationships, and regularization in mitigating catastrophic forgetting.

An important direction for future work is to establish a data-driven choice for regularization parameters as well as its theoretical guarantees. Our present result can help evaluate the data-driven tuning strategies by providing the interpretable results for oracle regularization parameters. Additionally, the techniques in this work may be helpful for theoretical analysis on continual learning in nonlinear model, which is also considered in our future research.

Acknowledgments and Disclosure of Funding

This research was partially supported by the National Natural Science Foundation of China (Grant No. 12271286). Part of Wenqing Su's work was conducted at Tsinghua University during the author's doctoral studies.

Appendix A. Technical Results for Main Proofs

A.1 Deterministic Equivalent

Before proving the main theorem, we introduce a technique of deterministic equivalence in random matrix theory (Hachem et al., 2007; Couillet et al., 2011).

Definition 7 *Let $A_n, B_n \in \mathbb{R}^{n\times n}$ be sequences of random or deterministic symmetric random matrices, we say A_n, B_n are equivalent, if for any sequence of deterministic matrices $C_n \in \mathbb{R}^{n\times n}$ such that*

$$\limsup_n \|C_n\|_{op} < \infty,$$

and for any sequence of deterministic vectors $v_n \in \mathbb{R}^n$ such that

$$\limsup_n \|v_n\| < \infty,$$

we have, almost surely,

$$\frac{1}{n}\mathrm{Tr}[C_n(A_n - B_n)] \to 0, \quad v_n^\top (A_n - B_n)v_n \to 0.$$

Here the operator norm is defined by $\|A\|_{op} := \sqrt{\lambda_{max}(A^\top A)}$ for any real matrix $A \in \mathbb{R}^{p\times q}$.

We write $A_n \asymp B_n$ if A_n, B_n are equivalent. For random matrix A_n, if there exists a (sequence of) deterministic matrix $\bar{A}_n$ such that $A_n \asymp \bar{A}_n$, then $\bar{A}_n$ is said to be a deterministic equivalent for A_n.

There are some useful rules of calculus for matrix equivalents.

Proposition 8
1. If $A_n \asymp B_n, C_n \asymp D_n$, then $A_n + C_n \asymp B_n + D_n$.
2. If $A_n \asymp B_n$, E_n is a sequence of deterministic matrix such that $\|E_n\|_{op} < \infty$, then $E_n A_n E_n^\top \asymp E_n B_n E_n^\top$.
3. If $A_n \asymp B_n$, then $\frac{1}{n}\mathrm{Tr}[A_n - B_n] \to 0$, almost surely.

Remark 9 *The notion of matrix equivalent varies in different literature. For instance, Dobriban and Sheng (2021) relaxed the restriction that A_n, B_n are symmetric, and instead required that*

$$\mathrm{Tr}[\Theta_n(A_n - B_n)] \to 0$$

almost surely for any Θ_n satisfying

$$\limsup_n \|\Theta_n\|_{tr} < \infty, \tag{13}$$

where the trace norm is defined by $\|A\|_{tr} = \mathrm{Tr}((A^\top A)^{1/2})$ for any matrix A. This condition is stronger than ours since $\Theta_n = \frac{1}{n}C_n$ and $\Theta_n = v_n v_n^\top$ are all special cases of (13). In our work, we are mainly interested in the traces and quadratic forms of symmetric random matrices, therefore, our requirement in the definition of matrix equivalent is enough for analysis. Yet, the rules of calculus in Proposition 8 are commonly true for any reasonable definition of matrix equivalent.

After introducing the concept of matrix equivalents, the Marčenko-Pastur Theorem about the sample covariance matrix $\hat{\Sigma}$ can be modified as the form of deterministic equivalent for resolvent matrix $Q(z) = (\hat{\Sigma} - zI_p)^{-1}$.

Theorem 10 (Theorem 2.6, Couillet and Liao (2022)) *Assume $X = Z\Sigma^{1/2}$, where $Z \in \mathbb{R}^{n\times p}$ has i.i.d. entries with zero mean, unit variance and finite eighth-order moment. Suppose $\Sigma \in \mathbb{R}^{p\times p}$ is a deterministic positive semi-definite matrix with bounded operator norm, then as $n, p \to \infty$ with $p/n \to \gamma \in (0, \infty)$, we have*

$$Q(z) \asymp \bar{Q}(z) := -\frac{1}{z}(I_p + \tilde{m}_{\gamma,p}(z)\Sigma)^{-1}, \tag{14}$$

where $\tilde{m}_{\gamma,p}(z)$ is the unique solution of

$$\tilde{m}_{\gamma,p}(z) = \Big(-z + \mathrm{Tr}\big[\Sigma(I_p + \tilde{m}_{\gamma,p}(z)\Sigma)^{-1}\big]\Big), \quad (z, \tilde{m}_{\gamma,p}(z)) \in \mathbb{C}^+ \times \mathbb{C}^+, \tag{15}$$

Define $\{\tilde{m}_{\gamma,p}(z), z \in (-\infty, 0)\}$ by (15) and the continuity of $\tilde{m}_{\gamma,p}(z)$, then (14) is true for $z \in (-\infty, 0)$. Besides, if ESD F_Σ converges weakly to some limit distribution H, then as $p \to \infty$, we have $\tilde{m}_{\gamma,p}(z) \to \tilde{m}_{H,\gamma}(z)$, where $\tilde{m}_{H,\gamma}(z)$ is defined by the unique solution of

$$\tilde{m}_{H,\gamma}(z) = \Big(-z + \gamma \int \frac{t}{1 + \tilde{m}_{H,\gamma(z)}t} dH(t)\Big)^{-1}, \quad (z, \tilde{m}_{\gamma,p}(z)) \in \mathbb{C}^+ \times \mathbb{C}^+. \tag{16}$$

Here we note that, if $\Sigma = I_p$, the results of Theorem 10 reduce to

$$Q(z) \asymp \bar{Q}(z) := m_\gamma(z) I_p, \tag{17}$$

where $m_\gamma(z)$ is the Stieltjes transform of limiting ESD. It is the unique solution of equation

$$z\gamma m_\gamma(z) - (1-\gamma-z)m_\gamma(z) + 1 = 0, \quad (z, \tilde{m}_{\gamma,p}(z)) \in \mathbb{C}^+ \times \mathbb{C}^+, \tag{18}$$

which coincides with (8).

In the rest of this subsection, we provide a deterministic equivalent for the quadratic form of resolvent matrix. This result will be applied repeatedly to deal with the products of resolvent matrix, which frequently appear in the expressions of bias and variance terms. For proof convenience, we consider $z \in (-\infty, 0)$ for the resolvent matrix $Q(z)$.

Theorem 11 *Assume $X = Z\Sigma^{1/2}$, where $Z \in \mathbb{R}^{n\times p}$ has i.i.d. entries with zero mean, unit variance and finite 16th-order moment. Let $\lambda > 0$ be a fixed real number, $Q(-\lambda) = (\frac{1}{n}X^\top X + \lambda I_p)^{-1}$ be the resolvent matrix, $A \in \mathbb{R}^{p\times p}$ is a symmetric deterministic matrix or symmetric random matrix independent of $Q(-\lambda)$, and having bounded operator norm. Suppose $\Sigma \in \mathbb{R}^{p\times p}$ is a deterministic positive semi-definite matrix with bounded operator norm, then as $n, p \to \infty$ with $p/n \to \gamma \in (0, \infty)$, we have*

$$Q(-\lambda)AQ(-\lambda) \asymp \bar{Q}A\bar{Q} + \frac{1}{n}\mathrm{Tr}\big[\Sigma\bar{Q}A\bar{Q}\big] \cdot \Big\{\big[1 + \frac{1}{n}\mathrm{Tr}(\Sigma\bar{Q})\big]^2 - \frac{1}{n}\mathrm{Tr}(\Sigma\bar{Q})^2\Big\}^{-1} \cdot \bar{Q}\Sigma\bar{Q}, \tag{19}$$

where $\bar{Q} := \bar{Q}(-\lambda)$ is defined in Theorem 10.

If we take $\Sigma = I_p$, the deterministic equivalent of resolvent matrix becomes (17), thus the following result is immediately derived from Theorem 11.

Corollary 12 *Assume $X \in \mathbb{R}^{n\times p}$ has i.i.d. entries with zero mean, unit variance and finite 16th-order moment. Let $\lambda > 0$ be a fixed real number, $Q(-\lambda) = (\frac{1}{n}X^\top X + \lambda I_p)^{-1}$ be the resolvent matrix, $A \in \mathbb{R}^{p\times p}$ is a symmetric deterministic matrix or a symmetric random matrix independent of $Q(-\lambda)$, and $\|A\|_{op}$ is bounded by a constant. Then as $n, p \to \infty$ with $p/n \to \gamma \in (0, \infty)$, we have*

$$Q(-\lambda)AQ(-\lambda) \asymp m_\gamma^2(-\lambda)A + \frac{1}{n}\mathrm{Tr}A \cdot \frac{m_\gamma'(-\lambda)m_\gamma^2(-\lambda)}{(1+\gamma m_\gamma(-\lambda))^2} I_p. \tag{20}$$

Taking $A = I_p$, the equation above becomes

$$Q^2(-\lambda) \asymp m_\gamma'(-\lambda) I_p. \tag{21}$$

A.2 Proof of Theorem 11

For the convenience of notation, we write $Q := Q(-\lambda)$, $\bar{Q} := \bar{Q}(-\lambda)$ and $\tilde{m} := \tilde{m}_{\gamma,p}(-\lambda)$. For a sequence of deterministic matrices $A_n \in \mathbb{R}^{n\times n}$, we write $A_n = o_{\|\cdot\|}(1)$ if $\|A_n\|_{op} \to 0$ as $n \to \infty$. The proof of this result can be divided into two steps:
i) Show that QAQ is concentrated in mean, that is, $QAQ \asymp \mathbb{E}(QAQ)$.
ii) Find a deterministic matrix R such that $\mathbb{E}(QAQ) = R + o_{\|\cdot\|}(1)$, therefore $\mathbb{E}(QAQ) \asymp R$.

Step 1: Concentration of QAQ

To prove that $QAQ \asymp \mathbb{E}(QAQ)$, it suffices to show that

$$\frac{1}{p}\mathrm{Tr}\big[C\big(QAQ - \mathbb{E}(QAQ)\big)\big] \to 0, a.s.$$

for $C \in \mathbb{R}^{p\times p}$ having bounded operator norm, and

$$v^\top\big(QAQ - \mathbb{E}(QAQ)\big)v \to 0, a.s.$$

for $v \in \mathbb{R}^p$ having bounded norm.

Similar to the proof in Lemma 15, we write

$$\frac{1}{p}\mathrm{Tr}\big[C\big(QAQ - \mathbb{E}(QAQ)\big)\big] = \frac{1}{p}\sum_{i=1}^{n}(\mathbb{E}_{\le i} - \mathbb{E}_{\le i-1})\mathrm{Tr}\big[C\big(QAQ - Q_{-i}AQ_{-i}\big)\big].$$

By (40), we have

$$\begin{aligned}\frac{1}{p}\mathrm{Tr}\big[C\big(QAQ - Q_{-i}AQ_{-i}\big)\big] =& -\frac{1}{pn}\mathrm{Tr}\big[CQ_{-i}A\frac{Q_{-i}x_ix_i^\top Q_{-i}}{1+\frac{1}{n}x_i^\top Q_{-i}x_i} + C\frac{Q_{-i}x_ix_i^\top Q_{-i}}{1+\frac{1}{n}x_i^\top Q_{-i}x_i}AQ_{-i}\big]\\ &+\frac{1}{pn^2}\mathrm{Tr}\big[C\frac{Q_{-i}x_ix_i^\top Q_{-i}AQ_{-i}x_ix_i^\top Q_{-i}}{(1+\frac{1}{n}x_i^\top Q_{-i}x_i)^2}\big].\end{aligned}$$

Let $Y_i = Y_{i,1} + Y_{i,2}$, where

$$\begin{aligned}Y_{i,1} &= -(\mathbb{E}_{\le i} - \mathbb{E}_{\le i-1})\frac{1}{pn}\mathrm{Tr}\Big[C\Big(Q_{-i}A\frac{Q_{-i}x_ix_i^\top Q_{-i}}{1+\frac{1}{n}x_i^\top Q_{-i}x_i} + \frac{Q_{-i}x_ix_i^\top Q_{-i}}{1+\frac{1}{n}x_i^\top Q_{-i}x_i}AQ_{-i}\Big)\Big],\\ Y_{i,2} &= (\mathbb{E}_{\le i} - \mathbb{E}_{\le i-1})\frac{1}{pn^2}\mathrm{Tr}\Big[C\frac{Q_{-i}x_ix_i^\top Q_{-i}AQ_{-i}x_ix_i^\top Q_{-i}}{(1+\frac{1}{n}x_i^\top Q_{-i}x_i)^2}\Big],\end{aligned}$$

Using the matrix inequality $A^\top B + B^\top A \preceq A^\top A + B^\top B$ and note that $\|Q_{-i}\|_{op} \le \lambda^{-1}$, we have

$$\begin{aligned}&\Big|\frac{1}{pn}\mathrm{Tr}\Big[C\Big(Q_{-i}A\frac{Q_{-i}x_ix_i^\top Q_{-i}}{1+\frac{1}{n}x_i^\top Q_{-i}x_i} + \frac{Q_{-i}x_ix_i^\top Q_{-i}}{1+\frac{1}{n}x_i^\top Q_{-i}x_i}AQ_{-i}\Big)\Big]\Big|\\ \le&\frac{\lambda^{-2}}{pn}\|C\|_{op}\Big\|A\frac{Q_{-i}x_ix_i^\top}{1+\frac{1}{n}x_i^\top Q_{-i}x_i} + \frac{x_ix_i^\top Q_{-i}}{1+\frac{1}{n}x_i^\top Q_{-i}x_i}A\Big\|_{tr}\\ \le&\frac{\lambda^{-2}}{pn}\|C\|_{op}\Big\|x_ix_i^\top + \frac{AQ_{-i}x_ix_i^\top Q_{-i}A}{1+\frac{1}{n}x_i^\top Q_{-i}x_i}\Big\|_{tr}\\ \le&\frac{\lambda^{-2}}{pn}\|C\|_{op}(\|x_i\|^2 + x_i^\top Q_{-i}A^2Q_{-i}x_i)\\ \le&\frac{\lambda^{-2}}{pn}\|C\|_{op}(1+\lambda^{-2}\|A\|_{op}^2)\|\Sigma\|_{op}\|z_i\|^2,\end{aligned}$$

where $x_i = \Sigma^{1/2} z_i$. For $k \geq 2$, if the entries of Z has k-th moment, we have $\mathbb{E}\|z_i\|^{2k} = O(n^k)$, therefore,

$$\mathbb{E}_{\leq i-1}\Big|\frac{1}{pn}\mathrm{Tr}\Big[C\Big(Q_{-i}A\frac{Q_{-i}x_ix_i^\top Q_{-i}}{1+\frac{1}{n}x_i^\top Q_{-i}x_i}+\frac{Q_{-i}x_ix_i^\top Q_{-i}}{1+\frac{1}{n}x_i^\top Q_{-i}x_i}AQ_{-i}\Big)\Big]\Big|^k = O(n^{-2k})\|z_i\|^{2k},$$

$$\mathbb{E}_{\leq i}\Big|\frac{1}{pn}\mathrm{Tr}\Big[C\Big(Q_{-i}A\frac{Q_{-i}x_ix_i^\top Q_{-i}}{1+\frac{1}{n}x_i^\top Q_{-i}x_i}+\frac{Q_{-i}x_ix_i^\top Q_{-i}}{1+\frac{1}{n}x_i^\top Q_{-i}x_i}AQ_{-i}\Big)\Big]\Big|^k = O(n^{-2k})\mathbb{E}\|z_i\|^{2k} = O(n^{-k}),$$

and

$$|Y_{i,1}|^k = 2^{k-1}\big(O(n^{-2k})\|z_i\|^{2k} + O(n^{-k})\big), \quad \mathbb{E}|Y_{i,1}|^k = O(n^{-k}).$$

For $Y_{i,2}$, note that

$$\begin{aligned}\Big|\frac{1}{pn^2}\mathrm{Tr}\Big[C\frac{Q_{-i}x_ix_i^\top Q_{-i}AQ_{-i}x_ix_i^\top Q_{-i}}{(1+\frac{1}{n}x_i^\top Q_{-i}x_i)^2}\Big]\Big| \leq& \frac{1}{pn^2}\|C\|_{op}\|A\|_{op}\Big\|\frac{Q_{-i}x_ix_i^\top Q_{-i}Q_{-i}x_ix_i^\top Q_{-i}}{(1+\frac{1}{n}x_i^\top Q_{-i}x_i)^2}\Big\|_{tr}\\ \leq& \frac{\lambda^{-2}}{pn^2}\|C\|_{op}\|A\|_{op}\frac{(x_i^\top Q_{-i}x_i)^2}{(1+\frac{1}{n}x_i^\top Q_{-i}x_i)^2}\\ \leq& \frac{\lambda^{-2}}{p}\|C\|_{op}\|A\|_{op} = O(n^{-1}),\end{aligned}$$

thus $|Y_{i,2}| = O(n^{-1})$. By Lemma 14, we have

$$\begin{aligned}\mathbb{E}\Big|\frac{1}{p}\mathrm{Tr}\big[C\big(QAQ-\mathbb{E}(QAQ)\big)\big]\Big|^k =& \mathbb{E}\Big|\sum_{i=1}^n Y_i\Big|^k \leq C_k\Big(\mathbb{E}\big(\sum_{i=1}^n \mathbb{E}_{\leq i-1}|Y_i|^2\big)^{k/2} + \sum_{i=1}^n \mathbb{E}|Y_i|^k\Big)\\ \leq& C_k\Big(\mathbb{E}\big[\sum_{i=1}^n 2\big(\mathbb{E}_{\leq i-1}|Y_{i,1}|^2 + \mathbb{E}_{\leq i-1}|Y_{i,2}|^2\big)\big]^{k/2} + 2^{k-1}\sum_{i=1}^n\big(\mathbb{E}|Y_{i,1}|^k + \mathbb{E}|Y_{i,2}|^k\big)\Big)\\ \leq& 2^{k-1}C_k\Big(\mathbb{E}\big[\sum_{i=1}^n \mathbb{E}_{\leq i-1}|Y_{i,1}|^2\big]^{k/2} + \mathbb{E}\big[\sum_{i=1}^n \mathbb{E}_{\leq i-1}|Y_{i,2}|^2\big]^{k/2} + \sum_{i=1}^n\big(\mathbb{E}|Y_{i,1}|^k + \mathbb{E}|Y_{i,2}|^k\big)\Big)\\ =& O(n^{-k/2}).\end{aligned}$$

Taking $k > 2$, by Markov's inequality and Borel-Cantelli lemma, we get

$$\frac{1}{p}\mathrm{Tr}\big[C\big(QAQ-\mathbb{E}(QAQ)\big)\big] \to 0, \quad a.s.$$

To prove that $v^\top\big(QAQ-\mathbb{E}(QAQ)\big)v \to 0, a.s.$, we write

$$v^\top\big(QAQ-\mathbb{E}(QAQ)\big)v = \sum_{i=1}^n(\mathbb{E}_{\leq i}-\mathbb{E}_{\leq i-1})v^\top\big(QAQ-Q_{-i}AQ_{-i}\big)v.$$

By (40), we have

$$\begin{aligned}v^\top\big(QAQ-Q_{-i}AQ_{-i}\big)v =& -\frac{1}{n}v^\top Q_{-i}A\frac{Q_{-i}x_ix_i^\top Q_{-i}}{1+\frac{1}{n}x_i^\top Q_{-i}x_i}v - \frac{1}{n}v^\top\frac{Q_{-i}x_ix_i^\top Q_{-i}}{1+\frac{1}{n}x_i^\top Q_{-i}x_i}AQ_{-i}v\\ &+\frac{1}{n^2}v^\top\frac{Q_{-i}x_ix_i^\top Q_{-i}AQ_{-i}x_ix_i^\top Q_{-i}}{(1+\frac{1}{n}x_i^\top Q_{-i}x_i)^2}v.\end{aligned}$$

Let $\tilde{Y}_i = Y_{i,3} + Y_{i,4}$, where

$$Y_{i,3} = -(\mathbb{E}_{\le i} - \mathbb{E}_{\le i-1})\frac{1}{n} v^\top \Big(Q_{-i} A \frac{Q_{-i} x_i x_i^\top Q_{-i}}{1 + \frac{1}{n} x_i^\top Q_{-i} x_i} + \frac{Q_{-i} x_i x_i^\top Q_{-i}}{1 + \frac{1}{n} x_i^\top Q_{-i} x_i} A Q_{-i} \Big) v,$$

$$Y_{i,4} = (\mathbb{E}_{\le i} - \mathbb{E}_{\le i-1})\frac{1}{n^2} v^\top \frac{Q_{-i} x_i x_i^\top Q_{-i} A Q_{-i} x_i x_i^\top Q_{-i}}{(1 + \frac{1}{n} x_i^\top Q_{-i} x_i)^2} v,$$

Using the matrix inequality $A^\top B + B^\top A \preceq A^\top A + B^\top B$ and note that $\|Q_{-i}\|_{op} \le \lambda^{-1}$, we have

$$\begin{aligned}
&\Big| \frac{1}{n} v^\top \Big(Q_{-i} A \frac{Q_{-i} x_i x_i^\top Q_{-i}}{1 + \frac{1}{n} x_i^\top Q_{-i} x_i} + \frac{Q_{-i} x_i x_i^\top Q_{-i}}{1 + \frac{1}{n} x_i^\top Q_{-i} x_i} A Q_{-i} \Big) v \Big| \\
\le & \frac{\lambda^{-2}}{n} \Big| v^\top \Big(A \frac{Q_{-i} x_i x_i^\top}{1 + \frac{1}{n} x_i^\top Q_{-i} x_i} + \frac{x_i x_i^\top Q_{-i}}{1 + \frac{1}{n} x_i^\top Q_{-i} x_i} A \Big) v \Big| \\
\le & \frac{\lambda^{-2}}{n} \Big| v^\top \Big(x_i x_i^\top + \frac{A Q_{-i} x_i x_i^\top Q_{-i} A}{1 + \frac{1}{n} x_i^\top Q_{-i} x_i} \Big) v \Big| \\
\le & \frac{\lambda^{-2}}{n} (1 + \lambda^{-2} \|A\|_{op}^2) \|\Sigma\|_{op} |v^\top z_i|^2.
\end{aligned}$$

Let $\mathbb{E} z_{11}^4 = m_4, \mathbb{E} z_{11}^{2k} = m_{2k}$, by Lemma 13, we have

$$\mathbb{E} \Big| z_i^\top v v^\top z_i - \|v\|^2 \Big|^k \le C_k \Big[(m_4 \|v\|^2)^{k/2} + m_{2k} (\|v\|^2)^{k/2} \Big] = O(1),$$

$$\mathbb{E} |v^\top z_i|^{2k} \le 2^k \Big(\mathbb{E} \Big| z_i^\top v v^\top z_i - \|v\|^2 \Big|^k + \|v\|^{2k} \Big) = O(1),$$

thus for $k \ge 2$, if the entries of Z has $2k$-th moment, we have

$$\mathbb{E}_{\le i-1} \Big| \frac{1}{n} v^\top \Big(Q_{-i} A \frac{Q_{-i} x_i x_i^\top Q_{-i}}{1 + \frac{1}{n} x_i^\top Q_{-i} x_i} + \frac{Q_{-i} x_i x_i^\top Q_{-i}}{1 + \frac{1}{n} x_i^\top Q_{-i} x_i} A Q_{-i} \Big) v \Big|^k = O(n^{-k}) |v^\top z_i|^{2k},$$

$$\mathbb{E}_{\le i} \Big| \frac{1}{n} v^\top \Big(Q_{-i} A \frac{Q_{-i} x_i x_i^\top Q_{-i}}{1 + \frac{1}{n} x_i^\top Q_{-i} x_i} + \frac{Q_{-i} x_i x_i^\top Q_{-i}}{1 + \frac{1}{n} x_i^\top Q_{-i} x_i} A Q_{-i} \Big) v \Big|^k = O(n^{-k}) \mathbb{E} |v^\top z_i|^{2k} = O(n^{-k}),$$

and

$$|Y_{i,3}|^k = 2^{k-1} \big(O(n^{-k}) |v^\top z_i|^{2k} + O(n^{-k}) \big), \quad \mathbb{E} |Y_{i,3}|^k = O(n^{-k}).$$

For $Y_{i,4}$, note that

$$\begin{aligned}
\Big| \frac{1}{n^2} v^\top \frac{Q_{-i} x_i x_i^\top Q_{-i} A Q_{-i} x_i x_i^\top Q_{-i}}{(1 + \frac{1}{n} x_i^\top Q_{-i} x_i)^2} v \Big| \le & \frac{\lambda^{-2}}{n^2} \|A\|_{op} \Big| \frac{x_i^\top Q_{-i}^2 x_i}{(1 + \frac{1}{n} x_i^\top Q_{-i} x_i)^2} v^\top x_i x_i^\top v \Big| \\
\le & \frac{\lambda^{-3}}{n} \|A\|_{op} \|\Sigma\|_{op} |v^\top z_i|^2,
\end{aligned}$$

thus

$$\mathbb{E}_{\leq i-1}\Big|\frac{1}{n^2}v^\top\frac{Q_{-i}x_ix_i^\top Q_{-i}AQ_{-i}x_ix_i^\top Q_{-i}}{(1+\frac{1}{n}x_i^\top Q_{-i}x_i)^2}v\Big|^k=O(n^{-k})|v^\top z_i|^{2k},$$

$$\mathbb{E}_{\leq i}\Big|\frac{1}{n^2}v^\top\frac{Q_{-i}x_ix_i^\top Q_{-i}AQ_{-i}x_ix_i^\top Q_{-i}}{(1+\frac{1}{n}x_i^\top Q_{-i}x_i)^2}v\Big|^k=O(n^{-k}),$$

and

$$|Y_{i,4}|^k=2^{k-1}\big(O(n^{-k})|v^\top z_i|^{2k}+O(n^{-k})\big),\quad \mathbb{E}|Y_{i,4}|^k=O(n^{-k}).$$

By Lemma 14, we have

$$\begin{aligned}\mathbb{E}\Big|v^\top\big(QAQ-\mathbb{E}(QAQ)\big)v\Big|^k=&\mathbb{E}\Big|\sum_{i=1}^n\tilde{Y}_i\Big|^k\leq C_k\Big(\mathbb{E}\big(\sum_{i=1}^n\mathbb{E}_{\leq i-1}|\tilde{Y}_i|^2\big)^{k/2}+\sum_{i=1}^n\mathbb{E}|\tilde{Y}_i|^k\Big)\\ \leq&2^{k-1}C_k\Big(\mathbb{E}\big[\sum_{i=1}^n\mathbb{E}_{\leq i-1}|Y_{i,3}|^2\big]^{k/2}+\mathbb{E}\big[\sum_{i=1}^n\mathbb{E}_{\leq i-1}|Y_{i,4}|^2\big]^{k/2}+\sum_{i=1}^n\big(\mathbb{E}|Y_{i,3}|^k+\mathbb{E}|Y_{i,4}|^k\big)\Big)\\ =&O(n^{-k/2}).\end{aligned}$$

Taking $k>2$, by Markov's inequality and Borel-Cantelli lemma, we get

$$v^\top\big(QAQ-\mathbb{E}(QAQ)\big)v\to 0,\quad a.s.$$

Step 2: Finding Deterministic Matrix

Note that

$$\begin{aligned}\mathbb{E}[QAQ]=&\mathbb{E}[QA\bar{Q}]+\mathbb{E}[QA(Q-\bar{Q})]\\ =&\bar{Q}A\bar{Q}+\mathbb{E}[QAQ(I_p-Q^{-1}\bar{Q})]\\ =&\bar{Q}A\bar{Q}+\mathbb{E}\big[QAQ\big(I_p-(\frac{1}{n}X^\top X+\lambda I_p)\bar{Q}\big)\big]\\ =&\bar{Q}A\bar{Q}+\mathbb{E}\big[QAQ(\bar{Q}^{-1}-\lambda I_p-\frac{1}{n}X^\top X)\bar{Q}\big]\\ =&\bar{Q}A\bar{Q}+\mathbb{E}\big[QAQ(\lambda\tilde{m}\Sigma-\frac{1}{n}X^\top X)\bar{Q}\big]\\ =&\bar{Q}A\bar{Q}+\mathbb{E}[QAQ]\frac{\Sigma\bar{Q}}{1+\frac{1}{n}\mathrm{Tr}(\Sigma\bar{Q})}-\mathbb{E}[QAQ\cdot\frac{1}{n}X^\top X]\bar{Q},\end{aligned}\tag{22}$$

the last equality holds since $\frac{1}{n}\mathrm{Tr}(\Sigma\bar{Q})=-(\frac{1}{-\lambda\tilde{m}}+1)$ from (15). Similarly,

$$\begin{aligned}\mathbb{E}[QAQ]=&\mathbb{E}[\bar{Q}AQ]+\mathbb{E}[(Q-\bar{Q})AQ]\\ =&\bar{Q}A\bar{Q}+\frac{\Sigma\bar{Q}}{1+\frac{1}{n}\mathrm{Tr}(\Sigma\bar{Q})}\mathbb{E}[QAQ]-\bar{Q}\mathbb{E}[\frac{1}{n}X^\top X\cdot QAQ].\end{aligned}\tag{23}$$

Let $Q_{-i} = (\frac{1}{n}\sum_{j\neq i} x_j x_j^\top - zI_p)^{-1}$ be the resolvent matrix without sample x_i. By Sherman-Morrison formula, we have

$$Q = Q_{-i} - \frac{\frac{1}{n}Q_{-i}x_i x_i^\top Q_{-i}}{1+\frac{1}{n}x_i^\top Q_{-i}x_i}$$

and

$$Qx_i = \frac{Q_{-i}x_i}{1+\frac{1}{n}x_i^\top Q_{-i}x_i}, \tag{24}$$

thus

$$\mathbb{E}[QAQ\cdot\frac{1}{n}X^\top X]\bar{Q} = \frac{1}{n}\sum_{i=1}^n \mathbb{E}\Big[\frac{QAQ_{-i}x_i x_i^\top}{1+\frac{1}{n}x_i^\top Q_{-i}x_i}\Big]\bar{Q}. \tag{25}$$

Let $\alpha = \frac{1}{n}\mathrm{Tr}[\mathbb{E}(\Sigma Q_{-i})]$, which is a constant for all i. Then we have

$$\begin{aligned}\frac{QAQ_{-i}x_i x_i^\top}{1+\frac{1}{n}x_i^\top Q_{-i}x_i} =& \frac{1}{1+\alpha}\Big[Q_{-i}x_i x_i^\top - \frac{Q_{-i}x_i x_i^\top(\frac{1}{n}x_i^\top Q_{-i}x_i-\alpha)}{1+\frac{1}{n}x_i^\top Q_{-i}x_i}\Big]\\ =& \frac{1}{1+\alpha}\Big[Q_{-i}x_i x_i^\top - Qx_i x_i^\top(\frac{1}{n}x_i^\top Q_{-i}x_i-\alpha)\Big].\end{aligned} \tag{26}$$

Let $d_i = \frac{1}{n}x_i^\top Q_{-i}x_i - \alpha$, $D = \mathrm{diag}\{d_i\}_{i=1}^n$, by (26), we obtain

$$\begin{aligned}\mathbb{E}[QAQ\cdot\frac{1}{n}X^\top X]\bar{Q} =& \frac{1}{(1+\alpha)n}\sum_{i=1}^n \mathbb{E}[QAQ_{-i}x_i x_i^\top - QAQx_i d_i x_i^\top]\bar{Q}\\ =& \frac{1}{(1+\alpha)n}\sum_{i=1}^n \mathbb{E}[QAQ_{-i}x_i x_i^\top]\bar{Q} - \frac{1}{(1+\alpha)n}\mathbb{E}[QAQX^\top DX]\bar{Q},\end{aligned} \tag{27}$$

and similarly,

$$\bar{Q}\mathbb{E}[\frac{1}{n}X^\top X\cdot QAQ] = \frac{1}{(1+\alpha)n}\sum_{i=1}^n \bar{Q}\mathbb{E}[x_i x_i^\top Q_{-i}AQ] - \frac{1}{(1+\alpha)n}\bar{Q}\mathbb{E}[X^\top DXQAQ], \tag{28}$$

We wish to control the operator norm of $\Delta := \frac{1}{n}\mathbb{E}[QAQX^\top DX]\bar{Q}$ plus its transport. Using the matrix inequality $A^\top B + B^\top A \preceq A^\top A + B^\top B$, we have

$$\begin{aligned}\Big\|\Delta+\Delta^\top\Big\| &= \frac{1}{n}\Big\|\mathbb{E}[n^{\epsilon/2}QAQX^\top D\cdot n^{-\epsilon/2}X\bar{Q} + n^{-\epsilon/2}\bar{Q}X^\top\cdot n^{\epsilon/2}(QAQX^\top D)^\top]\Big\|\\ &\le \frac{1}{n}\Big\|\mathbb{E}[n^{\epsilon}QAQX^\top D^2XQAQ + n^{-\epsilon}\bar{Q}X^\top X\bar{Q}]\Big\|\\ &\le n^{-1+\epsilon}\|\mathbb{E}[QAQX^\top D^2XQAQ]\| + n^{-1-\epsilon}\|\mathbb{E}[\bar{Q}X^\top X\bar{Q}]\|\\ &\le n^{-1+\epsilon}\|\mathbb{E}[\|D\|^2QAQX^\top XQAQ]\| + n^{-1-\epsilon}\|\mathbb{E}[\bar{Q}X^\top X\bar{Q}]\|\\ &\le C_1 n^{\epsilon}\mathbb{E}\|D\|^2 + C_2 n^{-\epsilon}\|\Sigma\|,\end{aligned}$$

the last inequality holds since $\|A\|$ is bounded by a constant, $\|\frac{1}{n}QX^\top X\| = \|I_p - \lambda Q\|$ is bounded by 2, $\|Q\| \le \lambda^{-1}$, and$\|\bar{Q}\| \le \|EQ\| + \|\bar{Q} - EQ\| \le \lambda^{-1} + o(1)$.

To control $n^\epsilon \mathbb{E}\|D\|^2$, we notice that

$$\begin{aligned} n^\epsilon \mathbb{E}\|D\|^2 =& n^\epsilon \mathbb{E} \max_i d_i^2 = n^\epsilon \int_0^\infty \mathbf{P}(\max_i d_i^2 > t) dt \\ \le& n^\epsilon \int_0^{n^{-\theta-\epsilon}} \mathbf{P}(\max_i d_i^2 > t) dt + n^\epsilon \int_{n^{-\theta-\epsilon}}^\infty n\mathbf{P}(d_1^2 > t) dt \\ \le& n^{-\theta} + n^{1+\epsilon} \int_{n^{-\theta-\epsilon}}^\infty \mathbf{P}(d_1^2 > t) dt. \end{aligned}$$

Since

$$\begin{aligned} \mathbf{P}(d_1^2 > t) =& \mathbf{P}\Big(\Big| \frac{1}{n} z_1^\top \Sigma^{1/2} Q_{-1} \Sigma^{1/2} z_1 - \frac{1}{n} \mathrm{Tr}[\mathbb{E}(\Sigma Q_{-1})] \Big|^2 > t \Big) \\ \le& \frac{1}{t^4} \mathbb{E} \Big| \frac{1}{n} z_1^\top \Sigma^{1/2} Q_{-1} \Sigma^{1/2} z_1 - \frac{1}{n} \mathrm{Tr}[\mathbb{E}(\Sigma Q_{-1})] \Big|^8 \\ \le& \frac{2^7}{t^4 n^8} \Big\{ \mathbb{E} \Big| z_1^\top \Sigma^{1/2} Q_{-1} \Sigma^{1/2} z_1 - \mathrm{Tr}[\Sigma Q_{-1}] \Big|^4 + \mathbb{E} \Big| \mathrm{Tr}[\Sigma (Q_{-1} - \mathbb{E} Q_{-1})] \Big|^8 \Big\}, \end{aligned}$$

taking $k = 8$ in Lemma 13, we have

$$\mathbb{E} \Big| z_1^\top \Sigma^{1/2} Q_{-1} \Sigma^{1/2} z_1 - \mathrm{Tr}[\Sigma Q_{-1}] \Big|^8 = \mathbb{E}_{-1} \mathbb{E}_1 \Big| z_1^\top \Sigma^{1/2} Q_{-1} \Sigma^{1/2} z_1 - \mathrm{Tr}[\Sigma^{1/2} Q_{-1} \Sigma^{1/2}] \Big|^8 \le C p^4$$

for some $C > 0$, and taking $k = 8$ in Lemma 15, we have

$$\mathbb{E} \Big| \frac{1}{p} \mathrm{Tr}[\Sigma (Q_{-1} - \mathbb{E} Q_{-1})] \Big|^8 = O(n^{-4}),$$

thus $\mathbf{P}(d_1^2 > t) \le C t^{-4} n^{-4}$ for some $C > 0$, and

$$n^\epsilon \mathbb{E}\|D\|^2 \le n^{-\theta} + C n^{4\epsilon + 3\theta - 3}.$$

Choose $\epsilon = \theta = 3/8$, we have $n^\epsilon \mathbb{E}\|D\|^2 = O(n^{-3/8})$, and

$$\Big\| \frac{1}{n} \mathbb{E}[QAQX^\top DX] \bar{Q} \Big\| = O(n^{-3/8}),$$

therefore, (27) and (28) implies

$$\begin{aligned} & \mathbb{E}[QAQ \cdot \frac{1}{n} X^\top X] \bar{Q} + \bar{Q} \mathbb{E}[\frac{1}{n} X^\top X \cdot QAQ] \\ =& \frac{1}{(1+\alpha)n} \sum_{i=1}^n \Big\{ \mathbb{E}[QAQ_{-i} x_i x_i^\top] \bar{Q} + \bar{Q} \mathbb{E}[x_i x_i^\top Q_{-i} AQ] \Big\} + o_{\|\cdot\|}(1). \end{aligned}$$

Applying Sherman-Morrison formula (40) and using the same technique as (26), we have

$$\begin{aligned} & \frac{1}{(1+\alpha)n} \sum_{i=1}^n \mathbb{E}[QAQ_{-i} x_i x_i^\top] \bar{Q} \\ =& \frac{1}{(1+\alpha)n} \Big\{ \sum_{i=1}^n \mathbb{E}[Q_{-i} A Q_{-i} x_i x_i^T \bar{Q}] - \sum_{i=1}^n \frac{Q_{-i} x_i n^{-1} x_i^\top Q_{-i} A Q_{-i} x_i x_i^\top \bar{Q}}{1 + \frac{1}{n} x_i^\top Q_{-i} x_i} \Big\} \\ =& \Delta_1 - \Delta_2, \end{aligned}$$

and

$$\mathbb{E}[QAQ \cdot \frac{1}{n} X^\top X]\bar{Q} + \bar{Q}\mathbb{E}[\frac{1}{n} X^\top X \cdot QAQ] = (\Delta_1 + \Delta_1^\top) - (\Delta_2 + \Delta_2^\top),$$

where

$$\Delta_1 = \frac{1}{(1+\alpha)n} \sum_{i=1}^n \mathbb{E}[Q_{-i} A Q_{-i} x_i x_i^T]\bar{Q},$$

$$\begin{aligned}\Delta_2 =& \frac{1}{(1+\alpha)^2 n} \sum_{i=1}^n \mathbb{E}[Q_{-i} x_i n^{-1} x_i^\top Q_{-i} A Q_{-i} x_i x_i^\top]\bar{Q} \\ & - \frac{1}{(1+\alpha)^2 n} \sum_{i=1}^n \mathbb{E}[Q_{-i} x_i n^{-1} x_i^\top Q_{-i} A Q x_i d_i x_i^\top]\bar{Q},\end{aligned}$$

since Q_{-i} and x_i are independent, we have

$$\begin{aligned}\Delta_1 =& \frac{1}{(1+\alpha)} \sum_{i=1}^n \mathbb{E}_{-i}[Q_{-i} A Q_{-i}] \mathbb{E}_i[n^{-1} x_i x_i^T]\bar{Q} \\ =& \frac{1}{(1+\alpha)} \sum_{i=1}^n \mathbb{E}_{-i}[Q_{-i} A Q_{-i}]\Sigma\bar{Q} \\ =& \frac{1}{1+\alpha}\mathbb{E}[QAQ]\Sigma\bar{Q} + o_{\|\cdot\|}(1).\end{aligned}$$

To analyze $\Delta_2 + \Delta_2^\top$, we first control the norm of $\tilde{\Delta} := \frac{1}{n}\sum_{i=1}^n \mathbb{E}[Q_{-i} x_i n^{-1} x_i^\top Q_{-i} A Q x_i d_i x_i^\top]\bar{Q}$ plus its transport. Substituting $Q_{-i} x_i$ by (24), we have

$$Q_{-i} x_i n^{-1} x_i^\top Q_{-i} A Q x_i d_i x_i^\top = Q x_i n^{-1} x_i^\top Q A Q x_i d_i x_i^\top \times (1 + \frac{1}{n} x_i^\top Q_{-i} x_i)^2.$$

Let $\tilde{d}_i = d_i \cdot (1 + \frac{1}{n} x_i^\top Q_{-i} x_i)^2 = d_i(1 + \alpha + d_i)^2$, $\tilde{D} = \mathrm{diag}\{|\tilde{d}_i|\}_{i=1}^n$, using similar matrix inequality techniques, we have

$$\begin{aligned}\left\|\tilde{\Delta} + \tilde{\Delta}^\top\right\| =& \left\|\frac{1}{n}\sum_{i=1}^n \mathbb{E}[Q x_i n^{-1} x_i^\top Q A Q x_i \tilde{d}_i x_i^\top]\bar{Q} + \bar{Q}[Q x_i n^{-1} x_i^\top Q A Q x_i \tilde{d}_i x_i^\top]^\top\right\| \\ \leq& \left\|\frac{1}{n}\mathbb{E}[n^{\epsilon/2}\frac{1}{n} Q X^\top X Q A Q X^\top \tilde{D} \cdot n^{-\epsilon/2} X\bar{Q}] + \mathbb{E}[n^{\epsilon/2}\frac{1}{n} Q X^\top X Q A Q X^\top \tilde{D} \cdot n^{-\epsilon/2} X\bar{Q}]^\top\right\| \\ \leq& C_1 n^\epsilon \mathbb{E}\|\tilde{D}\|^2 + C_2 n^{-\epsilon}\|\Sigma\|,\end{aligned}$$

and

$$n^\epsilon \mathbb{E}\|\tilde{D}\|^2 \leq n^{-\theta} + n^{1+\epsilon}\int_{n^{-\theta-\epsilon}}^\infty \mathbf{P}(\tilde{d}_1^2 > t)dt.$$

Since

$$\begin{aligned}\mathbf{P}(\tilde{d}_1^2 > t) \leq& \mathbf{P}\big(8d_1^2[(1+|\alpha|)^4 + d_1^4] > t\big)\\ \leq& \mathbf{P}\big(8d_1^2[(1+|\alpha|)^4 + d_1^4] > t, d_1^2 < (1+|\alpha|)^2\big) + \mathbf{P}\big(8d_1^2[(1+|\alpha|)^4 + d_1^4] > t, d_1^2 \geq (1+|\alpha|)^2\big)\\ \leq& \mathbf{P}\big(16d_1^2(1+|\alpha|)^4 > t\big) + \mathbf{P}\big(16d_1^6 > t\big)\\ \leq& C_1 t^{-4} n^{-4} + C_2 t^{-4/3} n^{-4},\end{aligned}$$

we have

$$n^{\epsilon}\mathbb{E}\|D\|^2 \leq n^{-\theta} + C_1 n^{4\epsilon+3\theta-3} + C_2 n^{\frac{4}{3}\epsilon+\frac{1}{3}\theta-3}.$$

Choose $\theta = \epsilon = 3/8$, we have $n^{\epsilon}\mathbb{E}\|D\|^2 = O(n^{-3/8})$, and

$$\Big\|\frac{1}{n}\mathbb{E}[QAQX^\top DX]\bar{Q}\Big\| = O(n^{-3/8}),$$

therefore,

$$\begin{aligned}\Delta_2 + \Delta_2^\top =& \frac{1}{(1+\alpha)^2 n}\sum_{i=1}^{n}\Big\{\mathbb{E}[Q_{-i}x_i n^{-1}x_i^\top Q_{-i}AQ_{-i}x_i x_i^\top]\bar{Q}\\ &+ \bar{Q}\mathbb{E}[Q_{-i}x_i n^{-1}x_i^\top Q_{-i}AQ_{-i}x_i x_i^\top]^\top\Big\} + o_{\|\cdot\|}(1).\end{aligned}$$

For the first term of $\Delta_2 + \Delta_2^\top$, note that

$$\begin{aligned}&\mathbb{E}[Q_{-i}x_i n^{-1}x_i^\top Q_{-i}AQ_{-i}x_i x_i^\top]\\ =&\mathbb{E}_{-i}\big[Q_{-i}\mathbb{E}_i[x_i n^{-1}x_i^\top Q_{-i}AQ_{-i}x_i x_i^\top]\big]\\ =&\mathbb{E}_{-i}\big[Q_{-i}\mathbb{E}_i[\frac{1}{n}\Sigma^{1/2}z_i z_i^\top \Sigma^{1/2}Q_{-i}AQ_{-i}\Sigma^{1/2}z_i z_i^\top\Sigma^{1/2}]\big]\\ =&\mathbb{E}_{-i}\big[Q_{-i}\Sigma^{1/2}[\frac{2}{n}\Sigma^{1/2}Q_{-i}AQ_{-i}\Sigma^{1/2} + \frac{1}{n}\mathrm{Tr}(\Sigma Q_{-i}AQ_{-i})I_p]\Sigma^{1/2}\big]\\ =&\mathbb{E}_{-i}\big[Q_{-i}\Sigma\cdot\frac{1}{n}\mathrm{Tr}(\Sigma Q_{-i}AQ_{-i})\big] + o_{\|\cdot\|}(1),\end{aligned}$$

thus

$$\begin{aligned}\Delta_2 + \Delta_2^\top =& \frac{1}{(1+\alpha)^2}\Big\{\mathbb{E}\big[Q\Sigma\cdot\frac{1}{n}\mathrm{Tr}(\Sigma QAQ)\big]\bar{Q} + \bar{Q}\big[Q\Sigma\cdot\frac{1}{n}\mathrm{Tr}(\Sigma QAQ)\big]^\top\Big\} + o_{\|\cdot\|}(1)\\ =&\frac{2}{(1+\alpha)^2}\cdot\frac{1}{n}\mathrm{Tr}(\mathbb{E}[\Sigma QAQ])\bar{Q}\Sigma\bar{Q} + o_{\|\cdot\|}(1).\end{aligned}$$

Combining the results above and noticing that

$$\alpha = \frac{1}{n}\mathrm{Tr}[\mathbb{E}(\Sigma Q_{-i})] = \frac{1}{n}\mathrm{Tr}[\mathbb{E}(\Sigma Q)] + o(1) = \frac{1}{n}\mathrm{Tr}[\Sigma\bar{Q}] + o(1),$$

we have

$$\begin{aligned}&\mathbb{E}[QAQ\cdot\frac{1}{n}X^\top X]\bar{Q} + \bar{Q}\mathbb{E}[\frac{1}{n}X^\top X\cdot QAQ]\\ =&\frac{\mathbb{E}[QAQ]\Sigma\bar{Q}}{1+\frac{1}{n}\mathrm{Tr}[\Sigma\bar{Q}]} + \frac{\Sigma\bar{Q}\mathbb{E}[QAQ]}{1+\frac{1}{n}\mathrm{Tr}[\Sigma\bar{Q}]} - \frac{2}{(1+\frac{1}{n}\mathrm{Tr}[\Sigma\bar{Q}])^2}\cdot\frac{1}{n}\mathrm{Tr}(\mathbb{E}[\Sigma QAQ])\bar{Q}\Sigma\bar{Q} + o_{\|\cdot\|}(1).\end{aligned}$$

Therefore, the sum of equations (22) and (23) becomes

$$\mathbb{E}[QAQ] = \bar{Q}A\bar{Q} + \frac{1}{(1+\frac{1}{n}\mathrm{Tr}[\Sigma\bar{Q}])^2}\cdot\frac{1}{n}\mathrm{Tr}(\mathbb{E}[\Sigma QAQ])\bar{Q}\Sigma\bar{Q} + o_{\|\cdot\|}(1).$$

Taking normalized trace in both sides of this equation, we get

$$\frac{1}{n}\mathrm{Tr}(\mathbb{E}[\Sigma QAQ]) = \frac{1}{n}\mathrm{Tr}(\Sigma\bar{Q}A\bar{Q}) + \frac{\frac{1}{n}\mathrm{Tr}[(\Sigma\bar{Q})^2]}{(1+\frac{1}{n}\mathrm{Tr}[\Sigma\bar{Q}])^2}\cdot\frac{1}{n}\mathrm{Tr}(\mathbb{E}[\Sigma QAQ]) + o(1).$$

and obtain the solution

$$\frac{1}{n}\mathrm{Tr}(\mathbb{E}[\Sigma QAQ]) = \Big\{1 - \frac{\frac{1}{n}\mathrm{Tr}[(\Sigma\bar{Q})^2]}{(1+\frac{1}{n}\mathrm{Tr}[\Sigma\bar{Q}])^2}\Big\}^{-1}\cdot\frac{1}{n}\mathrm{Tr}(\Sigma\bar{Q}A\bar{Q}) + o(1).$$

Thus

$$\mathbb{E}[QAQ] = \bar{Q}A\bar{Q} + \frac{\frac{1}{n}\mathrm{Tr}[\Sigma\bar{Q}A\bar{Q}]}{(1+\frac{1}{n}\mathrm{Tr}[\Sigma\bar{Q}])^2 - \frac{1}{n}\mathrm{Tr}[(\Sigma\bar{Q})^2]}\cdot\bar{Q}\Sigma\bar{Q} + o_{\|\cdot\|}(1).$$

Appendix B. Main Proofs

B.1 Proof of Lemma 1

For bias term, note that

$$\mathbb{E}(\hat{\beta}_T|X) = \mathbb{E}(\hat{\beta}_T^{\mathrm{ridge}}|X) + A_T\mathbb{E}(\hat{\beta}_{T-1}^{\mathrm{ridge}}|X) + \cdots + A_TA_{T-1}\cdots A_2\mathbb{E}(\hat{\beta}_1^{\mathrm{ridge}}|X).$$

Since $\mathbb{E}(\hat{\beta}_t^{\mathrm{ridge}}|X) = (\hat{\Sigma}_t+\lambda_tI_p)^{-1}\hat{\Sigma}_t\beta = (I_p - A_t)\beta$, we have

$$\begin{aligned}\mathbb{E}(\hat{\beta}_T|X) =& (I_p - A_T)\beta + A_T(I_p - A_{T-1})\beta + \cdots + A_TA_{T-1}\cdots A_2(I_p - A_1)\beta\\ =& \beta - A_TA_{T-1}\cdots A_1\beta.\end{aligned}$$

by (4), $B_X(\hat{\beta};\beta,\Sigma_0) = \beta^\top A_1A_2\cdots A_T\Sigma_0A_T\cdots A_2A_1\beta$.

For variance term, note that ridge estimators using different datasets are independent, we have

$$\mathrm{Cov}(\hat{\beta}_T|X) = \sum_{t=1}^{T} A_T\cdots A_{t+1}\mathrm{Cov}(\hat{\beta}_1^{\mathrm{ridge}}|X)A_{t+1}\cdots A_T.$$

Since

$$\begin{aligned}\mathrm{Cov}(\hat{\beta}_t^{\mathrm{ridge}}|X) &= \mathrm{Cov}[(\hat{\Sigma}_t+\lambda_tI_p)^{-1}(\frac{1}{n_t}X_t^\top\epsilon_t)|X]\\ &= \frac{\sigma^2}{n_t}(\hat{\Sigma}_t+\lambda_tI_p)^{-1}\hat{\Sigma}_t(\hat{\Sigma}_t+\lambda_tI_p)^{-1}\\ &= \frac{\sigma^2}{n_t}[(\hat{\Sigma}_t+\lambda_tI_p)^{-1} - \lambda_t(\hat{\Sigma}_t+\lambda_tI_p)^{-2}]\\ &= \frac{\sigma^2}{\lambda_tn_t}(A_t - A_t^2),\end{aligned}$$

we have

$$\mathrm{Cov}(\hat{\beta}_T|X) = \sigma^2 \sum_{t=1}^{T} \frac{1}{\lambda_t n_t} A_T \cdots A_{t+1}(A_t - A_t^2)A_{t+1} \cdots A_T.$$

Thus by (5),

$$V_X(\hat{\beta}; \beta, \Sigma_0) = \sigma^2 \sum_{t=1}^{T} \frac{1}{\lambda_t n_t} \mathrm{Tr}\big[A_T \cdots A_{t+1}(A_t - A_t^2)A_{t+1} \cdots A_T \Sigma_0\big].$$

B.2 Proof of Theorem 4

B.2.1 Bias Term

Let $Q_t = A_t/\lambda_t$, then Q_t is a resolvent matrix of sample covariance matrix $\hat{\Sigma}_t = \frac{1}{n_t} X_t^\top X_t$. Note that

$$B_X(\hat{\beta}; \beta) = \big(\prod_{t=1}^{T} \lambda_t^2\big)\beta^\top Q_1 Q_2 \cdots Q_T Q_T \cdots Q_2 Q_1 \beta,$$

Let $B_t = Q_t \cdots Q_T Q_T \cdots Q_t$. For $t < T$, since $B_t = Q_t B_{t+1} Q_t$ and B_{t+1} is independent of Q_t, by Corollary 12,

$$B_t \asymp m_{\gamma_t}^2(-\lambda_t) B_{t+1} + \frac{1}{n_t} \mathrm{Tr} B_{t+1} \cdot m_{\gamma_t}^2(-\lambda_t)\mu_t I_p, \tag{29}$$

where

$$\mu_t = m'_{\gamma_t}(-\lambda_t)/[1 + \gamma_t m_{\gamma_t}(-\lambda_t)]^2,$$

by the third rule of Proposition 8,

$$\frac{1}{p}\mathrm{Tr} B_t - m_{\gamma_t}^2(-\lambda_t)(1 + \gamma_t \mu_t) \cdot \frac{1}{p}\mathrm{Tr} B_{t+1} \to 0, \quad a.s. \tag{30}$$

By differentiating (18), we obtain

$$m'_\gamma(z) = m_\gamma^2(z) \Big/ \Big[1 - \frac{\gamma m_\gamma^2(z)}{(1 + \gamma m_\gamma(z))^2}\Big],$$

therefore,

$$m_{\gamma_t}^2(-\lambda_t)(1 + \gamma_t \mu_t) = m'_{\gamma_t}(-\lambda_t). \tag{31}$$

Applying (30) iteratively and using (31), we obtain

$$\frac{1}{p}\mathrm{Tr} B_t - \prod_{s=t}^{T} m'_{\gamma_s}(-\lambda_s) \to 0, \quad a.s., \quad 1 \le t \le T, \tag{32}$$

substituting (32) in (29), we get

$$B_t \asymp m_{\gamma_t}^2(-\lambda_t)B_{t+1} + \gamma_t m_{\gamma_t}^2(-\lambda_t)\mu_t \cdot \prod_{s=t+1}^{T} m'_{\gamma_s}(-\lambda_s)I_p, \quad 1 \le t \le T-1.$$

To simplify our formula, we denote $w_t = \gamma_t\mu_t \cdot \prod_{s=t+1}^{T} m'_{\gamma_s}(-\lambda_s)$, then by iteration,

$$\begin{aligned}
B_1 \asymp & m_{\gamma_1}^2(-\lambda_1)(B_2 + w_1 I_p) \\
\asymp & m_{\gamma_1}^2(-\lambda_1)\big[m_{\gamma_2}^2(-\lambda_2)(B_3 + w_2 I_p) + w_1 I_p\big] \\
\asymp & \cdots \asymp \prod_{t=1}^{T-1} m_{\gamma_t}^2(-\lambda_t)B_T + \sum_{t=1}^{T-1}\big(\prod_{s=1}^{t} m_{\gamma_s}^2(-\lambda_s)\big)w_t \cdot I_p \\
\asymp & \Big\{m'_{\gamma_T}(-\lambda_T)\prod_{t=1}^{T-1} m_{\gamma_t}^2(-\lambda_t) + \sum_{t=1}^{T-1}\Big[\gamma_t u_t \prod_{s=1}^{t} m_{\gamma_s}^2(-\lambda_s)\prod_{s=t+1}^{T} m'_{\gamma_s}(-\lambda_s)\Big]\Big\} \cdot I_p \\
= & \Big\{m'_{\gamma_T}(-\lambda_T)(1+\gamma_{T-1}\mu_{T-1})\prod_{t=1}^{T-1} m_{\gamma_t}^2(-\lambda_t) + \sum_{t=1}^{T-2}\Big[\gamma_t u_t \prod_{s=1}^{t} m_{\gamma_s}^2(-\lambda_s)\prod_{s=t+1}^{T} m'_{\gamma_s}(-\lambda_s)\Big]\Big\} \cdot I_p \\
= & \Big\{\prod_{t=T-1}^{T} m'_{\gamma_t}(-\lambda_t)\prod_{t=1}^{T-2} m_{\gamma_t}^2(-\lambda_t) + \sum_{t=1}^{T-2}\Big[\gamma_t u_t \prod_{s=1}^{t} m_{\gamma_s}^2(-\lambda_s)\prod_{s=t+1}^{T} m'_{\gamma_s}(-\lambda_s)\Big]\Big\} \cdot I_p \\
= & \cdots = \prod_{t=1}^{T} m'_{\gamma_t}(-\lambda_t) \cdot I_p
\end{aligned} \tag{33}$$

By the definition of deterministic equivalent, we have

$$\beta^\top B_1 \beta \to r^2 \prod_{t=1}^{T} m'_{\gamma_t}(-\lambda_t), \quad a.s.$$

thus

$$B_X(\hat{\beta}_T;\beta) \to r^2 \prod_{t=1}^{T}\big(\lambda_t^2 m'_{\gamma_t}(-\lambda_t)\big), \quad a.s.$$

B.2.2 Variance Term

For $1 \le s < t \le T$, Let $C_{s,t} = Q_t \cdots Q_{s+1}Q_sQ_{s+1}\cdots Q_t$, $D_{s,t} = Q_t \cdots Q_{s+1}Q_s^2Q_{s+1}\cdots Q_t$, then

$$V_X(\hat{\beta}_T;\beta) = \sigma^2 \sum_{t=1}^{T}\big(\prod_{s=t+1}^{T}\lambda_s^2\big) \cdot \frac{1}{n_t}\mathrm{Tr}\big[C_{t,T} - \lambda_t D_{t,T}\big].$$

For $s < t$, since $C_{s,t} = Q_tC_{s,t-1}Q_t$, and $C_{s,t-1}$ is independent of Q_t with bounded operator norm, by Corollary 12,

$$C_{s,t} \asymp m_{\gamma_t}^2(-\lambda_t)C_{s,t-1} + \frac{1}{n_t}\mathrm{Tr}C_{s,t-1} \cdot m_{\gamma_t}^2(-\lambda_t)\mu_t I_p,$$

By rules of calculus for deterministic equivalents, we have

$$\frac{1}{p}\mathrm{Tr}C_{s,t} - m^2_{\gamma_t}(-\lambda_t)(1+\gamma_t\mu_t)\cdot\frac{1}{p}\mathrm{Tr}C_{s,t-1} \to 0, \quad a.s. \tag{34}$$

Applying (34) iteratively and notice (31), we obtain

$$\frac{1}{p}\mathrm{Tr}C_{t,T} - \prod_{s=t+1}^{T} m'_{\gamma_s}(-\lambda_s)\cdot\frac{1}{p}\mathrm{Tr}Q_t \to 0, \quad a.s.$$

By (17), we have

$$\frac{1}{p}\mathrm{Tr}Q_t \to m_{\gamma_t}(-\lambda_t), \quad a.s.$$

thus

$$\frac{1}{n_t}\mathrm{Tr}C_{t,T} \to \prod_{s=t+1}^{T} m'_{\gamma_s}(-\lambda_s)\cdot\gamma_t m_{\gamma_t}(-\lambda_t), \quad a.s. \tag{35}$$

Using similar approach to $D_{s,t} = Q_t D_{s,t-1} Q_t$, we obtain

$$\frac{1}{p}\mathrm{Tr}D_{t,T} - \prod_{s=t}^{T} m'_{\gamma_s}(-\lambda_s) \to 0, \quad a.s.$$

thus

$$\frac{1}{n_t}\mathrm{Tr}D_{t,T} \to \prod_{s=t}^{T} m'_{\gamma_s}(-\lambda_s)\cdot\gamma_t, \quad a.s. \tag{36}$$

Substituting (35) and (36) in expressions of V_X, we get

$$V_X(\hat{\beta}_T;\beta) \to \sigma^2\sum_{t=1}^{T}\gamma_t\upsilon_t\prod_{s=t+1}^{T}\left[\lambda_s^2 m'_{\gamma_s}(-\lambda_s)\right], \quad a.s.$$

B.3 Proof of Theorem 3

B.3.1 Bias Term

Let $B_t = Q_t\cdots Q_T\Sigma_0 Q_T\cdots Q_t, 1\le t\le T$ and $B_{T+1} = \Sigma_0$. For $s\le t$, define $\bar{Q}_t = \lambda_t^{-1}(I_p+\tilde{m}_t\Sigma_t)^{-1}$, $H_{s,t} = \bar{Q}_s\cdots\bar{Q}_t\Sigma_t\bar{Q}_t\cdots\bar{Q}_s$, and

$$\begin{aligned}
\tilde{\mu}_t &= \left\{\left[1+\frac{1}{n_t}\mathrm{Tr}(\Sigma_t\bar{Q}_t)\right]^2 - \frac{1}{n_t}\mathrm{Tr}(\Sigma_t\bar{Q}_t)^2\right\}^{-1},\\
\tilde{\rho}_t &= \frac{1}{p}\mathrm{Tr}[\Sigma_t\bar{Q}_t B_{t+1}\bar{Q}_t] = \frac{1}{p}\mathrm{Tr}[H_{t,t}B_{t+1}],\\
\bar{B}_t &= \bar{Q}_t\cdots\bar{Q}_T\Sigma_0\bar{Q}_T\cdots\bar{Q}_t.
\end{aligned}$$

Notice that $B_t = Q_t B_{t+1} Q_t$ and B_{t+1} has bounded operator norm, by Theorem 11, we have

$$\begin{aligned}
B_t \asymp & \bar{Q}_t B_{t+1} \bar{Q}_t + \gamma_t \tilde{\rho}_t \tilde{\mu}_t H_{t,t} \\
\asymp & \bar{Q}_t (\bar{Q}_{t+1} B_{t+2} \bar{Q}_{t+1} + \gamma_{t+1} \tilde{\rho}_{t+1} \tilde{\mu}_{t+1} H_{t+1,t+1}) \bar{Q}_t + \gamma_t \tilde{\rho}_t \tilde{\mu}_t H_{t,t} \\
= & \bar{Q}_t \bar{Q}_{t+1} B_{t+2} \bar{Q}_{t+1} \bar{Q}_t + \gamma_{t+1} \tilde{\rho}_{t+1} \tilde{\mu}_{t+1} H_{t,t+1} + \gamma_t \tilde{\rho}_t \tilde{\mu}_t H_{t,t} \\
\asymp & \cdots \asymp \bar{Q}_t \cdots \bar{Q}_T B_{T+1} \bar{Q}_T \cdots \bar{Q}_t + \sum_{s=t}^{T} \gamma_s \tilde{\rho}_s \tilde{\mu}_s H_{t,s} \\
= & \bar{B}_t + \sum_{s=t}^{T} \gamma_s \tilde{\rho}_s \tilde{\mu}_s H_{t,s}.
\end{aligned}$$

For $1 \le t \le T-1$, since $H_{t,t}$ has bounded operator norm, we have

$$H_{t,t}^{1/2} B_{t+1} H_{t,t}^{1/2} \asymp H_{t,t}^{1/2} \bar{B}_{t+1} H_{t,t}^{1/2} + \sum_{s=t+1}^{T} \gamma_s \tilde{\rho}_s \tilde{\mu}_s H_{t,t}^{1/2} H_{t+1,s} H_{t,t}^{1/2},$$

by the property of deterministic equivalence,

$$\tilde{\rho}_t = \frac{1}{p} \mathrm{Tr}[H_{t,t} B_{t+1}] \to \frac{1}{p} \mathrm{Tr}[H_{t,t} \bar{B}_{t+1}] + \sum_{s=t+1}^{T} \gamma_s \tilde{\rho}_s \tilde{\mu}_s \frac{1}{p} \mathrm{Tr}[H_{t,t} H_{t+1,s}], \quad a.s.$$

By Assumption 4 and Assumption 5, we have $\frac{1}{p}\mathrm{Tr}[H_{t,t}\bar{B}_{t+1}] \to a_t, \frac{1}{p}\mathrm{Tr}[H_{t,t}H_{t+1,s}] \to a_{t,s}, \tilde{\mu}_t \to \mu_t$, thus $\tilde{\rho}_t \to \rho_t, a.s.$, where ρ_t is defined recursively by $\rho_T = a_T$ and

$$\rho_t = a_t + \sum_{s=t+1}^{T} \gamma_s \mu_s a_{t,s} \rho_s, \quad 1 \le t \le T-1.$$

Therefore, we have

$$B_1 \asymp \bar{B}_1 + \sum_{t=1}^{T} \gamma_t \rho_t \mu_t H_{1,t},$$

similar to the proof of Theorem 4, the deterministic equivalence above yields

$$\beta^\top B_1 \beta - \beta^\top \Big[\bar{B}_1 + \sum_{t=1}^{T} \gamma_t \rho_t \mu_t H_{1,t} \Big] \beta \to 0, \quad a.s.$$

and by definition of G_n,we have almost surely

$$\begin{aligned}
\beta^\top \Big[\bar{B}_1 + \sum_{t=1}^{T} \gamma_t \rho_t \mu_t H_{1,t} \Big] \beta = & \|\beta\|^2 \Big\{ \int \frac{s_0}{\prod_{t=1}^{T} \lambda_t^2 (1 + \tilde{m}_t s_t)^2} dG_n(\mathbf{s}) + \\
& \sum_{t=1}^{T} \gamma_t \mu_t \rho_t \int \frac{s_t}{\prod_{j=1}^{t} \lambda_j^2 (1 + \tilde{m}_j s_j)^2} dG_n(\mathbf{s}) \Big\} \\
\to & r^2 \Big(g_0 + \sum_{t=1}^{T} \gamma_t \mu_t \rho_t g_{1,t} \Big),
\end{aligned}$$

therefore,

$$B_X(\hat{\beta}_T;\beta) \to \tilde{B}_T(r,\boldsymbol{\gamma},\boldsymbol{\lambda},G,H) := r^2\Big(\prod_{j=1}^{T}\lambda_j^2\Big)\Big(g_1+\sum_{t=1}^{T}\gamma_t\mu_t\rho_t g_{1,t}\Big), \quad a.s.$$

B.3.2 Variance Term

For $1 \le s < t \le T$, Let $C_{s,t} = Q_t\cdots Q_{s+1}Q_sQ_{s+1}\cdots Q_t, D_{s,t} = Q_t\cdots Q_{s+1}Q_s^2Q_{s+1}\cdots Q_t$, then

$$V_X(\hat{\beta}_T;\beta) = \sigma^2\sum_{t=1}^{T}\big(\prod_{s=t+1}^{T}\lambda_s^2\big)\cdot\frac{1}{n_t}\mathrm{Tr}\big[(C_{t,T}-\lambda_t D_{t,T})\Sigma_0\big]. \tag{37}$$

For $s < t$, let $G_{s,t} = \bar{Q}_t\cdots\bar{Q}_s\Sigma_s\bar{Q}_s\cdots\bar{Q}_t$, $\bar{C}_{s,t} = \bar{Q}_t\cdots\bar{Q}_{s+1}\bar{Q}_s\bar{Q}_{s+1}\cdots\bar{Q}_t$, $\tilde{\rho}_{s,t}^{(1)} = p^{-1}\mathrm{Tr}(G_{t,t}C_{s,t-1})$. Notice that $C_{s,t} = Q_tC_{s,t-1}Q_t$ and $C_{s,t-1}$ has bounded operator norm, by Theorem 11, we have

$$\begin{aligned}
C_{s,t} \asymp& \bar{Q}_tC_{s,t-1}\bar{Q}_t + \gamma_t\tilde{\rho}_{s,t}^{(1)}\tilde{\mu}_tG_{t,t}\\
\asymp& \bar{Q}_t(\bar{Q}_{t-1}C_{s,t-2}\bar{Q}_{t-1} + \gamma_{t-1}\tilde{\rho}_{s,t-1}^{(1)}\tilde{\mu}_{t-1}G_{t-1,t-1})\bar{Q}_t + \gamma_t\tilde{\rho}_{s,t}^{(1)}\tilde{\mu}_tG_{t,t}\\
\asymp& \bar{Q}_t\bar{Q}_{t-1}C_{s,t-2}\bar{Q}_{t-1}\bar{Q}_t + \gamma_{t-1}\tilde{\rho}_{s,t-1}^{(1)}\tilde{\mu}_{t-1}G_{t-1,t} + \gamma_t\tilde{\rho}_{s,t}^{(1)}\tilde{\mu}_tG_{t,t}\\
\asymp& \cdots \asymp \bar{Q}_t\cdots\bar{Q}_{s+1}Q_s\bar{Q}_{s+1}\cdots\bar{Q}_t + \sum_{j=s+1}^{t}\gamma_j\tilde{\rho}_{s,j}^{(1)}\tilde{\mu}_jG_{j,t}\\
\asymp& \bar{C}_{s,t} + \sum_{j=s+1}^{t}\gamma_j\tilde{\rho}_{s,j}^{(1)}\tilde{\mu}_jG_{j,t}.
\end{aligned}$$

Since $G_{t,t}$ has bounded operator norm, we have

$$G_{t,t}^{1/2}C_{s,t-1}G_{t,t}^{1/2} \asymp G_{t,t}^{1/2}\bar{C}_{s,t-1}G_{t,t}^{1/2} + \sum_{j=s+1}^{t}\gamma_j\tilde{\rho}_{s,j}^{(1)}\tilde{\mu}_jG_{t,t}^{1/2}G_{j,t-1}G_{t,t}^{1/2}.$$

If we define $\tilde{\rho}_{s,s}^{(1)} = 0$, we have

$$\tilde{\rho}_{s,t}^{(1)} = \frac{1}{p}\mathrm{Tr}[G_{t,t}C_{s,t-1}] \to \frac{1}{p}\mathrm{Tr}[G_{t,t}\bar{C}_{s,t-1}] + \sum_{j=s}^{t-1}\gamma_j\tilde{\rho}_{s,j}^{(1)}\tilde{\mu}_j\frac{1}{p}\mathrm{Tr}[G_{t,t}G_{j,t-1}], \quad a.s.,$$

By Assumption 4 and Assumption 5, we have $\frac{1}{p}\mathrm{Tr}[G_{t,t}\bar{C}_{s,t-1}] \to b_{s,t}$, $\frac{1}{p}\mathrm{Tr}[G_{t,t}G_{j,t-1}] \to a_{j,t}$, thus $\tilde{\rho}_{s,t}^{(1)} \to \rho_{s,t}^{(1)}, a.s.$, where $\rho_{s,t}^{(1)}$ is defined recursively by $\rho_{s,s}^{(1)} = 0$ and

$$\rho_{s,t}^{(1)} = b_{s,t} + \sum_{j=s}^{t-1}\gamma_j\mu_ja_{j,t}\rho_{s,j}^{(1)}, \quad 1 \le s < t \le T.$$

Since $\frac{1}{p}\mathrm{Tr}[\bar{C}_{t,T}\Sigma_0] \to b_t$, $\frac{1}{p}\mathrm{Tr}[G_{j,T}\Sigma_0] \to a_j$, and

$$\frac{1}{p}\mathrm{Tr}[C_{t,T}\Sigma_0] \to \frac{1}{p}\mathrm{Tr}[\bar{C}_{t,T}\Sigma_0] + \sum_{j=t}^{T}\gamma_j\mu_j\rho_{t,j}^{(1)}\frac{1}{p}\mathrm{Tr}[G_{j,T}\Sigma_0], \quad a.s.$$

we have

$$\frac{1}{p}\mathrm{Tr}[C_{t,T}\Sigma_0] \to b_t + \sum_{j=t}^{T}\gamma_j\mu_j a_j\rho_{t,j}^{(1)} = L_{1,t}, \quad a.s. \tag{38}$$

Similarly, let $\bar{D}_{s,t} = \bar{Q}_t\cdots\bar{Q}_{s+1}\bar{Q}_s^2\bar{Q}_{s+1}\cdots\bar{Q}_t$, $\tilde{\rho}_{s,t}^{(2)} = p^{-1}\mathrm{Tr}(G_{t,t}D_{s,t-1})$, we have

$$\begin{aligned} D_{s,t} \asymp & \bar{Q}_t\cdots\bar{Q}_{s+1}Q_s^2\bar{Q}_{s+1}\cdots\bar{Q}_t + \sum_{j=s+1}^{t}\gamma_j\tilde{\rho}_{s,j}^{(2)}\tilde{\mu}_j G_{j,t} \\ \asymp & \bar{D}_{s,t} + \frac{1}{n_s}\mathrm{Tr}[G_{s,s}]\tilde{\mu}_s G_{s,t} + \sum_{j=s+1}^{t}\gamma_j\tilde{\rho}_{s,j}^{(2)}\tilde{\mu}_j G_{j,t}. \end{aligned}$$

If we define $\tilde{\rho}_{s,s}^{(2)} = \frac{1}{p}\mathrm{Tr}[G_{s,s}]$, we have

$$\tilde{\rho}_{s,t}^{(2)} = \frac{1}{p}\mathrm{Tr}[G_{t,t}D_{s,t-1}] \to \frac{1}{p}\mathrm{Tr}[G_{t,t}\bar{D}_{s,t-1}] + \sum_{j=s}^{t-1}\gamma_j\tilde{\rho}_{s,j}^{(2)}\tilde{\mu}_j\frac{1}{p}\mathrm{Tr}[G_{t,t}G_{j,t-1}], \quad a.s.,$$

By Assumption 4 and Assumption 5, we have $\frac{1}{p}\mathrm{Tr}[G_{t,t}\bar{D}_{s,t-1}] \to c_{s,t}$, $\frac{1}{p}\mathrm{Tr}[G_{t,t}G_{j,t-1}] \to a_{j,t}$, and $\frac{1}{p}\mathrm{Tr}[G_{s,s}] = c_{s,s}$, thus $\tilde{\rho}_{s,t}^{(2)} \to \rho_{s,t}^{(2)}, a.s.$, where $\rho_{s,t}^{(2)}$ is defined recursively by $\rho_{s,s}^{(2)} = c_{s,s}$ and

$$\rho_{s,t}^{(2)} = c_{s,t} + \sum_{j=s}^{t-1}\gamma_j\mu_j a_{j,t}\rho_{s,j}^{(2)}, \quad 1 \le s < t \le T.$$

Since $\frac{1}{p}\mathrm{Tr}[\bar{D}_{t,T}\Sigma_0] \to c_t$, we obtain

$$\frac{1}{p}\mathrm{Tr}[D_{t,T}\Sigma_0] \to c_t + \sum_{j=t}^{T}\gamma_j\mu_j a_j\rho_{t,j}^{(2)} = L_{2,t}, \quad a.s. \tag{39}$$

Substituting (38) and (39) in (37), we get

$$V_X(\hat{\beta}_T;\beta) \to \tilde{V}_T(\boldsymbol{\gamma},\boldsymbol{\lambda},H) := \sigma^2\sum_{t=1}^{T}\gamma_t\Big(\prod_{s=t+1}^{T}\lambda_s^2\Big)(L_{1,t} - \lambda_t L_{2,t}).$$

Appendix C. Auxiliary Lemmas

Lemma 13 (Lemma B.26, Bai and Silverstein (2010)) *Let $A \in \mathbb{R}^{n\times n}$ be nonrandom matrix, $x = (x_1, \cdots, x_n)^\top \in \mathbb{R}^n$ be random vector with independent entries. Assume that $\mathbb{E}x_i = 0, \mathbb{E}|x_i|^2 = 1$, and $\mathbb{E}|x_i|^l \le v_l$. Then, for any $k \ge 1$,*

$$\mathbb{E}|x^\top Ax - \mathrm{Tr}A|^k \le C_k\Big[\big(v_4\mathrm{Tr}(AA^\top)\big)^{k/2} + v_{2k}\mathrm{Tr}(AA^\top)^{k/2}\Big],$$

for some $C_k > 0$.

Lemma 14 (Lemma 2.13, Bai and Silverstein (2010)) *Let x_i be a martingale difference sequence with respect to the increasing σ-field $\{\mathcal{F}_i\}$. Then for $k \ge 2$,*

$$\mathbb{E}\Big|\sum_{i=1}^n x_i\Big|^k \le C_k\Big(\mathbb{E}\big(\sum_{i=1}^n \mathbb{E}_{\le i-1}|x_i|^2\big)^{k/2} + \sum_{i=1}^n \mathbb{E}|x_i|^k\Big).$$

for some $C_k > 0$, where $\mathbb{E}_{\le i}$ denotes the expectation taken over $\mathcal{F}_i$.

Lemma 15 *Let A be a deterministic matrix with bounded operator norm, Q be the resolvent matrix defined in Theorem 11. Then for $k \ge 2$,*

$$\mathbb{E}\Big|\frac{1}{p}\mathrm{Tr}[A(Q - \mathbb{E}Q)]\Big|^k = O(n^{-k/2}).$$

for some $C_k > 0$.

Proof Let $Q_{-i} = (\frac{1}{n}\sum_{j\ne i} x_j x_j^\top - zI_p)^{-1}$ be the resolvent matrix without sample x_i. Note that

$$\begin{aligned}\frac{1}{p}\mathrm{Tr}[A(Q - \mathbb{E}Q)] =& \sum_{i=1}^n \frac{1}{p}\mathrm{Tr}[A\mathbb{E}_{\le i}(Q)] - \sum_{i=1}^n \frac{1}{p}\mathrm{Tr}[A\mathbb{E}_{\le i-1}(Q)]\\ =& \frac{1}{p}\sum_{i=1}^n (\mathbb{E}_{\le i} - \mathbb{E}_{\le i-1})\mathrm{Tr}[A(Q - Q_{-i})].\end{aligned}$$

where $\mathbb{E}_{\le i}$ is the expectation taken over $\mathcal{F}_i$ generated by $z_1 \cdots, z_i$, and $\mathbb{E}_{\le 0}Q = Q$. By Sherman-Morrison formula, we have

$$Q = Q_{-i} - \frac{\frac{1}{n}Q_{-i}x_i x_i^\top Q_{-i}}{1 + \frac{1}{n}x_i^\top Q_{-i}x_i}, \tag{40}$$

thus

$$\frac{1}{p}\mathrm{Tr}[A(Q - Q_{-i})] = \frac{1}{pn}\cdot\frac{x_i^\top Q_{-i}AQ_{-i}x_i}{1 + \frac{1}{n}x_i^\top Q_{-i}x_i}.$$

Let $Y_i = (\mathbb{E}_{\le i} - \mathbb{E}_{\le i-1})\frac{1}{p}\mathrm{Tr}[A(Q - Q_{-i})]$, then $\{Y_i\}$ is a martingale difference sequence. Since

$$\Big|\frac{1}{pn}\cdot\frac{x_i^\top Q_{-i}AQ_{-i}x_i}{1 + \frac{1}{n}x_i^\top Q_{-i}x_i}\Big| \le \Big|\frac{1}{p}\cdot\frac{x_i^\top Q_{-i}AQ_{-i}x_i}{x_i^\top Q_{-i}x_i}\Big| \le \frac{1}{p}\cdot\|Q_{-i}^{1/2}AQ_{-i}^{1/2}\|_{op} \le p^{-1}\lambda^{-1}\|A\|_{op},$$

we have $|Y_i| \le 2p^{-1}\lambda^{-1}\|A\|_{op} = O(p^{-1})$. By Lemma 14,

$$\mathbb{E}\Big|\frac{1}{p}\mathrm{Tr}[A(Q-\mathbb{E}Q)]\Big|^k \le C_k\Big(\mathbb{E}\big(\sum_{i=1}^n \mathbb{E}_{\le i-1}|Y_i|^2\big)^{k/2} + \sum_{i=1}^n \mathbb{E}|Y_i|^k\Big) = O(n^{-k/2}).$$

■

References

Z. Bai and J. Silverstein. *Spectral Analysis of Large Dimensional Random Matrices, Second Edition.* Springer, New York, 2010.

P. Buzzega, M. Boschini, A. Porrello, and S. Calderara. Rethinking experience replay: a bag of tricks for continual learning. In *International Conference on Pattern Recognition*, pages 2180–2187, 2021.

R. Couillet and Z. Liao. *Random Matrix Methods for Machine Learning.* Cambridge University Press, Cambridge, 2022.

R. Couillet, M. Debbah, and J. W. Silverstein. A deterministic equivalent for the analysis of correlated MIMO multiple access channels. *IEEE Transactions on Information Theory*, 57(6):3493–3514, 2011.

M. De Lange, R. Aljundi, M. Masana, S. Parisot, X. Jia, A. Leonardis, G. Slabaugh, and T. Tuytelaars. A continual learning survey: Defying forgetting in classification tasks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(7):3366–3385, 2022.

L. H. Dicker and M. A. Erdogdu. Flexible results for quadratic forms with applications to variance components estimation. *The Annals of Statistics*, 45(1):386–414, 2017.

E. Dobriban and Y. Sheng. WONDER: Weighted one-shot distributed ridge regression in high dimensions. *Journal of Machine Learning Research*, 21(66):1–52, 2020.

E. Dobriban and Y. Sheng. Distributed linear regression by averaging. *The Annals of Statistics*, 49(2):918–943, 2021.

E. Dobriban and S. Wager. High-dimensional asymptotics of prediction: Ridge regression and classification. *The Annals of Statistics*, 46(1):247–279, 2018.

I. Evron, E. Moroshko, R. Ward, N. Srebro, and D. Soudry. How catastrophic can catastrophic forgetting be in linear regression? In *Conference on Learning Theory*, pages 4028–4079, 2022.

E. Fini, V. G. T. da Costa, X. Alameda-Pineda, E. Ricci, K. Alahari, and J. Mairal. Self-supervised models are continual learners. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9611–9620, 2022.

D. Goldfarb and P. Hand. Analysis of overparameterization in continual learning under a linear model. *arXiv preprint arXiv:2502.10442*, 2025.

M. B. Gurbuz and C. Dovrolis. NISPA: Neuro-inspired stability-plasticity adaptation for continual learning in sparse networks. In *International Conference on Machine Learning*, pages 8157–8174, 2022.

W. Hachem, P. Loubaton, and J. Najim. Deterministic equivalents for certain functionals of large random matrices. *The Annals of Applied Probability*, 17(3):875–930, 2007.

T. Hastie, A. Montanari, S. Rosset, and R. J. Tibshirani. Surprises in high-dimensional ridgeless least squares interpolation. *The Annals of Statistics*, 50(2):949–986, 2022.

S. C. Hoi, D. Sahoo, J. Lu, and P. Zhao. Online learning: A comprehensive survey. *Neurocomputing*, 459:249–289, 2021.

X. Jin, A. Sadhu, J. Du, and X. Ren. Gradient-based editing of memory examples for online task-free continual learning. In *Advances in Neural Information Processing Systems*, pages 29193–29205, 2021.

S. Jung, H. Ahn, S. Cha, and T. Moon. Continual learning with node-importance based adaptive group sparse regularization. In *Advances in Neural Information Processing Systems*, pages 3647–3658, 2020.

J. Kirkpatrick, R. Pascanu, N. Rabinowitz, J. Veness, G. Desjardins, A. A. Rusu, K. Milan, J. Quan, T. Ramalho, A. Grabska-Barwinska, D. Hassabis, C. Clopath, D. Kumaran, and R. Hadsell. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13):3521–3526, 2017.

S. Lee, S. Goldt, and A. Saxe. Continual learning in the teacher-student setup: Impact of task similarity. In *International Conference on Machine Learning*, pages 6109–6119, 2021.

H. Li, J. Wu, and V. Braverman. Fixed design analysis of regularization-based continual learning. In *Proceedings of The 2nd Conference on Lifelong Learning Agents*, pages 513–533, 2023.

Y. Li, M. Li, M. S. Asif, and S. Oymak. Provable and efficient continual representation learning. *arXiv preprint arXiv:2203.02026*, 2022.

Z. Li and D. Hoiem. Learning without forgetting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(12):2935–2947, 2018.

D. Lopez-Paz and M. Ranzato. Gradient episodic memory for continual learning. In *Advances in Neural Information Processing Systems*, page 6470–6479, 2017.

V. A. Marčenko and L. A. Pastur. Distribution of eigenvalues for some sets of random matrices. *Mathematics of The Ussr-sbornik*, 1(4):457–483, 1967.

J. L. McClelland, B. L. McNaughton, and R. CO'Reilly. Why there are complementary learning systems in the hippocampus and neocortex: insights from the successes and failures of connectionist models of learning and memory. *Psychological Rewiew*, 102(3): 419–457, 1995.

M. McCloskey and N. J. Cohen. Catastrophic interference in connectionist networks: The sequential learning problem. In *Psychology of learning and motivation*, volume 24, pages 109–165. 1989.

R. Ramesh and P. Chaudhari. Model zoo: A growing brain that learns continually. In *International Conference on Learning Representations*, 2022.

S.-A. Rebuffi, A. Kolesnikov, G. Sperl, and C. H. Lampert. iCaRL: Incremental classifier and representation learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5533–5542, 2017.

M. Riemer, I. Cases, R. Ajemian, M. Liu, I. Rish, Y. Tu, and G. Tesauro. Learning to learn without forgetting by maximizing transfer and minimizing interference. In *International Conference on Learning Representations*, 2019.

H. Ritter, A. Botev, and D. Barber. Online structured laplace approximations for overcoming catastrophic forgetting. In *Advances in Neural Information Processing Systems*, pages 3742–3752, 2018.

J. Silverstein and Z. Bai. On the empirical distribution of eigenvalues of a class of large dimensional random matrices. *Journal of Multivariate Analysis*, 54(2):175–192, 1995.

L. Wang, X. Zhang, H. Su, and J. Zhu. A comprehensive survey of continual learning: Theory, method and application. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(8):5362–5383, 2024.

Y. Wen, Z. Tan, K. Zheng, C. Xie, and W. Huang. Provable contrastive continual learning. In *International Conference on Machine Learning*, pages 52819–52838, 2024.

C. Yang, M. Tiomoko, and Z. Wang. Optimizing spca-based continual learning: A theoretical approach. In *International Conference on Learning Representations*, 2023.

J. Yao, S. Zheng, and Z. Bai. *Large Sample Covariance Matrices and High-Dimensional Data Analysis*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, Cambridge, 2015.

X. Zhao, H. Wang, W. Huang, and W. Lin. A statistical theory of regularization-based continual learning. In *International Conference on Machine Learning*, pages 61021–61039, 2024.