WHEN DO DATA MIXTURES IMPROVE SCALING LAWS? INSIGHTS FROM HIGH-DIMENSIONAL REGRESSION

Anonymous authors
Paper under double-blind review

ABSTRACT

Modern machine learning systems are trained on mixtures of data from different domains, and choosing the right mixture can substantially improve downstream performance. Despite an extensive literature on data mixing and reweighting, existing work is largely empirical and it remains unclear when auxiliary data genuinely improves scaling laws rather than merely providing more samples. To gain insight into this question, we study a high-dimensional mixed-data regression model with a shared regression function, heterogeneous covariances and noise levels, and dataset sizes that may grow at different rates. we establish the minimax risk under an ellipsoidal parameter constraint for the general covariances structure and derive deterministic equivalents for the test error of ridge regression under commutative covariances. We then specialize to a target domain and an auxiliary domain with aligned power-law covariance spectra, where the theory yields explicit scaling laws in terms of spectral decay, target regularity, and the relative growth of the two datasets. These laws identify regimes in which combining data mixtures provably yields a faster scaling rate than using either dataset alone. In particular, improving the scaling law requires a specific interplay between spectra and relative sample sizes of the domains. Our numerical experiments on language models exhibit the same qualitative phenomenon: appropriate data mixtures yield a faster decrease in target-domain test loss than training on either domain alone.

1 INTRODUCTION

Many successful machine learning systems learn from data spanning multiple domains (Grattafiori et al., 2024; Zitkovich et al., 2023), making the composition of their training sets a central design choice. Several approaches to optimize data mixtures have been proposed, including transferring mixture weights from smaller to larger models (Xie et al., 2023; Fan et al., 2024; Liu et al., 2025), predicting how the optimal mixture changes with the total training budget (Kang et al., 2025), and adjusting the weights during a training run according to the model's progress on different domains (Chen et al., 2023; 2025; Jiang et al., 2025). The dependence of the performance on model size, data volume, and mixture composition has been described by empirical scaling laws (Aghajanyan et al., 2023; Shukor et al., 2025; Sedova et al., 2026), with recent work suggesting that the mixture composition can affect the scaling exponent (Hamidieh et al., 2026). At the same time, Wang et al. (2026) have observed that the choice of the mixture may be sensitive to small changes in training hyperparameters. Together, these findings motivate a principled understanding of how and when data mixtures improve performance and, in particular, the scaling laws.

When data is drawn from a single distribution, a line of theoretical work has derived scaling laws for linear and kernel regression (Maloney et al., 2022; Bahri et al., 2024; Lin et al., 2024; 2025; Atanasov et al., 2026), as well as for simplified models of neural networks (Paquette et al., 2024; Bordelon et al., 2024; Ferbach et al., 2025; Ren et al., 2025; Defilippis et al., 2024). However, less is known about how these laws change when training on multiple data sources. In the model studied by Hashimoto (2021), mixture composition affects only the prefactor and leaves the scaling exponent unchanged. Later, Jain et al. (2024) derived scaling laws for mixtures of real and surrogate data, and Wu et al. (2026a) showed that, in weak-to-strong generalization, a student trained on teacher-generated labels can achieve a better scaling exponent than its teacher. Furthermore, in a stylized memorization model, Medvedev et al. (2026b) showed that optimizing training proportions can improve the scaling exponent relative to sampling from the test distribution. Most recently, Dai

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

& Zheng (2026) developed a theoretical model of how mixture proportions affect learning curves, predicting effective mixtures across model and dataset sizes. These results leave open the characterization of when combining datasets yields faster scaling rates than either dataset alone, and how these gains depend on the properties of the data sources and their relative sizes.

In this paper, we study a tractable setting of high-dimensional regression and identify regimes in which data mixtures improve scaling laws. We consider multiple training domains with a common regression function, different feature covariances and noise levels, and sample sizes growing at different rates. Our contributions can be summarized as follows:

- **Minimax risk and ridge performance.** We characterize both minimax statistical limits and the performance of a ridge estimator: for general covariances, Theorem 1 gives a variational characterization of the minimax prediction risk under an ellipsoidal parameter constraint, with matching bounds up to constants; for covariances that commute, Theorem 2 derives deterministic equivalents for the bias and variance of mixed ridge regression, which depend only on population-level quantities (covariances, sample sizes, noise levels, regularization).
- **Scaling laws for data mixtures.** We next specialize both minimax and ridge characterizations to two mixtures with power-law spectra, under source and capacity conditions (Theorems 3-4). This allows to identify regimes where *(i)* ridge regression achieves the minimax optimal exponent in the scaling law, and *(ii)* combining different mixtures strictly improves the scaling exponent as compared to using either dataset alone.
- **Numerical evidence from language models.** Finally, we demonstrate via numerical simulations that decoder-only transformers exhibit a qualitatively similar advantage in data mixing: in regimes consistent with our theory, combining target and auxiliary data produces a faster decrease in target-domain test loss than training on either dataset alone (Figure 3).

2 Related work

From an **empirical viewpoint**, existing work on data mixtures has studied how the composition of the training set affects performance and how to choose effective mixtures. DoReMi, DoGE, and Chameleon select domain weights using proxy models, gradient alignment, and feature statistics, respectively (Xie et al., 2023; Fan et al., 2024; Xie et al., 2025). Skill-It, Aioli, and ADO adapt data selection during training (Chen et al., 2023; 2025; Jiang et al., 2025), while RegMix and MixMin choose mixtures through performance prediction and surrogate optimization (Liu et al., 2025; Thudi et al., 2025). Empirical mixing laws model loss as a function of domain proportions and training scale (Ge et al., 2024; Ye et al., 2025; Shukor et al., 2025), and AutoScale extrapolates optimized mixtures to larger training budgets (Kang et al., 2025). Additional work on scaling laws examines competition and positive transfer between modalities (Aghajanyan et al., 2023), and data mixing with repetition when target data is scarce (Sedova et al., 2026). Our work provides a theoretical perspective on data mixtures: we identify conditions under which combining datasets improves the scaling exponent over either dataset alone, and then validate them on language model experiments.

From a **theoretical viewpoint**, work on data mixtures includes distribution-weighted combinations of predictors (Mansour et al., 2008; Hoffman et al., 2018), selecting weights that balance distribution mismatch and estimation error (Konstantinov & Lampert, 2019; Deng et al., 2023), and adaptive sampling across distributions (Haghtalab et al., 2022). In nonparametric regression, gains from combining mixture data have been characterized over Lipschitz classes (Schmidt-Hieber & Zamolodtchikov, 2024), and recent work on Hölder classes establishes minimax rates faster than those attainable from either dataset alone (Zhou & Yang, 2025; Zamolodtchikov, 2026). Whether faster rates are possible for high-dimensional regression remains, however, open. In particular, existing analyses of data mixtures in this setting do not establish that mixing improves the scaling exponent (Dai & Zheng, 2026; Jain et al., 2024), and for ridge regression with data labeled according to different objectives, Jagadeesan et al. (2025) have identified regimes in which scaling worsens at large dataset sizes. A related topic is covariate shift, studied through minimax analyses of kernel ridge regression (Ma et al., 2023), asymptotic characterizations of the risk in kernel and random-feature regression (Canatar et al., 2021; Tripuraneni et al., 2021), and minimum-norm interpolation (Mallinar et al., 2024). Under both covariate and model shift, Yang et al. (2025); Song et al. (2026) have established precise risk characterizations for pooled least squares and minimum-norm interpolation, in the proportional regime where number of samples and feature dimension grow at the

same rate. Our analysis accommodates decaying covariance spectra and datasets whose sizes may grow at different rates, and it yields scaling laws under source-capacity conditions (Caponnetto & De Vito, 2007; Rudi & Rosasco, 2017). Such conditions also underpin scaling laws for kernel and random feature regression (Cui et al., 2021; Defilippis et al., 2024; Atanasov et al., 2026). We use this framework to establish when data mixing improves the scaling law exponent, both in a minimax sense and for ridge regression.

3 Problem setup

We consider K independent training datasets. For each $i \in [K]$, we observe n_i independent samples $(\boldsymbol{x}_i^{(j)}, y_i^{(j)})_{j\in[n_i]}$ from the i-th dataset, where $\boldsymbol{x}_i^{(j)}$ are drawn from a distribution P_i on $\mathbb{R}^d$ with

$$\mathbb{E}_{P_i}[\boldsymbol{x}] = 0, \qquad \mathbb{E}_{P_i}[\boldsymbol{x}\boldsymbol{x}^\top] = \boldsymbol{C}_i, \qquad i = 1, \dots, K.$$

We set $N := \sum_{i=1}^K n_i$. The dimension d may be infinite, in which case $\boldsymbol{C}_i$ is a trace-class operator on ℓ_2. Denote $\boldsymbol{X}_i = [\boldsymbol{x}_i^{(1)}, \dots, \boldsymbol{x}_i^{(n_i)}]^\mathsf{T} \in \mathbb{R}^{n_i \times d}$ the design matrix. All domains share the same target parameter $\boldsymbol{\theta}_*$, while their noise levels may differ. The labels $\boldsymbol{y}_i = (y_i^{(j)})_{j\in[n_i]}$ are given by

$$\boldsymbol{y}_i = \boldsymbol{X}_i\boldsymbol{\theta}_* + \varepsilon_i, \qquad \varepsilon_i \sim \mathcal{N}(0, \sigma_{\varepsilon_i}^2 \boldsymbol{I}_{n_i}), \qquad i = 1, \dots, K.$$

The test distribution is a mixture $P = \sum_k \pi_k^* P_k$ of the same domains, with $\pi_k^* \geq 0$ and $\sum_k \pi_k^* = 1$; a single target domain corresponds to $\pi^* = \boldsymbol{e}_1$. For an estimator $\widehat{\boldsymbol{\theta}}$, we consider the excess test risk

$$\mathsf{R}_{\boldsymbol{\varepsilon}}(\widehat{\boldsymbol{\theta}}) := \mathbb{E}_{\boldsymbol{x}\sim P}\left[\left(\boldsymbol{x}^\mathsf{T}\widehat{\boldsymbol{\theta}} - \boldsymbol{x}^\mathsf{T}\boldsymbol{\theta}_*\right)^2\right] = \sum_{k=1}^K \pi_k^* \big\|\boldsymbol{C}_k^{1/2}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_*)\big\|_2^2.$$

In this paper, we characterize the minimax risk in this setting (i.e., the optimal risk over all estimators), and the risk of the following mixed ridge estimator

$$\widehat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} \left\{ \sum_{i=1}^K \|\boldsymbol{X}_i\boldsymbol{\theta} - \boldsymbol{y}_i\|_2^2 + \lambda\|\boldsymbol{\theta}\|_2^2 \right\} = \boldsymbol{G}\sum_{i=1}^K \boldsymbol{X}_i^\top \boldsymbol{y}_i, \tag{1}$$

where $\boldsymbol{G} = (\sum_{i=1}^K \boldsymbol{X}_i^\top \boldsymbol{X}_i + \lambda \boldsymbol{I})^{-1}$. Conditionally on designs, the risk of (1) averaged over label noise admits the bias-variance decomposition $\mathsf{R}(\widehat{\boldsymbol{\theta}}) := \mathbb{E}_{\boldsymbol{\varepsilon}}\mathsf{R}_{\boldsymbol{\varepsilon}}(\widehat{\boldsymbol{\theta}}) = \sum_k \pi_k^*(\mathsf{B}_k + \mathsf{V}_k)$ with

$$\mathsf{B}_k = \lambda^2\big\langle \boldsymbol{\theta}_*, \boldsymbol{G}\boldsymbol{C}_k\boldsymbol{G}\boldsymbol{\theta}_*\big\rangle, \qquad \mathsf{V}_k = \sum_{i=1}^K \sigma_{\varepsilon_i}^2 \operatorname{Tr}\big(\boldsymbol{C}_k\boldsymbol{G}\boldsymbol{X}_i^\top\boldsymbol{X}_i\boldsymbol{G}\big).$$

We assume the following concentration property on the designs.

Assumption 1 (Concentration of the designs)**.** *There exist constants $C, c, \eta > 0$ such that, for every $i \in [K]$, every PSD operator $\boldsymbol{A} \in \mathbb{R}^{d\times d}$, and every $t > 0$, a sample $\boldsymbol{x} \sim P_i$ satisfies*

$$\mathbb{P}(|\boldsymbol{x}^\top\boldsymbol{A}\boldsymbol{x} - \operatorname{Tr}(\boldsymbol{C}_i\boldsymbol{A})| > t\|\boldsymbol{C}_i^{1/2}\boldsymbol{A}\boldsymbol{C}_i^{1/2}\|_F) \leq Ce^{-ct^{1/\eta}}. \tag{2}$$

This condition covers several popular assumptions in the high-dimension regression literature, including independent sub-Gaussian coordinates and convex Lipschitz concentration (Cheng & Montanari, 2024; Misiakiewicz & Saeed, 2024).

Connection to kernel methods. Let $\psi = (\psi_j)_{j\geq 1}$ be the feature map of a reproducing kernel, and let the inputs of domain i be $\boldsymbol{z} \sim \rho_i$. Setting $\boldsymbol{x} = \psi(\boldsymbol{z})$ gives $\boldsymbol{C}_i = \mathbb{E}_{\rho_i}[\psi(\boldsymbol{z})\psi(\boldsymbol{z})^\top]$, and the shared-parameter model states that all domains are labeled by the same function $f_* = \langle\psi, \boldsymbol{\theta}_*\rangle$ in the RKHS. The estimator (1) is then kernel ridge regression under data mixture.

4 General results

Minimax error under ellipsoid constraint. Given a nonnegative self-adjoint operator $\boldsymbol{S}$ and a radius $R > 0$, consider $\boldsymbol{\Theta} = \{\boldsymbol{\theta} \in \mathrm{Dom}(\boldsymbol{S}) : \|\boldsymbol{S}^{1/2}\boldsymbol{\theta}\|_2 \leq R\}$, and define the minimax risk

$$\mathrm{R}_*(n_1, \dots, n_K) = \inf_{\widehat{\boldsymbol{\theta}}} \sup_{\boldsymbol{\theta}_*\in\boldsymbol{\Theta}} \sum_{k=1}^K \pi_k^* \,\mathbb{E}_{\boldsymbol{\theta}_*}\big[\|\boldsymbol{C}_k^{1/2}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_*)\|_2^2\big],$$

where the infimum is over all measurable estimators of the K datasets and the expectation is over designs and noise. We assume $\ker \boldsymbol{S} \subseteq \bigcap_i \ker \boldsymbol{C}_i$ and take inverse powers of $\boldsymbol{S}$ on $(\ker \boldsymbol{S})^\perp$. The relevant quantities are the test and training covariances rescaled by the ellipsoid constraint,

$$\boldsymbol{H}_i = \boldsymbol{S}^{-1/2}\boldsymbol{C}_i\boldsymbol{S}^{-1/2}, \qquad \boldsymbol{Q} = \sum_{k=1}^{K} \pi_k^* \boldsymbol{H}_k, \qquad \boldsymbol{M} = \sum_{i=1}^{K} \frac{n_i}{\sigma_{\varepsilon_i}^2}\boldsymbol{H}_i,$$

together with the largest signal-to-noise ratio permitted by the class, $E_0 = R^2 \max_i \|\boldsymbol{H}_i\|_{\mathrm{op}}/\sigma_{\varepsilon_i}^2$, and the variational functional

$$\mathcal{L}(\boldsymbol{M}) = \sup_{\substack{\boldsymbol{A}\succeq 0 \\ \mathrm{Tr}(\boldsymbol{A})\le R^2}} \mathrm{Tr}\Big[\boldsymbol{Q}\boldsymbol{A}^{1/2}\big(\boldsymbol{I} + \boldsymbol{A}^{1/2}\boldsymbol{M}\boldsymbol{A}^{1/2}\big)^{-1}\boldsymbol{A}^{1/2}\Big], \tag{3}$$

where in infinite dimension the supremum is over positive trace-class operators.

Theorem 1 (Minimax risk under an ellipsoid constraint)**.** *Under Assumption 1, there is a constant $C > 0$ depending only on the constants in that assumption such that, for every $n_1, \dots, n_K$,*

$$\frac{1}{\pi^2}\,\mathcal{L}(\boldsymbol{M}) \le \mathrm{R}_*(n_1, \dots, n_K) \le (1 + CE_0)\,\mathcal{L}(\boldsymbol{M}).$$

In particular, if E_0 is bounded by a constant, then $\mathrm{R}_*(n_1, \dots, n_K) \asymp \mathcal{L}(\boldsymbol{M})$ with constants independent of the dimension and of the sample sizes.

The proof compares the regression model with the Gaussian sequence model $\boldsymbol{y}_{\mathrm{seq}} = \boldsymbol{M}^{1/2}\boldsymbol{u} + \boldsymbol{g}$, $\boldsymbol{g} \sim \mathcal{N}(0, \boldsymbol{I})$, for the normalized parameter $\boldsymbol{u} = \boldsymbol{S}^{1/2}\boldsymbol{\theta}$. For the latter, we show that the minimax risk among *linear* estimators is exactly $\mathcal{L}(\boldsymbol{M})$, and the unrestricted minimax rate is within a factor of π^{-2} by a van Trees-type inequality (Gassiat & Stoltz, 2024). For the original regression model, the van Trees argument gives the same lower bound. We obtain the upper bound by constructing an explicit estimator and controlling the additional variance from the random designs, which introduces the factor $1 + CE_0$. The detailed proof is given in Appendix A.

Theorem 1 shows that, at the level of minimax rates, a collection of heterogeneous datasets is equivalent to a single Gaussian sequence model with information operator $\boldsymbol{M} = \sum_i n_i \sigma_{\varepsilon_i}^{-2}\boldsymbol{H}_i$. Domain i contributes information in proportion to $n_i/\sigma_{\varepsilon_i}^2$, with greater contributions along directions in which its covariance is large relative to the constraint. When $\boldsymbol{S}$ and the $\boldsymbol{C}_i$ commute, the variational problem reduces to a scalar program solved by truncation: in their common eigenbasis, ordering the coordinates by decreasing $[\boldsymbol{Q}]_{jj}$, one obtains (see Appendix C for a similar calculation in a power-law model)

$$\mathrm{R}_*(n_1, \dots, n_K) \asymp \inf_{m \ge 0} \Big\{ R^2 \max_{j>m} [\boldsymbol{Q}]_{jj} + \sum_{j \le m} \frac{[\boldsymbol{Q}]_{jj}}{[\boldsymbol{M}]_{jj}} \Big\}. \tag{4}$$

Each estimated coordinate costs its test weight divided by the total information that all domains carry about it, and the remaining coordinates cost the squared radius times their largest test weight.

Remark 4.1 (Tightness of the minimax bounds)**.** *The factor E_0 is $O(1)$ whenever $R = O(1)$, $\sigma_{\varepsilon_i}^2$ are bounded below, and the constraint dominates the covariances, $\max_i \|\boldsymbol{H}_i\|_{\mathrm{op}} = O(1)$. This covers Euclidean balls with bounded covariance spectra as well as the source-condition ellipsoids of Section 5. The lower bound uses Gaussian noise but only second moments of the designs. On ellipsoids, linear estimators are minimax within the Ibragimov–Hasminskii constant 1.25 (Donoho et al., 1990); our factor π^2 is not optimized, but the argument handles the weighted loss induced by $\boldsymbol{Q}$ and non-Gaussian designs.*

Deterministic equivalent for the mixed ridge estimator. Let $(\mu_1, \dots, \mu_K)$ be the unique positive solution of the fixed-point system

$$\mu_i = \frac{n_i}{1 + \mathrm{Tr}(\boldsymbol{C}_i\overline{\boldsymbol{G}})}, \qquad \overline{\boldsymbol{G}} = (\sum_{i=1}^{K} \mu_i \boldsymbol{C}_i + \lambda)^{-1}, \quad i \in [K].$$

The matrix $\overline{\boldsymbol{G}}$ is a deterministic proxy for the resolvent $\boldsymbol{G}$: each dataset is replaced by its population covariance scaled by an effective sample size $\mu_i \le n_i$, deflated by the effective dimension $\mathrm{Tr}(\boldsymbol{C}_i\overline{\boldsymbol{G}})$ that the dataset has to fit. For a deterministic $\boldsymbol{A}$, define $\tau_{\boldsymbol{A}} \in \mathbb{R}^K$ and the $K \times K$ matrix $\boldsymbol{L}$ by

$$\tau_{\boldsymbol{A}}[i] = \mathrm{Tr}(\boldsymbol{A}\overline{\boldsymbol{G}}\boldsymbol{C}_i\overline{\boldsymbol{G}}), \qquad \boldsymbol{L}_{ij} = \frac{n_i}{\mu_i^2}\mathbf{1}[i=j] - \mathrm{Tr}(\boldsymbol{C}_i\overline{\boldsymbol{G}}\boldsymbol{C}_j\overline{\boldsymbol{G}}).$$

Assumption 2. *There exists a constant $K_0 > 0$ such that for all $i \in [K]$, $\lambda\nu_\lambda^i(n_i) \geq n_i^{-K_0}$, where $\nu_\lambda^i(n_i)$ is defined as in* (34).

This condition imposes a polynomial lower bound on the sum of the ridge regularization and a covariance-dependent spectral term, and it is common in related work (Misiakiewicz & Saeed, 2024; Defilippis et al., 2024; Wu et al., 2026a). We note that the assumption is satisfied under the power-law spectra considered in Section 5.

Theorem 2 (Deterministic equivalent for mixed ridge)**.** *Under Assumption 1, assume $\boldsymbol{C}_k$ commute with each other. Let K be fixed, $\lambda > 0$ satisfy Assumption 2, and π^* be the deterministic test mixture weights. Define $\boldsymbol{C}_{\pi^*} = \sum_{k=1}^K \pi_k^* \boldsymbol{C}_k$, $\boldsymbol{A}_* = \boldsymbol{\theta}_* \boldsymbol{\theta}_*^\top$. Then $\boldsymbol{L}$ is invertible and, for any $D > 0$, with probability at least $1 - \sum_{i=1}^K n_i^{-D}$ over the training designs,*

$$|\mathsf{R}(\widehat{\theta}) - \overline{\mathsf{R}}| \leq C_{D,c_0} \left(\lambda^2 \varepsilon_2 V_{\boldsymbol{A}_*} + V_{\boldsymbol{C}_{\pi^*}} \sum_{i=1}^K \sigma_{\varepsilon_i}^2 \varepsilon_{3,i} \right), \tag{5}$$

$$\overline{\mathsf{R}} = \sum_{k=1}^K \pi_k^* \left(\overline{\mathsf{B}}_k + \overline{\mathsf{V}}_k \right), \ \overline{\mathsf{B}}_k = \lambda^2 \left[\langle \boldsymbol{\theta}_*, \overline{\boldsymbol{G}} \boldsymbol{C}_k \overline{\boldsymbol{G}} \boldsymbol{\theta}_* \rangle + \tau_{\boldsymbol{\theta}_* \boldsymbol{\theta}_*^\top}^\top \boldsymbol{L}^{-1} \tau_{\boldsymbol{C}_k} \right], \ \overline{\mathsf{V}}_k = \sum_{i=1}^K \sigma_{\varepsilon_i}^2 \, \tau_{\boldsymbol{C}_k}^\top \boldsymbol{L}^{-1} \boldsymbol{e}_i,$$

with $\boldsymbol{e}_i$ the i-th unit vector in $\mathbb{R}^K$, ν defined as in (34)*, $e_K = \left(\sum_{i=1}^K n_i^{-1} \right)^{1/2}$ and*

$$V_{\boldsymbol{A}} = \sum_{i=1}^K n_i \tau_{\boldsymbol{A}}[i], \ \varepsilon_2 = e_K^3 (\nu^{14} + \nu^8 \log^{4\eta+3/2}(N)), \ \varepsilon_{3,i} = e_K (e_K^2 \nu^{14} + \nu^8 + \nu^5 \log^{3\eta+3/2}(N)).$$

The proof (Appendix B.2) relies on computing the deterministic equivalents of functionals including $\mathrm{Tr}(\boldsymbol{A}\boldsymbol{G})$, $\mathrm{Tr}(\boldsymbol{A}\boldsymbol{G}\boldsymbol{C}_k\boldsymbol{G})$, and $\mathrm{Tr}(\boldsymbol{A}\boldsymbol{G}\boldsymbol{X}_k^\top \boldsymbol{X}_k \boldsymbol{G})$, where $\boldsymbol{A}$ is a fixed PSD matrix and $\boldsymbol{G}$ is the resolvent. The argument follow a similar decomposition and leave-one-out strategy as Misiakiewicz & Saeed (2024); Defilippis et al. (2024); Wu et al. (2026a). Here, we extend their analyses to heterogeneous data mixtures, by controlling a coupled family of covariance-weighted resolvent traces.

The error bound in the RHS of (5) is $O(e_K)$, which is additive as in (Wu et al., 2026a), due to asymmetric terms. Under power-law spectra and for optimal ridge regularization, we further show in Lemma 1 that the error bound is $o(\overline{R})$. More broadly, we conjecture that a multiplicative guarantee holds for a range of parameters, leaving this technical problem as a future direction. We also note that Theorem 2 requires all the covariances to commute, and conjecture this to be a technical requirement as well: we expect the deterministic equivalent in Theorem 2 to still hold for non-commutative covariances, possibly at the cost of a worse additive error. We refer to Figure 2 (discussed in Section 6) for an empirical demonstration of the validity of the deterministic equivalent predictions, well beyond our technical assumptions.

5 Scaling laws for data mixtures

We now specialize to $K = 2$ domains and to a test distribution equal to the first domain, $\pi^* = \boldsymbol{e}_1$. Domain 1 is the *target* and domain 2 the *auxiliary* domain. Assuming $\boldsymbol{C}_1$, $\boldsymbol{C}_2$, and $\boldsymbol{S}$ commute, and under a standard power-law decay condition on their eigenvalues, we characterize both the minimax rates and the rates achieved by ridge regression.

Power-law model. Without loss of generality, we consider $\boldsymbol{C}_1$ and $\boldsymbol{C}_2$ diagonal, take $d = \infty$ and aligned power-law spectra:

$$[\boldsymbol{C}_1]_{kk} = k^{-\alpha_1}, \qquad [\boldsymbol{C}_2]_{kk} = k^{-\alpha_2}, \qquad n_1 = n, \qquad n_2 = \lfloor n^{\gamma_2} \rfloor, \qquad \sigma_{\varepsilon_1}^2, \sigma_{\varepsilon_2}^2 = \Theta(1), \tag{6}$$

with $\alpha_1, \alpha_2 > 1$ and $\gamma_2 > 0$, and set $\delta = \alpha_1 - \alpha_2$. A positive δ means that the auxiliary spectrum has heavier tails: relative to the target, the auxiliary domain puts more mass on high-frequency directions. For the minimax rates, we consider the parameter class

$$\boldsymbol{\Theta} = \left\{ \boldsymbol{\theta} : \|\boldsymbol{C}_1^{1/2-r_1} \boldsymbol{\theta}\|_2 \leq R, \ \|\boldsymbol{C}_2^{1/2-r_2} \boldsymbol{\theta}\|_2 \leq R \right\}, \qquad s = \max \left\{ \alpha_1 r_1, \ \alpha_2 r_2 + \frac{\delta}{2} \right\},$$

which imposes a source condition with exponent $r_i > 0$ in each domain (Caponnetto & De Vito, 2007). Up to a change of radius, $\boldsymbol{\Theta}$ is the single ellipsoid $\{\boldsymbol{\theta} : \sum_k k^{2s-\alpha_1}\boldsymbol{\theta}[k]^2 \le R^2\}$ (Appendix C), so s is the regularity of the target in the target geometry, and $2s > \delta$. For mixed ridge regression, we consider the fixed signal $\boldsymbol{\theta}_*[k] = k^{-\beta}$, $\beta > 1/2$, which belongs to the ellipsoid above for every $s' < s = (\alpha_1 + 2\beta - 1)/2$ but not for $s' = s$ (this boundary membership is the source of the logarithmic factors in Theorem 4 below). In both cases, $[\boldsymbol{C}_1^{1/2}\boldsymbol{\theta}_*]_k \asymp k^{-(1+2s)/2}$: the target-only problem is the standard source-capacity setting with capacity exponent α_1.

Minimax rates. Let $\mathsf{R}^*(n_1, n_2) = \inf_{\widehat{\boldsymbol{\theta}}} \sup_{\boldsymbol{\theta}_* \in \boldsymbol{\Theta}} \mathbb{E}_{\boldsymbol{\theta}_*} \|\boldsymbol{C}_1^{1/2}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_*)\|_2^2$ be the minimax target risk, with $n_i = 0$ meaning that dataset i is unavailable. Define

$$\Gamma_{\mathrm{tar}} = \frac{2s}{1+2s}, \qquad \Gamma_{\mathrm{aux}} = \begin{cases} \dfrac{2s\gamma_2}{1+2s-\delta}, & \delta < 1, \\ \gamma_2, & \delta \ge 1, \end{cases} \qquad \gamma_{\mathrm{c}} = 1 - \frac{\delta}{1+2s}.$$

Theorem 3 (Minimax scaling law)**.** *Under Assumption 1 and the model* (6)*, the risks of the individual datasets satisfy* $\mathsf{R}^*(n, 0) \asymp n^{-\Gamma_{\mathrm{tar}}}$, $\mathsf{R}^*(0, n_2) \asymp n^{-\Gamma_{\mathrm{aux}}}$ *if* $\delta \neq 1$*, and* $\mathsf{R}^*(0, n_2) \asymp n^{-\Gamma_{\mathrm{aux}}} \log n$ *if* $\delta = 1$*. Furthermore, the risk for the data mixture satisfies*

$$\mathsf{R}^*(n, n_2) \asymp \begin{cases} \mathsf{R}^*(n, 0), & \gamma_2 \le \gamma_{\mathrm{c}}, \\ \mathsf{R}^*(0, n_2), & \gamma_2 > \gamma_{\mathrm{c}},\ \delta \le 1, \ \textit{or}\ \ \gamma_2 \ge 1,\ \delta > 1, \\ n^{-(\delta-1+\gamma_2)/\delta}, & \gamma_{\mathrm{c}} < \gamma_2 < 1,\ \delta > 1. \end{cases}$$

The proof (Appendix C) applies Theorem 1 with $\boldsymbol{S} = \boldsymbol{C}_1^{1-2s/\alpha_1}$, for which $E_0 = O(1)$, and evaluates the truncation formula (4), which here reads

$$\mathsf{R}^*(n, n_2) \asymp \inf_{m \ge 0} \Big\{ m^{-2s} + \sum_{j \le m} \frac{1}{I_j} \Big\}, \qquad I_j \asymp n + n_2 j^{\delta}. \tag{7}$$

The linear estimator in Appendix A.2 attains this rate but requires knowledge of the population covariances. We will next identify when ridge regression matches this rate without such knowledge.

Let us further comment on Theorem 3. In (7), I_j^{-1} can be regarded as the error of fitting coordinate j of the signal. When $\delta > 0$, I_j is dominated by $n_2 j^{\delta} \gg n$ above the crossover index $j_{\mathrm{c}} := (n/n_2)^{1/\delta} = n^{(1-\gamma_2)/\delta}$ where the error mainly comes from the auxiliary data. In contrast, below the crossover index j_{c}, $I_j \asymp n$. We discuss three cases:

(1) *Too few auxiliary samples ($\gamma_2 \le \gamma_{\mathrm{c}}$).* The target-only estimator learns $m_{\mathrm{tar}} = n^{1/(1+2s)}$ coordinates, and $\gamma_2 \le \gamma_{\mathrm{c}}$ is exactly $j_{\mathrm{c}} \ge m_{\mathrm{tar}}$. Thus, auxiliary data only helps with coordinates already well estimated in target data, and the mixed minimax rate does not improve on the target-only rate.

(2) *Light auxiliary tails ($\delta \le 1$).* Beyond j_{c}, the estimation cost $\sum_{j > j_{\mathrm{c}}} (n_2 j^{\delta})^{-1}$ is not summable and the variance is dominated by the highest coordinates being learnt, as if only auxiliary data was available. Thus, data mixing does not improve the minimax rate.

(3) *Heavy auxiliary tails ($\delta > 1$).* Here $\sum_{j > j_{\mathrm{c}}} (n_2 j^{\delta})^{-1} \asymp j_{\mathrm{c}}/n$ is summable: once auxiliary data takes over, all remaining coordinates are learned at a total cost comparable to that of the first j_{c} ones. Thus, the mixed minimax rate can improve upon both target-only and auxiliary-only rates.

In summary, the mixed minimax rate improves on both single-dataset minimax rates if and only if $\delta > 1$ and $\gamma_2 \in (\gamma_{\mathrm{c}}, 1)$. For $\delta > 0$, one can disregard the auxiliary data when $\gamma_2 \le \gamma_{\mathrm{c}}$, and disregard the target data when $\gamma_2 \ge 1$, without worsening the rate.

Remark 5.1 (Positive distribution shift)**.** *At equal sample sizes ($\gamma_2 = 1$) and $\delta > 0$, the auxiliary-only exponent $\Gamma_{\mathrm{aux}}(1) = 2s/(1 + 2s - \delta)$ for $\delta < 1$, or 1 for $\delta \ge 1$, is strictly larger than Γ_{tar}: n samples from the shifted distribution are more informative about the target than n target samples, because they put larger weight on high-frequency directions. This is an example of 'positive distribution shift', the observation that training over a different data distribution can help improve the test performance (Medvedev et al., 2026a;b). For $\delta < 0$, by contrast, $\gamma_{\mathrm{c}} > 1$: polynomially more auxiliary than target samples are needed before the rate changes.*

Scaling law of mixed ridge regression. We now turn to the estimator in (1) with the fixed signal $\boldsymbol{\theta}_*[k] = k^{-\beta}$, $s = (\alpha_1 + 2\beta - 1)/2$, and $\delta = \alpha_1 - \alpha_2$. Let $\overline{\mathsf{R}}_1(\lambda)$ denote the deterministic equivalent of the target risk given by Theorem 2, and $\overline{\mathsf{R}}_1^* = \inf_{\lambda>0} \overline{\mathsf{R}}_1(\lambda)$ the risk under optimal regularization. Define $(x)_+ := \max\{x, 0\}$, and consider the truncated coefficients

$$a = \min\{s, \alpha_1\}, \quad b = \min\{s, \alpha_2\}, \quad \kappa = \begin{cases} (a-b)/\delta, & \delta \neq 0, \\ 0, & \delta = 0, \end{cases} \quad \gamma_c^{\text{ridge}} = 1 - \frac{\delta}{1 + 2\max\{a, b\}}$$

and define

$$\Gamma_{\text{tar}}^{\text{ridge}} = \frac{2a}{1+2a}, \quad \Gamma_{\text{aux}}^{\text{ridge}} = \frac{2b\gamma_2}{2b + (1-\delta)_+}.$$

Theorem 4 (Ridge scaling law)**.** *Under Assumption 1 and the model* (6)*, the optimally regularized target risks of individual datasets satisfy* $\overline{\mathsf{R}}_1^*(n, 0) = \widetilde{\Theta}(n^{-\Gamma_{\text{tar}}^{\text{ridge}}})$ *and* $\overline{\mathsf{R}}_1^*(0, n_2) = \widetilde{\Theta}(n^{-\Gamma_{\text{aux}}^{\text{ridge}}})$*, with* $\widetilde{\Theta}(\cdot)$ *omitting logarithmic factors. Furthermore, the risk for the data mixture satisfies* $\overline{\mathsf{R}}_1^*(n, n_2) = \widetilde{\Theta}(n^{-\Gamma})$*, where*

$$\Gamma = \begin{cases} \Gamma_{\text{tar}}^{\text{ridge}} + \dfrac{2\kappa(\gamma_2 - 1)_+}{1+2a}, & \gamma_2 \leq \gamma_c^{\text{ridge}}, \\ \Gamma_{\text{aux}}^{\text{ridge}} + \dfrac{2(1-\delta)_+\kappa(1-\gamma_2)_+}{2b + (1-\delta)_+}, & \gamma_2 > \gamma_c^{\text{ridge}},\ \delta \leq 1,\ \textit{or} \\ & \gamma_2 \geq 1,\ \delta > 1, \\ \dfrac{\delta - 1 + \gamma_2}{\delta}, & \gamma_c^{\text{ridge}} < \gamma_2 < 1,\ \delta > 1. \end{cases} \tag{8}$$

The above rate is evaluated on the deterministic equivalent $\overline{\mathsf{R}}$ of Theorem 2. Lemma 1 below (proved in Appendix E) shows that, for the optimal ridge, the error bound in (5) is vanishing (i.e., $o(\overline{\mathsf{R}})$) and, thus, the same rate is achieved by the excess test risk $\mathsf{R}(\widehat{\theta})$ of the mixed ridge estimator.

Lemma 1 (Vanishing approximation error for optimal ridge.)**.** *In the setting of Theorem 4, there exists* λ_* *such that* $\overline{\mathsf{R}}_1(\lambda_*) = \widetilde{\Theta}(\overline{\mathsf{R}}_1^*(n, n_2))$*, and* $|\mathsf{R}_1(\widehat{\boldsymbol{\theta}}_{\lambda_*}) - \overline{\mathsf{R}}_1(\lambda_*)| = o(\overline{\mathsf{R}}_1^*(n, n_2))$*, with probability at least* $1 - \sum_{i=1}^2 n_i^{-D}$.

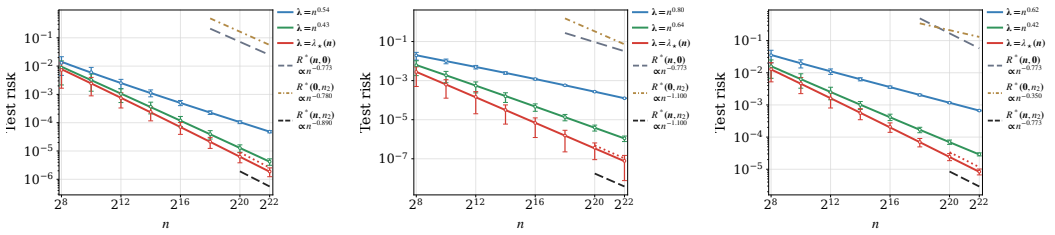
Comparing the rates for ridge of Theorem 4 with the minimax ones of Theorem 3 identifies the regimes in which the deterministic equivalent is minimax optimal.

Corollary 1 (Minimax optimality of ridge)**.** *In the setting of Theorem 4, we have that* $\overline{\mathsf{R}}_1^*(n, n_2) = \widetilde{\Theta}\big(\mathsf{R}^*(n, n_2)\big)$ *if and only if one of the following holds: (i)* $\gamma_2 \leq \gamma_{\text{c}}$ *and* $s \leq \alpha_1$*; (ii)* $\gamma_2 \geq \gamma_{\text{c}}$ *and either* $s \leq \alpha_2$ *or* $\delta \geq 1$*. In all other cases,* $\overline{\mathsf{R}}_1^*(n, n_2)$ *is polynomially larger than* $\mathsf{R}^*(n, n_2)$.

In particular, **whenever the mixed minimax rate is strictly better than the minimax rate of either dataset alone** (i.e., $\delta > 1$ and $\gamma_{\text{c}} < \gamma_2 < 1$), **ridge regression achieves the mixed minimax rate**. This conclusion holds for every regularity s, including $s > \alpha_1$, where target-only ridge is suboptimal because of saturation. Thus, a sufficiently heavier auxiliary spectral tail allows ridge to attain the improved minimax exponent using only $n_2 = o(n)$ auxiliary samples. We note that *(i)* $\gamma_2 > \gamma_{\text{c}}$ specifies how many auxiliary samples are needed to improve on the target-only risk, and *(ii)* $\gamma_2 < 1$ ensures that the target samples still improve on the auxiliary-only risk. This establishes a regime in terms of sample sizes where data mixtures are particularly effective.

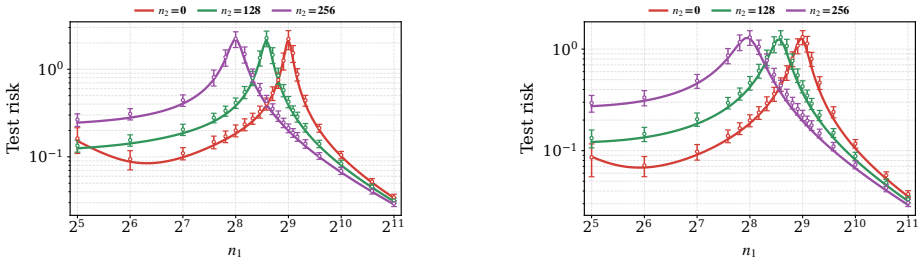
Remark 5.2 (Saturation induced by mixing)**.** *Ridge regression cannot attain the minimax rate when the source exponent* s *exceeds the capacity exponent of the spectrum it fits. This is the classical saturation of Tikhonov regularization (Bauer et al., 2007; Li et al., 2023), and it appears in the mixed regression setting of Theorem 4 through the truncations* $a = \min\{s, \alpha_1\}$ *and* $b = \min\{s, \alpha_2\}$.

For $\alpha_2 < s \leq \alpha_1$, $0 < \delta < 1$*, ridge is rate optimal on the target data alone, and when* $\gamma_2 > \gamma_{\text{c}}$*, auxiliary data still improves its exponent, but it no longer attains the mixed minimax rate. When* $\delta > 1$*, saturation is harmless as soon as* $\gamma_2 \geq \gamma_{\text{c}}$*. For* $s > \alpha_1$*, ridge is already suboptimal on the target data alone, and when* $\delta > 0$ *the threshold* $\gamma_{\text{c}}^{\text{ridge}}$ *lies below* γ_{c}*: auxiliary data partially compensates for saturation before the minimax rate itself changes. We expect iterated Tikhonov or early stopping to remove these truncations (Bauer et al., 2007).*



(a) $\gamma_2 = 0.78$ (b) $\gamma_2 = 1.1$ (c) $\gamma_2 = 0.35$

Figure 1: Test error (dots) of mixed ridge regression under power-law covariance, together with its corresponding deterministic equivalent from Theorem 2 (solid lines). The grey, golden and black dash lines are the minimax rates of target-only, auxiliary-only and mixture datasets from Theorem 3. The blue, green and red lines correspond to different regularizations with the optimal one in red. Here $\boldsymbol{x}_i \sim \mathcal{N}(0, \boldsymbol{C}_i)$, with $\boldsymbol{C}_i$ as in (6). We set $\alpha_1 = 3.2$, $\alpha_2 = 1.2$, $\beta = 0.6$, and $d = 512$.



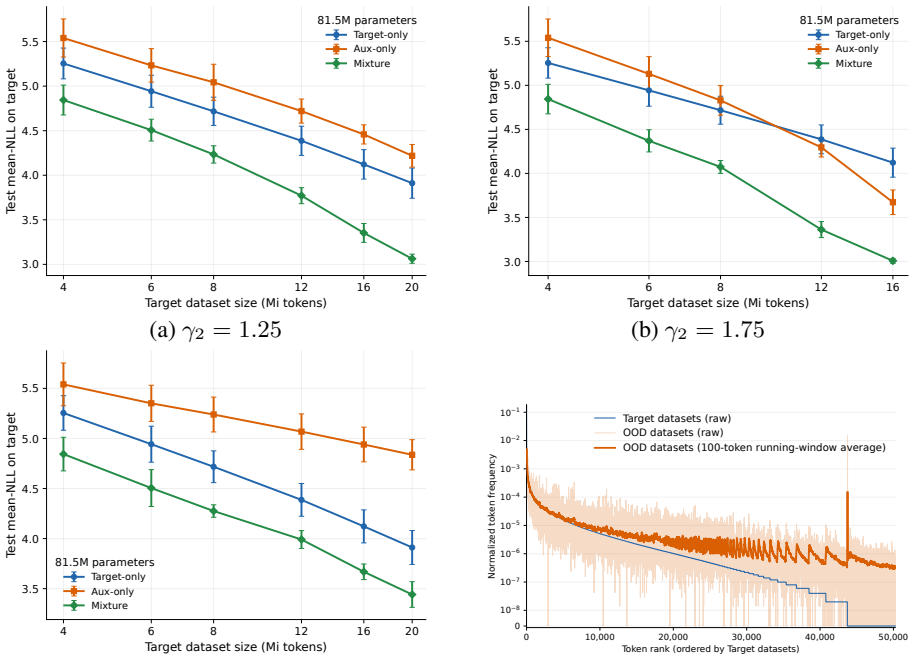
(a) ImageNet-100 (b) CIFAR-10

Figure 2: Test error (dots) of mixed ridge regression, together with its corresponding deterministic equivalent (solid lines). For CIFAR-10, features are extracted via a pretrained CLIP ViT-B/32 encoder, classes 0–4 form the target domain and classes 5–9 form the auxiliary domain. For ImageNet-100, features are extracted via a pretrained ResNet-18 encoder, and the corresponding class splits are 0–49 and 50–99. In both settings, we take $d = 512$ and generate labels according to $y = \boldsymbol{x}^\top \boldsymbol{\theta}_* + \xi$, with $\xi \sim \mathcal{N}(0, 0.1)$. The shared $\boldsymbol{\theta}_*$ is supported on the top-10 eigenvectors of $(\boldsymbol{C}_1 + \boldsymbol{C}_2)/2$, with random signs and coefficient magnitudes proportional to $j^{-1/2}$ for $j = 1, \ldots, 10$.

6 Numerical Experiments

Linear models. We consider a linear model following the setup of Section 3. Figure 1 indicates that the empirical risk of the mixed ridge estimator matches the deterministic equivalent of Theorem 2 and the scaling law predicted by Theorem 4. Furthermore, in panel (a), ridge regression achieves the mixed minimax rate and strictly improves upon the minimax rates of either dataset alone; in contrast, in panels (b) and (c), no improvement can be obtained and the rate of mixed ridge respectively coincides with the auxiliary-only and target-only rate. Next, we evaluate our deterministic equivalent on real image features from CIFAR-10 and ImageNet-100. The results in Figure 2 indicate that the validity of the deterministic equivalent of Theorem 2 goes well beyond the technical assumptions made therein.

Language modeling. Finally, we consider autoregressive language modeling. We pretrain an 81.5M-parameter GPT-2-style model (Radford et al., 2019) via nanoGPT (Karpathy, 2022) on the SlimPajama dataset (Cerebras Systems, 2023). We use n tokens from the `StackExchange` domain (target domain) and $n_2 = 4\left(\frac{n}{4\,\mathrm{Mi}}\right)^{\gamma_2}$ tokens from a fixed mixture of the remaining six domains (auxiliary domain). Full experimental details are in Appendix G.1. Figure 3 (panels (a)-(c)) shows that the improvement in the scaling law crucially relies on the relative growth rate of the token budget of the auxiliary domain: for $\gamma_2 = 1.25$, the scaling law of the data mixture improves upon that of either dataset alone; however, this improvement is not noticeable anymore already when $\gamma_2 = 0.75$ or $\gamma_2 = 1.75$. This qualitatively supports our theoretical finding that data mixtures improve scaling only within an appropriate relative-growth regime. Finally, the token-frequency distribution in panel (d) shows that the auxiliary domain places relatively more probability mass on tokens that are infrequent in the target domain. This provides a token-level analogue of the heavier-tailed auxiliary



(a) $\gamma_2 = 1.25$ (b) $\gamma_2 = 1.75$

(c) $\gamma_2 = 0.75$ (d) Token-frequency distributions

Figure 3: Target-domain test negative log-likelihood (NLL) scaling of an 81.5M-parameter GPT-2-style language model trained from scratch on SlimPajama. Panels (a)–(c) report mean NLL on a fixed target-domain test set as a function of the reference target budget n. We use $n \in \{4, 6, 8, 12, 16, 20\}$ Mi, except for the $\gamma_2 = 1.75$ experiment that uses $n \leq 16$ Mi. Panel (d) compares the normalized target and auxiliary token frequencies. Token types are ordered by decreasing target-domain frequency. The light auxiliary curve shows the raw frequencies, and the darker, thicker curve shows their 100-token running-window average.

covariance in our linear theory: the auxiliary data offer greater coverage of directions/tokens that are weakly represented in the target distribution. We also find that the scaling law of the data mixture does not improve when the token frequencies of the two datasets are too close with each other—an observation that agrees with our theory (see Figure 4 in Appendix G.2).

7 Conclusion

We study high-dimensional regression on data mixtures that share a regression function but differ in their covariances and noise levels. For general covariances, we characterize the minimax risk, and for commutative covariances, we derive deterministic equivalents for the test error of a mixed ridge estimator. Specializing these results to target and auxiliary domains with aligned power-law spectra, we derive scaling laws, thus establishing conditions under which mixing data is especially helpful. In particular, we show that, whenever the mixed minimax rate improves on the rate of either dataset alone, ridge regression is minimax optimal. This is the case when *(i)* the auxiliary spectrum is heavy tailed ($\delta > 1$), i.e., the two datasets cover complementary parts of the spectrum, and *(ii)* the auxiliary sample size has an appropriate growth rate ($\gamma_c < \gamma_2 < 1$). Our experiments on transformers trained on a mixture of text corpora display a phenomenology that qualitatively agrees with the theory.

Our work opens the door to several interesting future directions. A natural next step is to determine whether gradient descent methods attain the mixed minimax rates when ridge saturates, as recently shown for a single dataset by Wu et al. (2026b). Considering random-feature models (Defilippis et al., 2024; Bordelon et al., 2024), or features trained via a step of GD (Ba et al., 2022; Cui et al., 2024), would then identify how model size must scale with data budget to preserve the improvement in the scaling law. Finally, it would be valuable to jointly optimize domain proportions under data and compute constraints, as well as to characterize mixture transfer across model scales, motivated by recent theory on hyperparameter transfer (Ghosh et al., 2026).

AI USE STATEMENT

We used generative AI tools to polish the writing, implement straightforward parts of the code, fill in proof details after we had identified the key methods and arguments, check proofs for correctness, and improve the coverage of related work. We did not use generative AI tools to develop the paper's core ideas and define the problem setting. We take full responsibility for the final content of this work, including all text, claims, code, and other artifacts produced with the assistance of generative AI.

REPRODUCIBILITY STATEMENT

All proofs of theoretical results are provided in the appendices. All experiments were conducted on widely used, publicly available datasets, as discussed in Section 6 and Appendix G.

REFERENCES

Armen Aghajanyan, Lili Yu, Alexis Conneau, Wei-Ning Hsu, Karen Hambardzumyan, Susan Zhang, Stephen Roller, Naman Goyal, Omer Levy, and Luke Zettlemoyer. Scaling laws for generative mixed-modal language models. In *International Conference on Machine Learning*, volume 202, pp. 265–279. PMLR, 2023.

Alexander Atanasov, Jacob A. Zavatone-Veth, and Cengiz Pehlevan. Scaling and renormalization in high-dimensional regression. *Journal of Statistical Mechanics: Theory and Experiment*, 2026(4): 043404, 2026.

Jimmy Ba, Murat A Erdogdu, Taiji Suzuki, Zhichao Wang, Denny Wu, and Greg Yang. High-dimensional asymptotics of feature learning: How one gradient step improves the representation. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

Yasaman Bahri, Ethan Dyer, Jared Kaplan, Jaehoon Lee, and Utkarsh Sharma. Explaining neural scaling laws. *Proceedings of the National Academy of Sciences*, 121(27):e2311878121, 2024.

Frank Bauer, Sergei Pereverzev, and Lorenzo Rosasco. On regularization algorithms in learning theory. *Journal of complexity*, 23(1):52–72, 2007.

Blake Bordelon, Alexander Atanasov, and Cengiz Pehlevan. A dynamical model of neural scaling laws. In *International Conference on Machine Learning*, volume 235, pp. 4345–4382. PMLR, 2024.

Abdulkadir Canatar, Blake Bordelon, and Cengiz Pehlevan. Out-of-distribution generalization in kernel regression. In *Advances in Neural Information Processing Systems*, volume 34, pp. 12600–12612, 2021.

Andrea Caponnetto and Ernesto De Vito. Optimal rates for the regularized least-squares algorithm. *Foundations of Computational Mathematics*, 7(3):331–368, 2007.

Cerebras Systems. SlimPajama: A 627b token, cleaned and deduplicated version of RedPajama. Dataset release, June 2023. URL `https://www.cerebras.ai/blog/slimpajama-a-627b-token-cleaned-and-deduplicated-version-of-redpajama`.

Mayee F. Chen, Nicholas Roberts, Kush Bhatia, Jue Wang, Ce Zhang, Frederic Sala, and Christopher Ré. Skill-it! a data-driven skills framework for understanding and training language models. In *Advances in Neural Information Processing Systems*, volume 36, pp. 36000–36040, 2023.

Mayee F. Chen, Michael Y. Hu, Nicholas Lourie, Kyunghyun Cho, and Christopher Ré. Aioli: A unified optimization framework for language model data mixing. In *International Conference on Learning Representations*, 2025.

Chen Cheng and Andrea Montanari. Dimension free ridge regression. *The Annals of Statistics*, 52 (6):2879–2912, 2024.

Hugo Cui, Bruno Loureiro, Florent Krzakala, and Lenka Zdeborová. Generalization error rates in kernel regression: The crossover from the noiseless to noisy regime. In *Advances in Neural Information Processing Systems*, volume 34, pp. 10131–10143, 2021.

Hugo Cui, Luca Pesce, Yatin Dandi, Florent Krzakala, Yue Lu, Lenka Zdeborova, and Bruno Loureiro. Asymptotics of feature learning in two-layer networks after one gradient-step. In *International Conference on Machine Learning*, volume 235. PMLR, 2024.

Rui Dai and Shuran Zheng. Explaining data mixing scaling laws. In *International Conference on Machine Learning*, 2026.

Leonardo Defilippis, Bruno Loureiro, and Theodor Misiakiewicz. Dimension-free deterministic equivalents and scaling laws for random feature regression. In *Advances in Neural Information Processing Systems*, volume 37, pp. 104630–104693, 2024.

Yuyang Deng, Ilja Kuzborskij, and Mehrdad Mahdavi. Mixture weight estimation and model prediction in multi-source multi-target domain adaptation. In *Advances in Neural Information Processing Systems*, volume 36, pp. 4845–4898, 2023.

David L Donoho, Richard C Liu, and Brenda MacGibbon. Minimax risk over hyperrectangles, and implications. *The Annals of Statistics*, pp. 1416–1437, 1990.

Simin Fan, Matteo Pagliardini, and Martin Jaggi. Doge: Domain reweighting with generalization estimation. In *International Conference on Machine Learning*, pp. 12895–12915. PMLR, 2024.

Damien Ferbach, Katie Everett, Gauthier Gidel, Elliot Paquette, and Courtney Paquette. Dimension-adapted momentum outscales sgd. In *Advances in Neural Information Processing Systems*, volume 38, pp. 112780–112977, 2025.

Elisabeth Gassiat and Gilles Stoltz. The van Trees inequality in the spirit of Hájek and Le Cam. *Statistical Science*, 39(4):644–653, 2024.

Ce Ge, Zhijian Ma, Daoyuan Chen, Yaliang Li, and Bolin Ding. BiMix: A bivariate data mixing law for language model pretraining. *arXiv preprint arXiv:2405.14908*, 2024.

Nikhil Ghosh, Denny Wu, and Alberto Bietti. Understanding the mechanisms of fast hyperparameter transfer. In *International Conference on Learning Representations*, 2026.

Aaron Grattafiori, Abhimanyu Dubey, et al. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Nika Haghtalab, Michael Jordan, and Eric Zhao. On-demand sampling: Learning optimally from multiple distributions. In *Advances in Neural Information Processing Systems*, volume 35, pp. 406–419, 2022.

Kimia Hamidieh, Lester Mackey, and David Alvarez-Melis. Domain-aware scaling laws uncover data synergy. In *Conference on Language Modeling*, 2026.

Tatsunori Hashimoto. Model performance scaling with multiple data sources. In *International Conference on Machine Learning*, volume 139, pp. 4107–4116. PMLR, 2021.

Judy Hoffman, Mehryar Mohri, and Ningshan Zhang. Algorithms and theory for multiple-source adaptation. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Meena Jagadeesan, Michael I. Jordan, and Jacob Steinhardt. Safety versus performance: How multi-objective learning reduces barriers to market entry. *Proceedings of the National Academy of Sciences*, 122(42):e2510004122, 2025.

Ayush Jain, Andrea Montanari, and Eren Sasoglu. Scaling laws for learning with real and surrogate data. In *Advances in Neural Information Processing Systems*, volume 37, pp. 110246–110289, 2024.

Yiding Jiang, Allan Zhou, Zhili Feng, Sadhika Malladi, and J. Zico Kolter. Adaptive data optimization: Dynamic sample selection with scaling laws. In *International Conference on Learning Representations*, 2025.

Iain M Johnstone. Function estimation and gaussian sequence models. Unpublished manuscript, 2002.

Feiyang Kang, Yifan Sun, Bingbing Wen, Si Chen, Dawn Song, Rafid Mahmood, and Ruoxi Jia. Autoscale: Scale-aware data mixing for pre-training llms. In *Conference on Language Modeling*, 2025.

Andrej Karpathy. nanoGPT. GitHub repository, 2022. URL https://github.com/karpathy/nanoGPT.

Nikola Konstantinov and Christoph Lampert. Robust learning from untrusted sources. In *International conference on machine learning*, pp. 3488–3498. PMLR, 2019.

Michel Ledoux and Michel Talagrand. *Probability in Banach Spaces: Isoperimetry and Processes*. Springer, Berlin, Heidelberg, 1991. doi: 10.1007/978-3-642-20212-4.

Yicheng Li, Haobo Zhang, and Qian Lin. On the saturation effect of kernel ridge regression. In *The Eleventh International Conference on Learning Representations*, 2023.

Licong Lin, Jingfeng Wu, Sham M Kakade, Peter L Bartlett, and Jason D Lee. Scaling laws in linear regression: Compute, parameters, and data. In *Advances in Neural Information Processing Systems*, volume 37, pp. 60556–60606, 2024.

Licong Lin, Jingfeng Wu, and Peter Bartlett. Improved scaling laws in linear regression via data reuse. In *Advances in Neural Information Processing Systems*, volume 38, pp. 90713–90759, 2025.

Qian Liu, Xiaosen Zheng, Niklas Muennighoff, Guangtao Zeng, Longxu Dou, Tianyu Pang, Jing Jiang, and Min Lin. Regmix: Data mixture as regression for language model pre-training. In *International Conference on Learning Representations*, volume 2025, pp. 38305–38339, 2025.

Cong Ma, Reese Pathak, and Martin J. Wainwright. Optimally tackling covariate shift in RKHS-based nonparametric regression. *The Annals of Statistics*, 51(2):738–761, 2023.

Neil Rohit Mallinar, Austin Zane, Spencer Frei, and Bin Yu. Minimum-norm interpolation under covariate shift. In *International Conference on Machine Learning*, volume 235, pp. 34543–34585. PMLR, 2024.

Alexander Maloney, Daniel A Roberts, and James Sully. A solvable model of neural scaling laws. *arXiv:2210.16859*, 2022.

Yishay Mansour, Mehryar Mohri, and Afshin Rostamizadeh. Domain adaptation with multiple sources. In *Advances in Neural Information Processing Systems*, volume 21, 2008.

Marko Medvedev, Idan Attias, Elisabetta Cornacchia, Theodor Misiakiewicz, Gal Vardi, and Nathan Srebro. Positive distribution shift as a framework for understanding tractable learning. In *International Conference on Machine Learning*, 2026a.

Marko Medvedev, Kaifeng Lyu, Zhiyuan Li, and Nathan Srebro. Shift is good: Mismatched data mixing improves test performance. In *International Conference on Artificial Intelligence and Statistics*, volume 300, pp. 4690–4698. PMLR, 2026b.

Theodor Misiakiewicz and Basil Saeed. A non-asymptotic theory of kernel ridge regression: deterministic equivalents, test error, and gcv estimator. *arXiv preprint arXiv:2403.08938*, 2024.

Elliot Paquette, Courtney Paquette, Lechao Xiao, and Jeffrey Pennington. 4+ 3 phases of compute-optimal neural scaling laws. In *Advances in Neural Information Processing Systems*, volume 37, pp. 16459–16537, 2024.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. Technical report, OpenAI, 2019. URL https://cdn.openai.com/better-language-models/language-models.pdf.

Yunwei Ren, Eshaan Nichani, Denny Wu, and Jason Lee. Emergence and scaling laws in sgd learning of shallow neural networks. In *Advances in Neural Information Processing Systems*, volume 38, pp. 38227–38309, 2025.

Alessandro Rudi and Lorenzo Rosasco. Generalization properties of learning with random features. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Johannes Schmidt-Hieber and Petr Zamolodtchikov. Local convergence rates of the nonparametric least squares estimator with applications to transfer learning. *Bernoulli*, 30(3):1845–1877, 2024.

Anastasiia Sedova, Skyler Seto, Natalie Schluter, and Pierre Ablin. Scaling laws for mixture pretraining under data constraints. *arXiv preprint arXiv:2605.12715*, 2026.

Mustafa Shukor, Louis Bethune, Dan Busbridge, David Grangier, Enrico Fini, Alaaeldin El-Nouby, and Pierre Ablin. Scaling laws for optimal data mixtures. In *Advances in Neural Information Processing Systems*, volume 38, pp. 129554–129579, 2025.

Maurice Sion. On general minimax theorems. *Pacific Journal of Mathematics*, 8(1):171 – 176, 1958.

Yanke Song, Kenneth Gu, Sohom Bhattacharya, and Pragya Sur. Generalization error of min-norm interpolators in transfer learning. *arXiv preprint arXiv:2406.13944*, 2026.

Anvith Thudi, Evianne Rovers, Yangjun Ruan, Tristan Thrush, and Chris J. Maddison. MixMin: Finding data mixtures via convex minimization. In *International Conference on Machine Learning*, volume 267, pp. 59492–59508. PMLR, 2025.

Nilesh Tripuraneni, Ben Adlam, and Jeffrey Pennington. Overparameterization improves robustness to covariate shift in high dimensions. In *Advances in Neural Information Processing Systems*, volume 34, pp. 13883–13897, 2021.

Joel A. Tropp. An introduction to matrix concentration inequalities. *Foundations and Trends in Machine Learning*, 8(1–2):1–230, 2015.

Jiachen Tianhao Wang, Tong Wu, Kaifeng Lyu, James Y Zou, Dawn Song, Ruoxi Jia, and Prateek Mittal. Can small training runs reliably guide data curation? rethinking proxy-model practice. In *International Conference on Learning Representations*, volume 2026, pp. 122921–122964, 2026.

Diyuan Wu, Lehan Chen, Theodor Misiakiewicz, and Marco Mondelli. Improved scaling laws via weak-to-strong generalization in random feature ridge regression. In *International Conference on Machine Learning*, 2026a.

Jingfeng Wu, Peter L Bartlett, Sham M Kakade, Jason D Lee, and Bin Yu. Risk comparisons in linear regression: Implicit regularization dominates explicit regularization. In *Conference on Learning Theory*, 2026b.

Sang Michael Xie, Hieu Pham, Xuanyi Dong, Nan Du, Hanxiao Liu, Yifeng Lu, Percy S Liang, Quoc V Le, Tengyu Ma, and Adams Wei Yu. Doremi: Optimizing data mixtures speeds up language model pretraining. In *Advances in Neural Information Processing Systems*, volume 36, pp. 69798–69818, 2023.

Wanyun Xie, Francesco Tonin, and Volkan Cevher. Chameleon: A flexible data-mixing framework for language model pretraining and finetuning. In *International Conference on Machine Learning*, pp. 68725–68745. PMLR, 2025.

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

Jiasheng Ye, Peiju Liu, Tianxiang Sun, Jun Zhan, Yunhua Zhou, and Xipeng Qiu. Data mixing laws: Optimizing data mixtures by predicting language modeling performance. In *International Conference on Learning Representations*, 2025.

Petr Zamolodtchikov. A minimax theory of nonparametric regression under covariate shift. *arXiv preprint arXiv:2603.05897*, 2026.

Ling Zhou and Yuhong Yang. On a synergistic learning phenomenon in nonparametric domain adaptation. *arXiv preprint arXiv:2511.17009*, 2025.

Brianna Zitkovich, Tianhe Yu, et al. RT-2: Vision-language-action models transfer web knowledge to robotic control. In *Proceedings of the 7th Conference on Robot Learning*, volume 229, pp. 2165–2183, 2023.

A Proof of Theorem 1

First, define the normalized weight $\boldsymbol{u} = \boldsymbol{S}^{1/2}\boldsymbol{\theta}, \widehat{\boldsymbol{u}} = \boldsymbol{S}^{1/2}\widehat{\boldsymbol{\theta}}$, and given any PSD matrix $\boldsymbol{N}$, we construct an auxiliary Gaussian sequence model (Johnstone, 2002),

$$\boldsymbol{y}_{\boldsymbol{N},\mathrm{seq}} = \boldsymbol{N}^{1/2}\boldsymbol{u} + \boldsymbol{g}, \quad \boldsymbol{g} \sim \mathcal{N}(0, \boldsymbol{I}), \tag{9}$$

where $\boldsymbol{N} \in \mathbb{R}^{d\times d}$ and $\boldsymbol{u}, \boldsymbol{g}, \boldsymbol{y}_{\boldsymbol{N},\mathrm{seq}} \in \mathbb{R}^d$.

Define the unrestricted and linear minimax errors of the Gaussian sequence model as

$$\mathrm{R}_{*,\mathrm{seq}}(\boldsymbol{N}) = \inf_{\widehat{\boldsymbol{u}}} \sup_{\|\boldsymbol{u}\|_2\le R} \mathbb{E}_{\boldsymbol{u}}\|\boldsymbol{Q}^{1/2}(\widehat{\boldsymbol{u}}(\boldsymbol{y}_{\boldsymbol{N},\mathrm{seq}}) - \boldsymbol{u})\|_2^2,$$
$$\mathrm{R}^{\mathrm{lin}}_{*,\mathrm{seq}}(\boldsymbol{N}) = \inf_{\boldsymbol{B}} \sup_{\|\boldsymbol{u}\|_2\le R} \mathbb{E}_{\boldsymbol{u}}\|\boldsymbol{Q}^{1/2}(\boldsymbol{B}\boldsymbol{y}_{\boldsymbol{N},\mathrm{seq}} - \boldsymbol{u})\|_2^2,$$

where we recall $\boldsymbol{Q} = \sum_{k=1}^K \pi_k^* \boldsymbol{H}_k$.

We first have the following variational characterization of the minimax error of the Gaussian sequence model (9).

Theorem 5 (Exact Gaussian linear minimax error)**.** *For the Gaussian sequence model* (9) *defined by any PSD matrix* $\boldsymbol{N}$,

$$\pi^{-2}\mathcal{L}(\boldsymbol{N}) \le \mathrm{R}_{*,\mathrm{seq}}(\boldsymbol{N}) \le \mathrm{R}^{\mathrm{lin}}_{*,\mathrm{seq}}(\boldsymbol{N}) = \mathcal{L}(\boldsymbol{N}), \tag{10}$$

where $L(\boldsymbol{N}) = \sup_{\substack{\boldsymbol{A}\succeq 0\\ \mathrm{Tr}(\boldsymbol{A})\le R^2}} \mathrm{Tr}\Big[\boldsymbol{Q}\boldsymbol{A}^{1/2}(\boldsymbol{I} + \boldsymbol{A}^{1/2}\boldsymbol{N}\boldsymbol{A}^{1/2})^{-1}\boldsymbol{A}^{1/2}\Big]$.

We prove Theorem 5 in Appendix A.1, and next in Appendix A.2, we explain the way to extend the proof in Theorem 5 to the original regression problem in Section 3, which proves Theorem 1. We first prove these result for $d < \infty$. The extension to the infinite-dimensional case is treated in Appendix A.3.

A.1 Proof of Theorem 5

Proof of the upper bound. First of all, it is clear by definition that $\mathrm{R}_{*,\mathrm{seq}}(\boldsymbol{N}) \le \mathrm{R}^{\mathrm{lin}}_{*,\mathrm{seq}}(\boldsymbol{N})$, and we characterize $\mathrm{R}^{\mathrm{lin}}_{*,\mathrm{seq}}(\boldsymbol{N})$ in the rest of the proof.

For any $\boldsymbol{B}$, computing the test error gives

$$\begin{aligned}\mathrm{R}^{\mathrm{lin}}_{*,\mathrm{seq}}(\boldsymbol{N}) &= \inf_{\boldsymbol{B}} \sup_{\|\boldsymbol{u}\|_2\le R} \mathbb{E}_{\boldsymbol{u}}\|\boldsymbol{Q}^{1/2}(\boldsymbol{B}\boldsymbol{Y}_{\boldsymbol{N}} - \boldsymbol{u})\|_2^2\\ &= \inf_{\boldsymbol{B}} \sup_{\|\boldsymbol{u}\|_2} \boldsymbol{u}^\top(\boldsymbol{I} - \boldsymbol{N}^{1/2}\boldsymbol{B}^\top)\boldsymbol{Q}(\boldsymbol{I} - \boldsymbol{B}\boldsymbol{N}^{1/2})\boldsymbol{u} + \mathrm{Tr}(\boldsymbol{Q}\boldsymbol{B}\boldsymbol{B}^\top).\end{aligned}$$

Note that for every any PSD matrix $\boldsymbol{D}$,

$$\sup_{\|\boldsymbol{u}\|_2\le R} \boldsymbol{u}^\top\boldsymbol{D}\boldsymbol{u} = R^2\|\boldsymbol{D}\|_{\mathrm{op}} = \sup_{\substack{\boldsymbol{A}\succeq 0\\ \mathrm{Tr}(\boldsymbol{A})\le R^2}} \mathrm{Tr}(\boldsymbol{D}\boldsymbol{A}).$$

Consequently, we could rewrite the test error as

$$F(\boldsymbol{B}, \boldsymbol{A}) = \mathrm{Tr}\Big[\boldsymbol{Q}\big((\boldsymbol{I} - \boldsymbol{B}\boldsymbol{N}^{1/2})\boldsymbol{A}(\boldsymbol{I} - \boldsymbol{N}^{1/2}\boldsymbol{B}^\top) + \boldsymbol{B}\boldsymbol{B}^\top\big)\Big],$$

which implies

$$\mathrm{R}^{\mathrm{lin}}_{*,\mathrm{seq}}(\boldsymbol{N}) = \inf_{\boldsymbol{B}} \sup_{\substack{\boldsymbol{A}\succeq 0 \\ \mathrm{Tr}(\boldsymbol{A})\le R^2}} F(\boldsymbol{B},\boldsymbol{A}).$$

To characterize the exact solution of the above minimax problem, we first lower bound the minimax solution by

$$\inf_{\boldsymbol{B}} \sup_{\substack{\boldsymbol{A}\succeq 0 \\ \mathrm{Tr}(\boldsymbol{A})\le R^2}} F(\boldsymbol{B},\boldsymbol{A}) \ge \sup_{\substack{\boldsymbol{A}\succeq 0 \\ \mathrm{Tr}(\boldsymbol{A})\le R^2}} \inf_{\boldsymbol{B}} F(\boldsymbol{B},\boldsymbol{A}) = \sup_{\substack{\boldsymbol{A}\succeq 0 \\ \mathrm{Tr}(\boldsymbol{A})\le R^2}} F(\boldsymbol{B}_{\boldsymbol{A}},\boldsymbol{A}) = \mathcal{L}(\boldsymbol{N}),$$

where for any $\boldsymbol{A}$, we define $\boldsymbol{B}_{\boldsymbol{A}} = \arg\min_{\boldsymbol{B}} F(\boldsymbol{B},\boldsymbol{A})$. The observation is that for any $\boldsymbol{A}$ in the admissible regime, $\boldsymbol{B}_{\boldsymbol{A}}$ is achievable and the exact form can be characterized. To see this, note that $F(\boldsymbol{B},\boldsymbol{A})$ is quadratic in $\boldsymbol{B}$ and collecting all the terms depends on $\boldsymbol{B}$ into a quadratic form gives

$$F(\boldsymbol{B},\boldsymbol{A}) = \mathrm{Tr}[\boldsymbol{Q}\boldsymbol{P}(\boldsymbol{A},\boldsymbol{N})] + \mathrm{Tr}[\boldsymbol{Q}(\boldsymbol{B}-\boldsymbol{B}_{\boldsymbol{A}})\boldsymbol{V}_{\boldsymbol{A}}(\boldsymbol{B}-\boldsymbol{B}_{\boldsymbol{A}})^\top], \tag{11}$$

where we define

$$\begin{aligned}\boldsymbol{P}(\boldsymbol{A},\boldsymbol{N}) &= \boldsymbol{A}^{1/2}(\boldsymbol{I}+\boldsymbol{A}^{1/2}\boldsymbol{N}\boldsymbol{A}^{1/2})^{-1}\boldsymbol{A}^{1/2},\\ \boldsymbol{V}_{\boldsymbol{A}} &= \boldsymbol{I}+\boldsymbol{N}^{1/2}\boldsymbol{A}\boldsymbol{N}^{1/2}, \quad \boldsymbol{B}_{\boldsymbol{A}} = \boldsymbol{P}(\boldsymbol{A},\boldsymbol{N})\boldsymbol{N}^{1/2},\end{aligned}$$

and use the fact that

$$\begin{aligned}\boldsymbol{P}(\boldsymbol{A},\boldsymbol{N}) &= \boldsymbol{A}-\boldsymbol{A}\boldsymbol{N}^{1/2}(\boldsymbol{I}+\boldsymbol{N}^{1/2}\boldsymbol{A}\boldsymbol{N}^{1/2})^{-1}\boldsymbol{N}^{1/2}\boldsymbol{A},\\ \boldsymbol{P}(\boldsymbol{A},\boldsymbol{N})\boldsymbol{N}^{1/2} &= \boldsymbol{A}\boldsymbol{N}^{1/2}(\boldsymbol{I}+\boldsymbol{N}^{1/2}\boldsymbol{A}\boldsymbol{N}^{1/2})^{-1}.\end{aligned} \tag{12}$$

Next, note that for any given $\boldsymbol{B}$, the feasible regime of $\boldsymbol{A}$ is compact and the optimum is achievable. Thus, define

$$\boldsymbol{A}_* = \arg\max_{\substack{\boldsymbol{A}\succeq 0 \\ \mathrm{Tr}(\boldsymbol{A})\le R^2}} F(\boldsymbol{B}_{\boldsymbol{A}},\boldsymbol{A}) = \arg\max_{\substack{\boldsymbol{A}\succeq 0 \\ \mathrm{Tr}(\boldsymbol{A})\le R^2}} \mathrm{Tr}[\boldsymbol{Q}\boldsymbol{P}(\boldsymbol{A},\boldsymbol{N})].$$

It is clear by definition that $\boldsymbol{B}_* := \boldsymbol{B}_{\boldsymbol{A}_*} = \arg\min_{\boldsymbol{B}} F(\boldsymbol{B},\boldsymbol{A}_*)$. We then show that

$$\boldsymbol{A}_* \in \arg\max_{\substack{\boldsymbol{A}\succeq 0 \\ \mathrm{Tr}(\boldsymbol{A})\le R^2}} F(\boldsymbol{B}_*,\boldsymbol{A}),$$

which implies

$$\inf_{\boldsymbol{B}} \sup_{\substack{\boldsymbol{A}\succeq 0 \\ \mathrm{Tr}(\boldsymbol{A})\le R^2}} F(\boldsymbol{B},\boldsymbol{A}) = \sup_{\substack{\boldsymbol{A}\succeq 0 \\ \mathrm{Tr}(\boldsymbol{A})\le R^2}} \inf_{\boldsymbol{B}} F(\boldsymbol{B},\boldsymbol{A}) = \mathcal{L}(\boldsymbol{N}).$$

To see this, it is sufficient to show that for any feasible $\boldsymbol{A}$, $F(\boldsymbol{B}_*,\boldsymbol{A}) - F(\boldsymbol{B}_*,\boldsymbol{A}_*) \le 0$. Recall that $F(\boldsymbol{B},\boldsymbol{A}) = \mathrm{Tr}\Big[\boldsymbol{Q}\big((\boldsymbol{I}-\boldsymbol{B}\boldsymbol{N}^{1/2})\boldsymbol{A}(\boldsymbol{I}-\boldsymbol{N}^{1/2}\boldsymbol{B}^\top)+\boldsymbol{B}\boldsymbol{B}^\top\big)\Big]$. Then, a direct computation gives

$$F(\boldsymbol{B}_*,\boldsymbol{A}) - F(\boldsymbol{B}_*,\boldsymbol{A}_*) = \mathrm{Tr}[\boldsymbol{Q}(\boldsymbol{I}-\boldsymbol{B}_*\boldsymbol{N}^{1/2})(\boldsymbol{A}-\boldsymbol{A}_*)(\boldsymbol{I}-\boldsymbol{N}^{1/2}\boldsymbol{B}_*^\top)].$$

Let us define $f(\boldsymbol{A}) = \mathrm{Tr}[\boldsymbol{Q}\boldsymbol{P}(\boldsymbol{A},\boldsymbol{N})]$. Then, differentiating (12) in the direction $\boldsymbol{\Delta}$ gives

$$\begin{aligned}Df_{\boldsymbol{A}_*}[\boldsymbol{\Delta}] := \frac{\mathrm{d}}{\mathrm{d}t} f(\boldsymbol{A}_*+t\Delta)\big|_{t=0} &= \mathrm{Tr}[\boldsymbol{Q}(\boldsymbol{I}-\boldsymbol{P}(\boldsymbol{A}_*,\boldsymbol{N})\boldsymbol{N})\boldsymbol{\Delta}(\boldsymbol{I}-\boldsymbol{N}\boldsymbol{P}(\boldsymbol{A}_*,\boldsymbol{N}))]\\ &= \mathrm{Tr}[\boldsymbol{Q}(\boldsymbol{I}-\boldsymbol{B}_*\boldsymbol{N}^{1/2})\boldsymbol{\Delta}(\boldsymbol{I}-\boldsymbol{N}^{1/2}\boldsymbol{B}_*^\top)],\end{aligned}$$

which implies $Df_{\boldsymbol{A}_*}[\boldsymbol{A}-\boldsymbol{A}_*] \le 0$.

Since by definition $\boldsymbol{A}_* = \arg\max_{\substack{\boldsymbol{A}\succeq 0 \\ \mathrm{Tr}(\boldsymbol{A})\le R^2}} \mathrm{Tr}[\boldsymbol{Q}\boldsymbol{P}(\boldsymbol{A},\boldsymbol{N})]$, for any feasible direction $\boldsymbol{A}-\boldsymbol{A}_*$, we have $F(\boldsymbol{B}_*,\boldsymbol{A}) - F(\boldsymbol{B}_*,\boldsymbol{A}_*) = Df_{\boldsymbol{A}_*}[\boldsymbol{A}-\boldsymbol{A}_*] \le 0$, which concludes the proof.

Proof of the lower bound. We prove the lower bound via probabilistic methods. In particular, we construct a prior distribution $p(\boldsymbol{u})$ supported on $\{\boldsymbol{u} : \|\boldsymbol{u}\|_2 \le R\}$, and use the fact that

$$\mathrm{R}_{*,\mathrm{seq}}(\boldsymbol{N}) \ge \inf_{\widehat{\boldsymbol{u}}} \mathbb{E}_{\boldsymbol{u}\sim \mathrm{p}_{\boldsymbol{u}}} \mathbb{E}_{\boldsymbol{u}} \|\boldsymbol{Q}^{1/2}(\widehat{\boldsymbol{u}}(\boldsymbol{y}_{\boldsymbol{N},\mathrm{seq}}) - \boldsymbol{u})\|_2^2.$$

For the construction, fix a feasible finite-rank $\boldsymbol{A} = \sum_{\ell=1}^{r} a_\ell \boldsymbol{v}_\ell \boldsymbol{v}_\ell^\top \in \mathbb{R}^{d\times d}$, where the $\boldsymbol{v}_\ell$ are orthonormal, $a_\ell > 0$, and $\sum_\ell a_\ell \le R^2$. Let $\boldsymbol{T}$ have columns $\sqrt{a_\ell}\boldsymbol{v}_\ell$, so $\boldsymbol{T}\boldsymbol{T}^\top = \boldsymbol{A}$ and $\boldsymbol{T}^\top\boldsymbol{T} = \mathrm{Diag}(a_1,\dots,a_r)$. Consider $\boldsymbol{\omega}$ independent coordinates with density

$$q(w) = \cos^2(\pi w/2)\mathbf{1}\{|w| \le 1\},$$

which have the following properties

$$\int_{-1}^{1} q'(w)\,dw = 0, \qquad \int_{-1}^{1} \frac{q'(w)^2}{q(w)}\,dw = \pi^2 \int_{-1}^{1} \sin^2(\pi w/2)\,dw = \pi^2.$$

Let $\boldsymbol{u} = \boldsymbol{T}\boldsymbol{\omega}$. This prior is supported on the parameter ball, since

$$\|\boldsymbol{u}\|_2^2 = \sum_{\ell=1}^{r} a_\ell \omega_\ell^2 \le \sum_{\ell=1}^{r} a_\ell = \mathrm{Tr}(\boldsymbol{A}) \le R^2.$$

Next, for such prior distribution $p(\boldsymbol{u})$, we provide a lower bound on the minimax risk. Let $p(\boldsymbol{y}_{\boldsymbol{N},\mathrm{seq}}|\boldsymbol{u})$ be the probability distribution of $\boldsymbol{y}_{\boldsymbol{N},\mathrm{seq}}$ given $\boldsymbol{u}$. Since the randomness of $\boldsymbol{u}$ only comes from $\boldsymbol{\omega}$, we rewrite the joint distribution as $p(\boldsymbol{y}_{\boldsymbol{N},\mathrm{seq}}, \boldsymbol{\omega})$ and define the score function w.r.t. $\boldsymbol{\omega}$ as

$$\zeta = \nabla_{\boldsymbol{\omega}} \log p(\boldsymbol{y}_{\boldsymbol{N},\mathrm{seq}}, \boldsymbol{\omega}) = \nabla_{\boldsymbol{\omega}} \log p(\boldsymbol{y}_{\boldsymbol{N},\mathrm{seq}}|\boldsymbol{\omega}) + \nabla_{\boldsymbol{\omega}} \log q(\boldsymbol{\omega}).$$

We then compute two quantities:

$$\begin{aligned}
\mathbb{E}[\boldsymbol{e}\zeta^\top] &= \iint \boldsymbol{e}\big(\nabla_{\boldsymbol{\omega}} p(\boldsymbol{y}_{\boldsymbol{N},\mathrm{seq}}, \boldsymbol{\omega})\big)^\top \, d\mu(\boldsymbol{y}_{\boldsymbol{N},\mathrm{seq}})\, d\boldsymbol{\omega} \\
&= -\iint \frac{\partial \boldsymbol{e}}{\partial \boldsymbol{\omega}} p(\boldsymbol{y}_{\boldsymbol{N},\mathrm{seq}}, \boldsymbol{\omega})\, d\mu(\boldsymbol{y}_{\boldsymbol{N},\mathrm{seq}})\, d\boldsymbol{\omega} \\
&= \boldsymbol{Q}^{1/2}\boldsymbol{T} \iint p(\boldsymbol{y}_{\boldsymbol{N},\mathrm{seq}}, \boldsymbol{\omega})\, d\mu(\boldsymbol{y}_{\boldsymbol{N},\mathrm{seq}})\, d\boldsymbol{\omega} = \boldsymbol{Q}^{1/2}\boldsymbol{T},
\end{aligned}$$

$$\begin{aligned}
\boldsymbol{J} := \mathbb{E}[\zeta\zeta^\top] &= \mathbb{E}\big[(\nabla_{\boldsymbol{\omega}} \log p(\boldsymbol{y}_{\boldsymbol{N},\mathrm{seq}} \mid \boldsymbol{\omega}))(\nabla_{\boldsymbol{\omega}} \log p(\boldsymbol{y}_{\boldsymbol{N},\mathrm{seq}} \mid \boldsymbol{\omega}))^\top\big] \\
&\quad + \mathbb{E}\big[(\nabla_{\boldsymbol{\omega}} \log q(\boldsymbol{\omega}))(\nabla_{\boldsymbol{\omega}} \log q(\boldsymbol{\omega}))^\top\big] \\
&\quad + \underbrace{\mathbb{E}\big[(\nabla_{\boldsymbol{\omega}} \log p(\boldsymbol{y}_{\boldsymbol{N},\mathrm{seq}} \mid \boldsymbol{\omega}))(\nabla_{\boldsymbol{\omega}} \log q(\boldsymbol{\omega}))^\top\big]}_{=0} \\
&\quad + \underbrace{\mathbb{E}\big[(\nabla_{\boldsymbol{\omega}} \log q(\boldsymbol{\omega}))(\nabla_{\boldsymbol{\omega}} \log p(\boldsymbol{y}_{\boldsymbol{N},\mathrm{seq}} \mid \boldsymbol{\omega}))^\top\big]}_{=0} \\
&= \boldsymbol{T}^\top \boldsymbol{N} \boldsymbol{T} + \pi^2 \boldsymbol{I}_r,
\end{aligned}$$

where we use the identities

$$\mathbb{E}\big[\nabla_{\boldsymbol{\omega}} \log p(\boldsymbol{y}_{\boldsymbol{N},\mathrm{seq}} \mid \boldsymbol{\omega}) \,\big|\, \boldsymbol{\omega}\big] = \int \nabla_{\boldsymbol{\omega}} p(\boldsymbol{y}_{\boldsymbol{N},\mathrm{seq}} \mid \boldsymbol{\omega})\, d\mu(\boldsymbol{y}_{\boldsymbol{N},\mathrm{seq}}) = \nabla_{\boldsymbol{\omega}} 1 = 0,$$

$$\mathbb{E}\left[\frac{q'(\omega_\ell)}{q(\omega_\ell)}\right] = \int_{-1}^{1} q'(w)\,dw = q(1) - q(-1) = 0,$$

$$\mathbb{E}\left[\frac{q'(\omega_\ell)}{q(\omega_\ell)}\frac{q'(\omega_m)}{q(\omega_m)}\right] = \begin{cases} \displaystyle\int_{-1}^{1} \frac{q'(w)^2}{q(w)}\,dw = \pi^2 \int_{-1}^{1} \sin^2(\pi w/2)\,dw = \pi^2, & \ell = m, \\ \displaystyle\left(\int_{-1}^{1} q'(w)\,dw\right)^2 = 0, & \ell \ne m. \end{cases}$$

For an arbitrary estimator $\widehat{\boldsymbol{u}}$, define $\boldsymbol{e} = \boldsymbol{Q}^{1/2}(\widehat{\boldsymbol{u}} - \boldsymbol{T}\boldsymbol{\omega})$, and it is sufficient to lower bound $\mathbb{E}[\|\boldsymbol{e}\|_2^2]$.

Note that $\mathbb{E}[\|\boldsymbol{e} - \boldsymbol{Q}^{1/2}\boldsymbol{T}\boldsymbol{J}^{-1}\zeta\|_2^2] \geq 0$. Expanding a nonnegative square yields

$$\begin{aligned}
0 &\leq \mathbb{E}\|\boldsymbol{e} - \boldsymbol{Q}^{1/2}\boldsymbol{T}\boldsymbol{J}^{-1}\boldsymbol{\zeta}\|_2^2 \\
&= \mathbb{E}[\|\boldsymbol{e}\|_2^2] - 2\mathrm{Tr}(\boldsymbol{Q}^{1/2}\boldsymbol{T}\boldsymbol{J}^{-1}\mathbb{E}[\zeta\boldsymbol{e}^\top]) + \mathrm{Tr}[\boldsymbol{J}^{-1}\boldsymbol{T}^\top\boldsymbol{Q}\boldsymbol{T}\boldsymbol{J}^{-1}]\mathbb{E}[\zeta\zeta^\top] \\
&= \mathbb{E}\|\boldsymbol{e}\|_2^2 - \mathrm{Tr}[\boldsymbol{Q}\boldsymbol{T}(\boldsymbol{T}^\top\boldsymbol{N}\boldsymbol{T} + \pi^2\boldsymbol{I}_r)^{-1}\boldsymbol{T}^\top].
\end{aligned}$$

Recall by definition $\boldsymbol{A} = \boldsymbol{T}\boldsymbol{T}^\top$. Then, the resulting lower bound is

$$\begin{aligned}
\mathbb{E}\|\boldsymbol{e}\|_2^2 \geq& \mathrm{Tr}[\boldsymbol{Q}\boldsymbol{T}(\boldsymbol{T}^\top\boldsymbol{N}\boldsymbol{T} + \pi^2\boldsymbol{I}_r)^{-1}\boldsymbol{T}^\top] \\
\geq& \pi^{-2}\mathrm{Tr}[\boldsymbol{Q}\boldsymbol{P}(\boldsymbol{A}, \boldsymbol{N})].
\end{aligned}$$

Taking the maximal $\boldsymbol{A}$ concludes the proof.

A.2 Proof of Theorem 1

Recall the normalized weights $\boldsymbol{u} = \boldsymbol{S}^{1/2}\boldsymbol{\theta}$ and $\widehat{\boldsymbol{u}} = \boldsymbol{S}^{1/2}\widehat{\boldsymbol{\theta}}$. Denote the full regression data by $\mathcal{D} = \{(\boldsymbol{X}_i, \boldsymbol{y}_i)\}_{i=1}^K$. The parameter constraint and the test error can be written in the same form as for the Gaussian sequence model, which gives

$$\mathrm{R}_*(n_1, \ldots, n_K) = \inf_{\widehat{\boldsymbol{u}}} \sup_{\|\boldsymbol{u}\|_2 \leq R} \mathbb{E}_{\boldsymbol{u}}\|\boldsymbol{Q}^{1/2}(\widehat{\boldsymbol{u}}(\mathcal{D}) - \boldsymbol{u})\|_2^2.$$

Here the expectation includes both the designs and the observation noise, and the infimum is over all estimators based on $\mathcal{D}$. We also recall

$$\boldsymbol{M} = \sum_{i=1}^K \frac{n_i}{\sigma_{\varepsilon_i}^2}\boldsymbol{H}_i, \qquad E_0 = R^2 \max_{1\leq i\leq K} \frac{\|\boldsymbol{H}_i\|_{\mathrm{op}}}{\sigma_{\varepsilon_i}^2}, \qquad \upsilon = 1 + \kappa E_0.$$

Proof of the upper bound. Our goal is to construct a regression estimator whose worst-case risk is bounded by $\upsilon\mathcal{L}(\boldsymbol{M})$, where $\upsilon = 1 + \kappa E_0$ and κ is a finite constant depending only on the concentration constants in Assumption 1. Since the minimax risk is the infimum over all estimators, it suffices to construct $\widehat{\boldsymbol{u}}_{\mathrm{reg}}$ such that

$$\sup_{\|\boldsymbol{u}\|_2 \leq R} \mathbb{E}_{\boldsymbol{u}}\|\boldsymbol{Q}^{1/2}(\widehat{\boldsymbol{u}}_{\mathrm{reg}} - \boldsymbol{u})\|_2^2 \leq \upsilon\mathcal{L}(\boldsymbol{M}).$$

We do so by comparing its risk with that of an optimal linear estimator in an auxiliary Gaussian sequence model. All regression expectations below include both the random designs and the observation noise.

Define the statistic

$$\boldsymbol{z} = \boldsymbol{S}^{-1/2}\sum_{i=1}^K \frac{1}{\sigma_{\varepsilon_i}^2}\boldsymbol{X}_i^\top\boldsymbol{y}_i.$$

We will show that, uniformly over $\|\boldsymbol{u}\|_2 \leq R$,

$$\mathbb{E}_{\boldsymbol{u}}\boldsymbol{z} = \boldsymbol{M}\boldsymbol{u}, \qquad \mathrm{Cov}_{\boldsymbol{u}}(\boldsymbol{z}) \preceq \upsilon\boldsymbol{M}. \tag{13}$$

We first explain how these moment bounds yield the desired estimator, and then verify them.

Set $\boldsymbol{N} = \boldsymbol{M}/\upsilon$. By (13), the rescaled statistic $\boldsymbol{z}/\upsilon$ has mean $\boldsymbol{N}\boldsymbol{u}$ and covariance at most $\boldsymbol{N}$. To see its connection with the Gaussian sequence model, recall that

$$\boldsymbol{y}_{\boldsymbol{N},\mathrm{seq}} = \boldsymbol{N}^{1/2}\boldsymbol{u} + \boldsymbol{g}, \qquad \boldsymbol{g} \sim \mathcal{N}(0, \boldsymbol{I}).$$

Multiplying the Gaussian observation by $\boldsymbol{N}^{1/2}$ gives

$$\boldsymbol{N}^{1/2}\boldsymbol{y}_{\boldsymbol{N},\mathrm{seq}} = \boldsymbol{N}\boldsymbol{u} + \boldsymbol{N}^{1/2}\boldsymbol{g}.$$

The moment bounds (13) gives

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{u}}[\boldsymbol{z}/\upsilon] &= \boldsymbol{N}\boldsymbol{u} = \mathbb{E}_{\boldsymbol{u}}[\boldsymbol{N}^{1/2}\boldsymbol{y}_{\boldsymbol{N},\mathrm{seq}}], \\
\mathrm{Cov}_{\boldsymbol{u}}(\boldsymbol{z}/\upsilon) &\preceq \boldsymbol{N} = \mathrm{Cov}_{\boldsymbol{u}}(\boldsymbol{N}^{1/2}\boldsymbol{y}_{\boldsymbol{N},\mathrm{seq}}).
\end{aligned}$$

This motivates us to apply the same linear estimators constructed for the sequence model to the regression model here. To choose this linear transformation, take

$$\boldsymbol{A}_* \in \underset{\substack{\boldsymbol{A}\succeq 0\\ \mathrm{Tr}(\boldsymbol{A})\le R^2}}{\arg\max}\ \mathrm{Tr}[\boldsymbol{Q}\boldsymbol{P}(\boldsymbol{A},\boldsymbol{N})], \qquad \boldsymbol{B}_* = \boldsymbol{P}(\boldsymbol{A}_*,\boldsymbol{N})\boldsymbol{N}^{1/2}.$$

By (10) and the optimizer identified in (11), the Gaussian estimator

$$\boldsymbol{B}_*\boldsymbol{y}_{\boldsymbol{N},\mathrm{seq}} = \boldsymbol{P}(\boldsymbol{A}_*,\boldsymbol{N})(\boldsymbol{N}^{1/2}\boldsymbol{y}_{\boldsymbol{N},\mathrm{seq}})$$

has worst-case risk $\mathcal{L}(\boldsymbol{N})$. The matrices defining this estimator depend only on the model quantities and the parameter constraint, not on the unknown $\boldsymbol{u}$. Replacing its input $\boldsymbol{N}^{1/2}\boldsymbol{y}_{\boldsymbol{N},\mathrm{seq}}$ by $\boldsymbol{z}/v$, define

$$\widehat{\boldsymbol{u}}_{\mathrm{reg}} = \frac{1}{v}\boldsymbol{P}(\boldsymbol{A}_*,\boldsymbol{N})\boldsymbol{z}, \qquad \widehat{\boldsymbol{\theta}}_{\mathrm{reg}} = \boldsymbol{S}^{-1/2}\widehat{\boldsymbol{u}}_{\mathrm{reg}}.$$

Standard bias-variance decomposition for linear estimators gives, for every $\|\boldsymbol{u}\|_2 \le R$,

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{u}}\|\boldsymbol{Q}^{1/2}(\widehat{\boldsymbol{u}}_{\mathrm{reg}} - \boldsymbol{u})\|_2^2 \\
&\quad = \|\boldsymbol{Q}^{1/2}(\boldsymbol{P}(\boldsymbol{A}_*,\boldsymbol{N})\boldsymbol{N} - \boldsymbol{I})\boldsymbol{u}\|_2^2 \\
&\qquad + v^{-2}\mathrm{Tr}[\boldsymbol{Q}\boldsymbol{P}(\boldsymbol{A}_*,\boldsymbol{N})\,\mathrm{Cov}_{\boldsymbol{u}}(\boldsymbol{z})\boldsymbol{P}(\boldsymbol{A}_*,\boldsymbol{N})] \\
&\quad \le \|\boldsymbol{Q}^{1/2}(\boldsymbol{P}(\boldsymbol{A}_*,\boldsymbol{N})\boldsymbol{N} - \boldsymbol{I})\boldsymbol{u}\|_2^2 + \mathrm{Tr}[\boldsymbol{Q}\boldsymbol{P}(\boldsymbol{A}_*,\boldsymbol{N})\boldsymbol{N}\boldsymbol{P}(\boldsymbol{A}_*,\boldsymbol{N})] \\
&\quad = \mathbb{E}_{\boldsymbol{u}}\|\boldsymbol{Q}^{1/2}(\boldsymbol{B}_*\boldsymbol{y}_{\boldsymbol{N},\mathrm{seq}} - \boldsymbol{u})\|_2^2.
\end{aligned}$$

Taking the supremum over the parameter ball gives

$$\mathrm{R}_*(n_1,\dots,n_K) \le \sup_{\|\boldsymbol{u}\|_2\le R} \mathbb{E}_{\boldsymbol{u}}\|\boldsymbol{Q}^{1/2}(\widehat{\boldsymbol{u}}_{\mathrm{reg}} - \boldsymbol{u})\|_2^2 \le \mathcal{L}(\boldsymbol{N}) = \mathcal{L}(\boldsymbol{M}/v).$$

It remains to verify (13) and compare $\mathcal{L}(\boldsymbol{M}/v)$ with $\mathcal{L}(\boldsymbol{M})$. For the mean, using $\boldsymbol{y}_i = \boldsymbol{X}_i\boldsymbol{S}^{-1/2}\boldsymbol{u} + \varepsilon_i$ and the independence and zero mean of the observation noise gives

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{u}}\boldsymbol{z} &= \boldsymbol{S}^{-1/2}\sum_{i=1}^{K}\frac{1}{\sigma_{\varepsilon_i}^2}\mathbb{E}[\boldsymbol{X}_i^\top\boldsymbol{X}_i]\boldsymbol{S}^{-1/2}\boldsymbol{u} \\
&= \sum_{i=1}^{K}\frac{n_i}{\sigma_{\varepsilon_i}^2}\boldsymbol{H}_i\boldsymbol{u} = \boldsymbol{M}\boldsymbol{u}.
\end{aligned}$$

For the covariance, consider one observation $(\boldsymbol{x}, y)$ from domain i, where $y = \langle \boldsymbol{x}, \boldsymbol{S}^{-1/2}\boldsymbol{u}\rangle + \varepsilon$. Its contribution in a direction $\boldsymbol{w}$ is

$$\langle \boldsymbol{x}, \boldsymbol{S}^{-1/2}\boldsymbol{w}\rangle y = \langle \boldsymbol{x}, \boldsymbol{S}^{-1/2}\boldsymbol{w}\rangle\langle \boldsymbol{x}, \boldsymbol{S}^{-1/2}\boldsymbol{u}\rangle + \langle \boldsymbol{x}, \boldsymbol{S}^{-1/2}\boldsymbol{w}\rangle\varepsilon.$$

The first term fluctuates because the inputs are random, and controlling its variance requires a fourth-moment bound, which is obtained from Assumption 1. More precisely, for $\boldsymbol{x} \sim P_i$ and any $\boldsymbol{w}$ with $\boldsymbol{w}^\top\boldsymbol{H}_i\boldsymbol{w} > 0$, applying that assumption to $\boldsymbol{S}^{-1/2}\boldsymbol{w}\boldsymbol{w}^\top\boldsymbol{S}^{-1/2}$ gives

$$P_i\Big[\Big|\langle \boldsymbol{x}, \boldsymbol{S}^{-1/2}\boldsymbol{w}\rangle^2 - \boldsymbol{w}^\top\boldsymbol{H}_i\boldsymbol{w}\Big| > t\boldsymbol{w}^\top\boldsymbol{H}_i\boldsymbol{w}\Big] \le C\exp\Big(-ct^{1/\eta}\Big), \qquad t > 0,$$

where C, c, η are the concentration constants in the assumption. Integrating this tail bound yields

$$\begin{aligned}
\mathbb{E}_{P_i}\langle \boldsymbol{x}, \boldsymbol{S}^{-1/2}\boldsymbol{w}\rangle^4 &= (\boldsymbol{w}^\top\boldsymbol{H}_i\boldsymbol{w})^2 + \mathbb{E}_{P_i}\Big[\big(\langle \boldsymbol{x}, \boldsymbol{S}^{-1/2}\boldsymbol{w}\rangle^2 - \boldsymbol{w}^\top\boldsymbol{H}_i\boldsymbol{w}\big)^2\Big] \\
&\le \left(1 + 2C\int_0^\infty t\exp\Big(-ct^{1/\eta}\Big)\,dt\right)(\boldsymbol{w}^\top\boldsymbol{H}_i\boldsymbol{w})^2 \\
&= \kappa(\boldsymbol{w}^\top\boldsymbol{H}_i\boldsymbol{w})^2.
\end{aligned}$$

This defines a finite constant κ depending only on the concentration constants. If $\boldsymbol{w}^\top \boldsymbol{H}_i \boldsymbol{w} = 0$, then $\langle \boldsymbol{x}, \boldsymbol{S}^{-1/2}\boldsymbol{w}\rangle = 0$ almost surely, so the same bound holds. Applying Cauchy–Schwarz now gives

$$\begin{aligned}\mathbb{E}_{P_i}\Big[\langle \boldsymbol{x}, \boldsymbol{S}^{-1/2}\boldsymbol{w}\rangle^2 \langle \boldsymbol{x}, \boldsymbol{S}^{-1/2}\boldsymbol{u}\rangle^2\Big] &\le \Big(\mathbb{E}_{P_i}\langle \boldsymbol{x}, \boldsymbol{S}^{-1/2}\boldsymbol{w}\rangle^4 \mathbb{E}_{P_i}\langle \boldsymbol{x}, \boldsymbol{S}^{-1/2}\boldsymbol{u}\rangle^4\Big)^{1/2} \\ &\le \kappa(\boldsymbol{w}^\top \boldsymbol{H}_i \boldsymbol{w})(\boldsymbol{u}^\top \boldsymbol{H}_i \boldsymbol{u}).\end{aligned}$$

Since ε is independent of $\boldsymbol{x}$, with mean zero and variance $\sigma_{\varepsilon_i}^2$, we have

$$\begin{aligned}\mathrm{Var}_{\boldsymbol{u}}\Big(\langle \boldsymbol{x}, \boldsymbol{S}^{-1/2}\boldsymbol{w}\rangle y\Big) &= \mathbb{E}_{P_i}\Big[\langle \boldsymbol{x}, \boldsymbol{S}^{-1/2}\boldsymbol{w}\rangle^2 \langle \boldsymbol{x}, \boldsymbol{S}^{-1/2}\boldsymbol{u}\rangle^2\Big] - (\boldsymbol{w}^\top \boldsymbol{H}_i \boldsymbol{u})^2 + \sigma_{\varepsilon_i}^2 \boldsymbol{w}^\top \boldsymbol{H}_i \boldsymbol{w} \\ &\le \big(\kappa \boldsymbol{u}^\top \boldsymbol{H}_i \boldsymbol{u} + \sigma_{\varepsilon_i}^2\big)\boldsymbol{w}^\top \boldsymbol{H}_i \boldsymbol{w} \\ &\le v\sigma_{\varepsilon_i}^2 \boldsymbol{w}^\top \boldsymbol{H}_i \boldsymbol{w}.\end{aligned}$$

The last inequality holds uniformly over the parameter ball because

$$\frac{\boldsymbol{u}^\top \boldsymbol{H}_i \boldsymbol{u}}{\sigma_{\varepsilon_i}^2} \le \frac{R^2 \|\boldsymbol{H}_i\|_{\mathrm{op}}}{\sigma_{\varepsilon_i}^2} \le E_0, \qquad \|\boldsymbol{u}\|_2 \le R,$$

and $v = 1 + \kappa E_0$. Thus, the additional variance caused by the random inputs is controlled uniformly over all admissible parameters.

Let $\boldsymbol{x}_{i,j}^\top$ denote the j-th row of $\boldsymbol{X}_i$. Independence of the observations implies

$$\begin{aligned}\boldsymbol{w}^\top \mathrm{Cov}_{\boldsymbol{u}}(\boldsymbol{z})\boldsymbol{w} &= \sum_{i=1}^{K}\sum_{j=1}^{n_i} \frac{1}{\sigma_{\varepsilon_i}^4} \mathrm{Var}_{\boldsymbol{u}}\Big(\langle \boldsymbol{x}_{i,j}, \boldsymbol{S}^{-1/2}\boldsymbol{w}\rangle y_{i,j}\Big) \\ &\le v \sum_{i=1}^{K} \frac{n_i}{\sigma_{\varepsilon_i}^2} \boldsymbol{w}^\top \boldsymbol{H}_i \boldsymbol{w} \\ &= v\boldsymbol{w}^\top \boldsymbol{M}\boldsymbol{w}.\end{aligned}$$

Since this holds for every $\boldsymbol{w}$, it proves $\mathrm{Cov}_{\boldsymbol{u}}(\boldsymbol{z}) \preceq v\boldsymbol{M}$ and completes the verification of (13).

Finally, we compare the Gaussian linear minimax risks at $\boldsymbol{M}/v$ and $\boldsymbol{M}$. Since $v \ge 1$, for every feasible $\boldsymbol{A}$,

$$\boldsymbol{I} + v^{-1}\boldsymbol{A}^{1/2}\boldsymbol{M}\boldsymbol{A}^{1/2} \succeq v^{-1}(\boldsymbol{I} + \boldsymbol{A}^{1/2}\boldsymbol{M}\boldsymbol{A}^{1/2}).$$

Inversion reverses the PSD order, and congruence by $\boldsymbol{A}^{1/2}$ therefore gives

$$\boldsymbol{P}(\boldsymbol{A}, \boldsymbol{M}/v) \preceq v\boldsymbol{P}(\boldsymbol{A}, \boldsymbol{M}).$$

Taking the trace against $\boldsymbol{Q}$ and then the supremum over feasible $\boldsymbol{A}$ yields $\mathcal{L}(\boldsymbol{M}/v) \le v\mathcal{L}(\boldsymbol{M})$. Combining this comparison with the risk bound for the constructed estimator proves

$$\mathrm{R}_*(n_1, \ldots, n_K) \le \mathcal{L}(\boldsymbol{M}/v) \le v\mathcal{L}(\boldsymbol{M}). \tag{14}$$

If the designs are Gaussian, the same construction admits a sharper constant. In this case, the exact fourth-moment identity gives

$$\begin{aligned}\mathrm{Var}_{\boldsymbol{u}}\Big(\langle \boldsymbol{x}, \boldsymbol{S}^{-1/2}\boldsymbol{w}\rangle y\Big) &= (\boldsymbol{u}^\top \boldsymbol{H}_i \boldsymbol{u} + \sigma_{\varepsilon_i}^2)\boldsymbol{w}^\top \boldsymbol{H}_i \boldsymbol{w} + (\boldsymbol{w}^\top \boldsymbol{H}_i \boldsymbol{u})^2 \\ &\le (2\boldsymbol{u}^\top \boldsymbol{H}_i \boldsymbol{u} + \sigma_{\varepsilon_i}^2)\boldsymbol{w}^\top \boldsymbol{H}_i \boldsymbol{w} \\ &\le (1 + 2E_0)\sigma_{\varepsilon_i}^2 \boldsymbol{w}^\top \boldsymbol{H}_i \boldsymbol{w},\end{aligned}$$

where we used $(\boldsymbol{w}^\top \boldsymbol{H}_i \boldsymbol{u})^2 \le (\boldsymbol{w}^\top \boldsymbol{H}_i \boldsymbol{w})(\boldsymbol{u}^\top \boldsymbol{H}_i \boldsymbol{u})$. Consequently, (13) holds with $v = 1 + 2E_0$, and repeating the estimator construction and risk comparison with this choice gives (14) with the sharper factor.

Proof of the lower bound. We aim to prove $\mathrm{R}_*(n_1, \dots, n_K) \geq \pi^{-2}\mathcal{L}(\boldsymbol{M})$, and follow the same construction proof as for the Gaussian sequence model.

We recall the construction below. Fix a nonzero feasible matrix $\boldsymbol{A}$ with spectral decomposition

$$\boldsymbol{A} = \sum_{\ell=1}^{r} a_\ell \boldsymbol{v}_\ell \boldsymbol{v}_\ell^\top, \qquad a_\ell > 0, \qquad \sum_{\ell=1}^{r} a_\ell \leq R^2,$$

where the $\boldsymbol{v}_\ell$ are orthonormal. As in Appendix A.1, let $\boldsymbol{T}$ have columns $\sqrt{a_\ell}\boldsymbol{v}_\ell$, and set $\boldsymbol{u} = \boldsymbol{T}\boldsymbol{\omega}$, where the coordinates of $\boldsymbol{\omega}$ are independent with density

$$q(w) = \cos^2(\pi w/2)\mathbf{1}\{|w| \leq 1\}.$$

Then $\boldsymbol{T}\boldsymbol{T}^\top = \boldsymbol{A}$, and this prior is supported on the parameter ball because

$$\|\boldsymbol{u}\|_2^2 = \sum_{\ell=1}^{r} a_\ell \omega_\ell^2 \leq \sum_{\ell=1}^{r} a_\ell = \mathrm{Tr}(\boldsymbol{A}) \leq R^2.$$

Consequently, for every estimator $\widehat{\boldsymbol{u}}$,

$$\sup_{\|\boldsymbol{u}\|_2 \leq R} \mathbb{E}_{\boldsymbol{u}} \|\boldsymbol{Q}^{1/2}(\widehat{\boldsymbol{u}}(\mathcal{D}) - \boldsymbol{u})\|_2^2 \geq \mathbb{E}_{\boldsymbol{\omega}} \mathbb{E}_{\mathcal{D}|\boldsymbol{\omega}} \|\boldsymbol{Q}^{1/2}(\widehat{\boldsymbol{u}}(\mathcal{D}) - \boldsymbol{T}\boldsymbol{\omega})\|_2^2,$$

where $\mathcal{D}$ denotes the whole dataset.

Write $q(\boldsymbol{\omega}) = \prod_{\ell=1}^{r} q(\omega_\ell)$ and $p(\mathcal{D}, \boldsymbol{\omega}) = p(\mathcal{D} \mid \boldsymbol{\omega})q(\boldsymbol{\omega})$. Here the density of the data is taken with respect to the product of the design law and the Lebesgue measure for the responses, denoted by μ. This reference measure is independent of $\boldsymbol{\omega}$. Define the joint score and estimation error by

$$\zeta = \nabla_{\boldsymbol{\omega}} \log p(\mathcal{D}, \boldsymbol{\omega}), \qquad \boldsymbol{e} = \boldsymbol{Q}^{1/2}(\widehat{\boldsymbol{u}}(\mathcal{D}) - \boldsymbol{T}\boldsymbol{\omega}).$$

We show that the following identity holds in the regression case as well, and the rest follows exactly as in Appendix A.1:

$$\boldsymbol{J} := \mathbb{E}[\zeta\zeta^\top] = \boldsymbol{T}^\top \boldsymbol{M}\boldsymbol{T} + \pi^2 \boldsymbol{I}_r, \qquad \mathbb{E}[\boldsymbol{e}\zeta^\top] = \boldsymbol{Q}^{1/2}\boldsymbol{T}.$$

First, we compute $\mathbb{E}[\zeta\zeta^\top]$. The joint score decomposes as

$$\zeta = \nabla_{\boldsymbol{\omega}} \log p(\mathcal{D} \mid \boldsymbol{\omega}) + \nabla_{\boldsymbol{\omega}} \log q(\boldsymbol{\omega}).$$

The design distribution does not depend on $\boldsymbol{\omega}$, and the conditional response distribution is

$$\boldsymbol{y}_i \mid \boldsymbol{X}_i, \boldsymbol{\omega} \sim \mathcal{N}(\boldsymbol{X}_i \boldsymbol{S}^{-1/2}\boldsymbol{T}\boldsymbol{\omega}, \sigma_{\varepsilon_i}^2 \boldsymbol{I}_{n_i}), \qquad i = 1, \dots, K.$$

Differentiating the conditional Gaussian likelihood therefore gives

$$\begin{aligned}\nabla_{\boldsymbol{\omega}} \log p(\mathcal{D} \mid \boldsymbol{\omega}) &= \sum_{i=1}^{K} \frac{1}{\sigma_{\varepsilon_i}^2} \boldsymbol{T}^\top \boldsymbol{S}^{-1/2} \boldsymbol{X}_i^\top (\boldsymbol{y}_i - \boldsymbol{X}_i \boldsymbol{S}^{-1/2}\boldsymbol{T}\boldsymbol{\omega}) \\ &= \sum_{i=1}^{K} \frac{1}{\sigma_{\varepsilon_i}^2} \boldsymbol{T}^\top \boldsymbol{S}^{-1/2} \boldsymbol{X}_i^\top \varepsilon_i.\end{aligned}$$

This expression has mean zero conditionally on the designs and $\boldsymbol{\omega}$. The cross terms between domains vanish by independence and the zero mean of the observation noises. Using their covariances and then averaging over the designs yields

$$\begin{aligned}&\mathbb{E}\big[(\nabla_{\boldsymbol{\omega}} \log p(\mathcal{D} \mid \boldsymbol{\omega}))(\nabla_{\boldsymbol{\omega}} \log p(\mathcal{D} \mid \boldsymbol{\omega}))^\top \,\big|\, \boldsymbol{\omega}\big] \\ &= \sum_{i=1}^{K} \frac{1}{\sigma_{\varepsilon_i}^4} \boldsymbol{T}^\top \boldsymbol{S}^{-1/2} \mathbb{E}[\boldsymbol{X}_i^\top \varepsilon_i \varepsilon_i^\top \boldsymbol{X}_i] \boldsymbol{S}^{-1/2}\boldsymbol{T} \\ &= \sum_{i=1}^{K} \frac{1}{\sigma_{\varepsilon_i}^2} \boldsymbol{T}^\top \boldsymbol{S}^{-1/2} \mathbb{E}[\boldsymbol{X}_i^\top \boldsymbol{X}_i] \boldsymbol{S}^{-1/2}\boldsymbol{T} \\ &= \sum_{i=1}^{K} \frac{n_i}{\sigma_{\varepsilon_i}^2} \boldsymbol{T}^\top \boldsymbol{H}_i \boldsymbol{T} = \boldsymbol{T}^\top \boldsymbol{M}\boldsymbol{T}.\end{aligned}$$

This is precisely the average Fisher information obtained in the Gaussian sequence model with $\boldsymbol{N} = \boldsymbol{M}$ under the same parameterization $\boldsymbol{u} = \boldsymbol{T}\boldsymbol{\omega}$. The calculation uses only the second moments of the designs; the Gaussian assumption is used for the observation noise.

For the prior contribution, the one-dimensional density satisfies

$$\int_{-1}^{1} q'(w)\,dw = 0, \qquad \int_{-1}^{1} \frac{q'(w)^2}{q(w)}\,dw = \pi^2 \int_{-1}^{1} \sin^2(\pi w/2)\,dw = \pi^2.$$

Independence of the prior coordinates therefore gives

$$\mathbb{E}[\nabla_{\boldsymbol{\omega}} \log q(\boldsymbol{\omega})] = 0, \qquad \mathbb{E}\big[(\nabla_{\boldsymbol{\omega}} \log q(\boldsymbol{\omega}))(\nabla_{\boldsymbol{\omega}} \log q(\boldsymbol{\omega}))^\top\big] = \pi^2 \boldsymbol{I}_r.$$

Moreover, the likelihood score has mean zero conditionally on $\boldsymbol{\omega}$, so its cross terms with the prior score vanish. Combining the two contributions proves

$$\boldsymbol{J} = \mathbb{E}[\zeta\zeta^\top] = \boldsymbol{T}^\top \boldsymbol{M}\boldsymbol{T} + \pi^2 \boldsymbol{I}_r.$$

Next, we verify $\mathbb{E}[\boldsymbol{e}\zeta^\top] = \boldsymbol{Q}^{1/2}\boldsymbol{T}$. With the data held fixed, $\widehat{\boldsymbol{u}}(\mathcal{D})$ does not depend on $\boldsymbol{\omega}$, and hence

$$\frac{\partial \boldsymbol{e}}{\partial \boldsymbol{\omega}} = -\boldsymbol{Q}^{1/2}\boldsymbol{T}.$$

Since $q(-1) = q(1) = 0$, the boundary terms vanish when integrating by parts in each prior coordinate. Using $\zeta\, p(\mathcal{D}, \boldsymbol{\omega}) = \nabla_{\boldsymbol{\omega}} p(\mathcal{D}, \boldsymbol{\omega})$, we obtain

$$\begin{aligned}
\mathbb{E}[\boldsymbol{e}\zeta^\top] &= \iint \boldsymbol{e}(\nabla_{\boldsymbol{\omega}} p(\mathcal{D}, \boldsymbol{\omega}))^\top \, d\mu(\mathcal{D})\, d\boldsymbol{\omega} \\
&= -\iint \frac{\partial \boldsymbol{e}}{\partial \boldsymbol{\omega}} p(\mathcal{D}, \boldsymbol{\omega})\, d\mu(\mathcal{D})\, d\boldsymbol{\omega} \\
&= \boldsymbol{Q}^{1/2}\boldsymbol{T}.
\end{aligned}$$

The calculation first applies when $\boldsymbol{Q}^{1/2}\widehat{\boldsymbol{u}}$ is bounded. For an estimator with finite average risk under the prior, truncation and the finite second moment of ζ justify passage to the L_2 limit. An estimator with infinite average risk already satisfies the desired lower bound. Thus the identity, and consequently the preceding risk bound, applies to every estimator with finite average risk.

It remains to express the lower bound in terms of the variational objective. Since $\boldsymbol{T}\boldsymbol{T}^\top = \boldsymbol{A}$, the resolvent identities give

$$\begin{aligned}
\boldsymbol{T}(\boldsymbol{T}^\top \boldsymbol{M}\boldsymbol{T} + \pi^2 \boldsymbol{I}_r)^{-1}\boldsymbol{T}^\top &= \pi^{-2}\boldsymbol{T}(\boldsymbol{I}_r + \pi^{-2}\boldsymbol{T}^\top \boldsymbol{M}\boldsymbol{T})^{-1}\boldsymbol{T}^\top \\
&= \pi^{-2}\boldsymbol{P}(\boldsymbol{A}, \boldsymbol{M}/\pi^2).
\end{aligned}$$

Furthermore, $\boldsymbol{M}/\pi^2 \preceq \boldsymbol{M}$, so inversion reverses the PSD order in the definition of $\boldsymbol{P}$ and yields

$$\boldsymbol{P}(\boldsymbol{A}, \boldsymbol{M}/\pi^2) \succeq \boldsymbol{P}(\boldsymbol{A}, \boldsymbol{M}).$$

Combining these relations with the comparison between worst-case and average risk, we obtain, for every estimator and every nonzero feasible $\boldsymbol{A}$,

$$\begin{aligned}
\sup_{\|\boldsymbol{u}\|_2 \le R} \mathbb{E}_{\boldsymbol{u}} \|\boldsymbol{Q}^{1/2}(\widehat{\boldsymbol{u}}(\mathcal{D}) - \boldsymbol{u})\|_2^2 &\ge \mathbb{E}\|\boldsymbol{e}\|_2^2 \\
&\ge \pi^{-2}\mathrm{Tr}[\boldsymbol{Q}\boldsymbol{P}(\boldsymbol{A}, \boldsymbol{M}/\pi^2)] \\
&\ge \pi^{-2}\mathrm{Tr}[\boldsymbol{Q}\boldsymbol{P}(\boldsymbol{A}, \boldsymbol{M})].
\end{aligned}$$

Taking the infimum over estimators and then the supremum over feasible $\boldsymbol{A}$ proves

$$\mathrm{R}_*(n_1, \ldots, n_K) \ge \pi^{-2}\mathcal{L}(\boldsymbol{M}/\pi^2) \ge \pi^{-2}\mathcal{L}(\boldsymbol{M}).$$

The matrix $\boldsymbol{A} = 0$ contributes zero and does not change the supremum.

Combining this lower bound with (14) proves Theorem 1. Finally, Theorem 5 gives

$$\pi^{-2}\mathcal{L}(\boldsymbol{M}) \le \mathrm{R}_{*,\mathrm{seq}}(\boldsymbol{M}) \le \mathcal{L}(\boldsymbol{M}),$$

and therefore

$$\pi^{-2}\mathrm{R}_{*,\mathrm{seq}}(\boldsymbol{M}) \le \mathrm{R}_*(n_1, \ldots, n_K) \le \pi^2 v\, \mathrm{R}_{*,\mathrm{seq}}(\boldsymbol{M}).$$

Thus, when the concentration constants and E_0 are uniformly bounded, the unrestricted minimax risks of the regression and Gaussian sequence models have the same order.

A.3 Extension to infinite dimension

We work on $\mathcal{H} = (\ker \boldsymbol{S})^\perp$ and set $D = \mathrm{Dom}(\boldsymbol{S}^{-1/2}) = \mathrm{Ran}(\boldsymbol{S}^{1/2})$. This is dense in $\mathcal{H}$, and normalized parameters lie in $D \cap \{\boldsymbol{u} : \|\boldsymbol{u}\|_2 \le R\}$. All normalized covariance forms are understood as bounded operators. Let $\mathcal{A}$ be the positive trace-class operators of trace at most R^2.

For bounded $\boldsymbol{N} \succeq 0$ and a Hilbert–Schmidt operator W, define

$$F(W, \boldsymbol{A}) = \|(W\boldsymbol{N}^{1/2} - \boldsymbol{Q}^{1/2})\boldsymbol{A}^{1/2}\|_{\mathrm{HS}}^2 + \|W\|_{\mathrm{HS}}^2,$$
$$J_{\boldsymbol{N}}(W) = \sup_{\boldsymbol{A} \in \mathcal{A}} F(W, \boldsymbol{A}) = R^2 \|W\boldsymbol{N}^{1/2} - \boldsymbol{Q}^{1/2}\|_{\mathrm{op}}^2 + \|W\|_{\mathrm{HS}}^2.$$

The same completion of squares as in finite dimension gives

$$\min_W F(W, \boldsymbol{A}) = \mathrm{Tr}[\boldsymbol{Q}\boldsymbol{P}(\boldsymbol{A}, \boldsymbol{N})],$$
$$W_{\boldsymbol{A}} = \boldsymbol{Q}^{1/2}\boldsymbol{P}(\boldsymbol{A}, \boldsymbol{N})\boldsymbol{N}^{1/2}, \qquad \|W_{\boldsymbol{A}}\|_{\mathrm{HS}}^2 \le R^2 \|\boldsymbol{Q}\|_{\mathrm{op}}.$$

Thus all these minimizers lie in the weakly compact Hilbert–Schmidt ball $\mathcal{B}$ of radius $R\sqrt{\|\boldsymbol{Q}\|_{\mathrm{op}}}$. The function F is convex and weakly lower semicontinuous in W, and affine and trace-norm continuous in $\boldsymbol{A}$. Sion's minimax theorem (Sion, 1958) therefore yields

$$\min_W J_{\boldsymbol{N}}(W) = \min_{W \in \mathcal{B}} \sup_{\boldsymbol{A} \in \mathcal{A}} F(W, \boldsymbol{A}) = \sup_{\boldsymbol{A} \in \mathcal{A}} \min_{W \in \mathcal{B}} F(W, \boldsymbol{A}) = \mathcal{L}(\boldsymbol{N}).$$

The first equality follows from $J_{\boldsymbol{N}}(W) \ge \|W\|_{\mathrm{HS}}^2$ and $J_{\boldsymbol{N}}(0) = R^2\|\boldsymbol{Q}\|_{\mathrm{op}}$. Let W_* be a minimizer. Projecting its output onto $(\ker \boldsymbol{Q})^\perp$ and its input onto $(\ker \boldsymbol{N})^\perp$ cannot increase $J_{\boldsymbol{N}}$. Since $\boldsymbol{Q}^{1/2}D$ and $\boldsymbol{N}^{1/2}D$ are dense in these respective subspaces, there are finite-rank operators $L_m = \sum_j a_{mj} \otimes b_{mj}$, with $a_{mj}, b_{mj} \in D$, such that $W_m = \boldsymbol{Q}^{1/2} L_m \boldsymbol{N}^{1/2} \to W_*$ in Hilbert–Schmidt norm. Here $(a \otimes b)h = a\langle b, h\rangle$. Boundedness of $\boldsymbol{N}$ implies $J_{\boldsymbol{N}}(W_m) \to \mathcal{L}(\boldsymbol{N})$.

For the regression upper bound, set $v = 1 + \kappa E_0$ and $\boldsymbol{N} = \boldsymbol{M}/v$. For $b \in D$, the scalar statistic

$$z(b) = \sum_{i=1}^{K} \sum_{r=1}^{n_i} \sigma_{\varepsilon_i}^{-2} \langle \boldsymbol{x}_i^{(r)}, \boldsymbol{S}^{-1/2} b\rangle y_i^{(r)}$$

satisfies $\mathbb{E}_{\boldsymbol{u}} z(b) = \langle b, \boldsymbol{M}\boldsymbol{u}\rangle$ and $\mathrm{Var}_{\boldsymbol{u}} z(b) \le v\langle b, \boldsymbol{M}b\rangle$, by the scalar calculation proving (13). Consequently the well-defined finite-rank estimator $\widehat{\boldsymbol{\theta}}_m = \boldsymbol{S}^{-1/2} \sum_j a_{mj} z(b_{mj})/v$ satisfies, with $\boldsymbol{C}_{\mathrm{test}} = \sum_k \pi_k^* \boldsymbol{C}_k$,

$$\sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \mathbb{E}_{\boldsymbol{\theta}} \|\boldsymbol{C}_{\mathrm{test}}^{1/2}(\widehat{\boldsymbol{\theta}}_m - \boldsymbol{\theta})\|_2^2 \le R^2 \|\boldsymbol{Q}^{1/2}(L_m\boldsymbol{N} - \boldsymbol{I})\|_{\mathrm{op}}^2 + \|\boldsymbol{Q}^{1/2} L_m \boldsymbol{N}^{1/2}\|_{\mathrm{HS}}^2$$
$$= J_{\boldsymbol{N}}(W_m) \longrightarrow \mathcal{L}(\boldsymbol{M}/v).$$

Hence $\mathrm{R}_* \le \mathcal{L}(\boldsymbol{M}/v) \le v\mathcal{L}(\boldsymbol{M})$.

For the lower bound, choose finite-rank orthogonal projections Π_m onto the first m elements of an orthonormal basis contained in D. For any $\boldsymbol{A} \in \mathcal{A}$, the feasible operators $\boldsymbol{A}_m = \Pi_m \boldsymbol{A} \Pi_m = T_m T_m^*$ converge to $\boldsymbol{A}$ in trace norm; choose T_m with orthogonal columns in D. The preceding cosine-prior argument applies to $\boldsymbol{\theta} = \boldsymbol{S}^{-1/2} T_m \boldsymbol{\omega}$. For arbitrary estimators, apply its score calculation directly to $e = \boldsymbol{C}_{\mathrm{test}}^{1/2}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta})$: the target derivative is $G_m = \boldsymbol{C}_{\mathrm{test}}^{1/2} \boldsymbol{S}^{-1/2} T_m$, with $G_m^* G_m = T_m^* \boldsymbol{Q} T_m$. It therefore gives

$$\mathrm{R}_* \ge \pi^{-2} \mathrm{Tr}[\boldsymbol{Q}\boldsymbol{P}(\boldsymbol{A}_m, \boldsymbol{M}/\pi^2)].$$

The resolvent identity makes $\boldsymbol{P}(\boldsymbol{A}, \boldsymbol{N})$ trace-norm continuous in $\boldsymbol{A}$ for bounded $\boldsymbol{N}$. Letting $m \to \infty$ and taking the supremum over $\boldsymbol{A}$ yields $\mathrm{R}_* \ge \pi^{-2}\mathcal{L}(\boldsymbol{M}/\pi^2) \ge \pi^{-2}\mathcal{L}(\boldsymbol{M})$, completing the proof.

B Proof of Theorem 2

B.1 Invertibility of $\boldsymbol{L}$

Recall that $\boldsymbol{L} = \boldsymbol{D} - \boldsymbol{K}$, where $\boldsymbol{D} = \mathrm{diag}(n_1/\mu_1^2, \ldots, n_K/\mu_K^2)$ and $\boldsymbol{K}_{ij} = \mathrm{Tr}(\boldsymbol{C}_i \overline{\boldsymbol{G}} \boldsymbol{C}_j \overline{\boldsymbol{G}})$. Since the covariance matrices are positive semidefinite, $K_{ij} \ge 0$ and $\boldsymbol{K}_{ij} = \boldsymbol{K}_{ji}$. Using $\overline{\boldsymbol{G}}^{-1} =$

$\sum_{j=1}^{K} \mu_j \boldsymbol{C}_j + \lambda \boldsymbol{I}$ and the fixed-point equation $n_i/\mu_i = 1 + \mathrm{Tr}(\boldsymbol{C}_i \overline{\boldsymbol{G}})$, we obtain

$$\begin{aligned}\sum_{j=1}^{K} \mu_j \boldsymbol{K}_{ij} &= \mathrm{Tr}\left(\boldsymbol{C}_i \overline{\boldsymbol{G}} \left(\sum_{j=1}^{K} \mu_j \boldsymbol{C}_j\right) \overline{\boldsymbol{G}}\right) \\ &= \mathrm{Tr}(\boldsymbol{C}_i \overline{\boldsymbol{G}}) - \lambda \mathrm{Tr}(\boldsymbol{C}_i \overline{\boldsymbol{G}}^2) \\ &= \frac{n_i}{\mu_i} - 1 - \lambda \mathrm{Tr}(\boldsymbol{C}_i \overline{\boldsymbol{G}}^2).\end{aligned}$$

For any $\boldsymbol{v} \in \mathbb{R}^K$, the Cauchy-Schwartz inequality $2v_i v_j \le (\mu_j/\mu_i) v_i^2 + (\mu_i/\mu_j) v_j^2$, together with the nonnegativity and symmetry of $\boldsymbol{K}$, yields

$$\begin{aligned}\boldsymbol{v}^\top \boldsymbol{K} \boldsymbol{v} &\le \frac{1}{2} \sum_{i,j=1}^{K} \boldsymbol{K}_{ij} \left(\frac{\mu_j}{\mu_i} v_i^2 + \frac{\mu_i}{\mu_j} v_j^2\right) \\ &= \sum_{i=1}^{K} \frac{v_i^2}{\mu_i} \sum_{j=1}^{K} \mu_j \boldsymbol{K}_{ij} \\ &= \sum_{i=1}^{K} \left(\frac{n_i}{\mu_i^2} - \frac{1 + \lambda \mathrm{Tr}(\boldsymbol{C}_i \overline{\boldsymbol{G}}^2)}{\mu_i}\right) v_i^2.\end{aligned}$$

Consequently, since $\lambda > 0$, $\mathrm{Tr}(\boldsymbol{C}_i \overline{\boldsymbol{G}}^2) \ge 0$, and $\mu_i > 0$, every nonzero $\boldsymbol{v} \in \mathbb{R}^K$ satisfies

$$\begin{aligned}\boldsymbol{v}^\top \boldsymbol{L} \boldsymbol{v} = \boldsymbol{v}^\top (\boldsymbol{D} - \boldsymbol{K}) \boldsymbol{v} &= \sum_{i=1}^{K} \frac{n_i}{\mu_i^2} v_i^2 - \boldsymbol{v}^\top \boldsymbol{K} \boldsymbol{v} \\ &\ge \sum_{i=1}^{K} \frac{n_i}{\mu_i^2} v_i^2 - \sum_{i=1}^{K} \left(\frac{n_i}{\mu_i^2} - \frac{1 + \lambda \mathrm{Tr}(\boldsymbol{C}_i \overline{\boldsymbol{G}}^2)}{\mu_i}\right) v_i^2 \\ &= \sum_{i=1}^{K} \frac{1 + \lambda \mathrm{Tr}(\boldsymbol{C}_i \overline{\boldsymbol{G}}^2)}{\mu_i} v_i^2 > 0\end{aligned}$$

Thus $\boldsymbol{L}$ is positive definite and hence invertible.

B.2 Proof of Theorem 2

Proof. Conditionally on the designs,

$$\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_* = -\lambda \boldsymbol{G} \boldsymbol{\theta}_* + \boldsymbol{G} \sum_{i=1}^{K} \boldsymbol{X}_i^\top \varepsilon_i.$$

Averaging over the independent and centered training noises, we get

$$\mathsf{R}(\widehat{\boldsymbol{\theta}}) = \sum_{k=1}^{K} \pi_k^* B_k + \sum_{i=1}^{K} \sigma_{\varepsilon_i}^2 \mathrm{Tr}(\boldsymbol{C}_{\pi^*} \boldsymbol{G} \boldsymbol{X}_i^\top \boldsymbol{X}_i \boldsymbol{G}), \quad \text{where} \quad B_k = \lambda^2 \mathrm{Tr}(\boldsymbol{A}_* \boldsymbol{G} \boldsymbol{C}_k \boldsymbol{G}).$$

Using the second bound from Theorem 6 with $\boldsymbol{A} := \boldsymbol{A}_*$, we get for each k,

$$|B_k - \lambda^2 \boldsymbol{e}_k^\top \boldsymbol{D} \boldsymbol{L}^{-1} \tau_{\boldsymbol{A}_*}| \le C_{D,c_0} \lambda^2 \varepsilon_2 V_{\boldsymbol{A}_*} \quad \text{with probability at least } 1 - \sum_{i=1}^{K} n_i^{-D}.$$

Now we compute the deterministic equivalent $\boldsymbol{e}_k^\top \boldsymbol{D} \boldsymbol{L}^{-1} \tau_{\boldsymbol{A}_*}$. Recall that $\boldsymbol{K} := \boldsymbol{D} - \boldsymbol{L}$. We have

$$\boldsymbol{D} \boldsymbol{L}^{-1} = \boldsymbol{I} + \boldsymbol{K} \boldsymbol{L}^{-1}, \qquad \boldsymbol{K} \boldsymbol{e}_k = \tau_{\boldsymbol{C}_k}.$$

Therefore,

$$\begin{aligned}\boldsymbol{e}_k^\top \boldsymbol{D} \boldsymbol{L}^{-1} \tau_{\boldsymbol{A}_*} &= \tau_{\boldsymbol{A}_*}[k] + \tau_{\boldsymbol{C}_k}^\top \boldsymbol{L}^{-1} \tau_{\boldsymbol{A}_*} \\ &= \boldsymbol{\theta}_*^\top \overline{\boldsymbol{G}} \boldsymbol{C}_k \overline{\boldsymbol{G}} \boldsymbol{\theta}_* + \tau_{\boldsymbol{A}_*}^\top \boldsymbol{L}^{-1} \tau_{\boldsymbol{C}_k} \\ &= \frac{1}{\lambda^2} \overline{B}_k.\end{aligned}$$

Summing over $k \in [K]$, we have

$$\left| \sum_{k=1}^{K} \pi_k^* (B_k - \overline{B}_k) \right| \le C_{D,c_0} \lambda^2 \varepsilon_2 V_{\boldsymbol{A}_*}.$$

For the variance, using the third bound from Theorem 6 with $\boldsymbol{A} = \boldsymbol{C}_{\pi^*}$ and target dataset i,

$$\left| \frac{1}{n_i} \mathrm{Tr}(\boldsymbol{C}_{\pi^*} \boldsymbol{G} \boldsymbol{X}_i^\top \boldsymbol{X}_i \boldsymbol{G}) - \frac{1}{n_i} \boldsymbol{e}_i^\top \boldsymbol{L}^{-1} \tau_{\boldsymbol{C}_{\pi^*}} \right| \le C_{D,c_0} \varepsilon_{3,i} n_i^{-1} V_{\boldsymbol{C}_{\pi^*}}.$$

Multiplying by n_i and summing over $i \in [K]$ gives

$$\left| \sum_{i=1}^{K} \sigma_{\varepsilon_i}^2 (\mathrm{Tr}(\boldsymbol{C}_{\pi^*} \boldsymbol{G} \boldsymbol{X}_i^\top \boldsymbol{X}_i \boldsymbol{G}) - \boldsymbol{e}_i^\top \boldsymbol{L}^{-1} \tau_{\boldsymbol{C}_{\pi^*}}) \right| \le C_{D,c_0} V_{\boldsymbol{C}_{\pi^*}} \sum_{i=1}^{K} \sigma_{\varepsilon_i}^2 \varepsilon_{3,i}.$$

Direct calculation then gives

$$\begin{aligned} \sum_{i=1}^{K} \sigma_{\varepsilon_i}^2 \boldsymbol{e}_i^\top \boldsymbol{L}^{-1} \tau_{\boldsymbol{C}_*} &= \sum_{k=1}^{K} \pi_k^* \sum_{i=1}^{K} \sigma_{\varepsilon_i}^2 \tau_{\boldsymbol{C}_k}^\top \boldsymbol{L}^{-1} \boldsymbol{e}_i \\ &= \sum_{k=1}^{K} \pi_k^* \overline{V}_k. \end{aligned}$$

Taking union bounds then implies with probability at least $1 - \sum_{i=1}^{K} n_i^{-D}$,

$$|\mathsf{R}(\widehat{\boldsymbol{\theta}}) - \overline{\mathsf{R}}| \le \varepsilon_n, \qquad \text{where} \qquad \varepsilon_n = C_{D,c_0} \left(\lambda^2 \varepsilon_2 V_{\boldsymbol{\theta}_* \boldsymbol{\theta}_*^\top} + V_{\boldsymbol{C}_{\pi^*}} \sum_{i=1}^{K} \sigma_{\varepsilon_i}^2 \varepsilon_{3,i} \right).$$

□

C Proof of Theorem 3

The source constraints satisfy

$$\|\boldsymbol{C}_i^{1/2-r_i} \boldsymbol{\theta}\|_2^2 \asymp \sum_{j \ge 1} j^{\alpha_i(2r_i-1)} \boldsymbol{\theta}[j]^2, \qquad i \in \{1,2\}.$$

By the definition of s,

$$\max\{\alpha_1(2r_1-1), \alpha_2(2r_2-1)\} = 2s - \alpha_1.$$

The larger exponent dominates the smaller one for every $j \ge 1$. Consequently, if

$$\boldsymbol{\Theta}_0(\rho) = \left\{ \boldsymbol{\theta} : \sum_{j \ge 1} j^{2s-\alpha_1} \boldsymbol{\theta}[j]^2 \le \rho^2 \right\},$$

then there exist constants $c, C > 0$ such that $\boldsymbol{\Theta}_0(cR) \subseteq \boldsymbol{\Theta} \subseteq \boldsymbol{\Theta}_0(CR)$. We may therefore work with $\boldsymbol{\Theta}_0(R)$, since fixed changes in the radius only change the constants in the bounds below.

For this ellipsoid, choose the diagonal operator $\boldsymbol{S}$ with $[\boldsymbol{S}]_{j,j} = j^{2s-\alpha_1}$, so that the constraint is $\|\boldsymbol{S}^{1/2}\boldsymbol{\theta}\|_2 \le R$. The normalized covariances satisfy

$$[\boldsymbol{H}_1]_{j,j} = [\boldsymbol{S}^{-1/2} \boldsymbol{C}_1 \boldsymbol{S}^{-1/2}]_{j,j} = j^{-2s}, \qquad [\boldsymbol{H}_2]_{j,j} = [\boldsymbol{S}^{-1/2} \boldsymbol{C}_2 \boldsymbol{S}^{-1/2}]_{j,j} = j^{-(2s-\delta)}.$$

Since $s > 0$ and $2s - \delta > 0$, both normalized covariances are bounded. The fixed positive noise variances imply $E_0 = \max_{i=1,2} \frac{R^2 \|\boldsymbol{H}_i\|_{op}}{\sigma_{\varepsilon_i}^2} = O(1)$. Thus, Theorem 1 applies with constants independent of the number of coordinates.

Define $\boldsymbol{Q} = \boldsymbol{H}_1$ and $\boldsymbol{M} = n_1 \sigma_{\varepsilon_1}^{-2} \boldsymbol{H}_1 + n_2 \sigma_{\varepsilon_2}^{-2} \boldsymbol{H}_2$. Recall that

$$\mathcal{L}(\boldsymbol{M}) = \sup_{\substack{\boldsymbol{A} \succeq 0 \\ \mathrm{Tr}(\boldsymbol{A}) \le R^2}} F(\boldsymbol{A}), \qquad F(\boldsymbol{A}) := \mathrm{Tr}\left[\boldsymbol{Q} \boldsymbol{A}^{1/2} (\boldsymbol{I} + \boldsymbol{A}^{1/2} \boldsymbol{M} \boldsymbol{A}^{1/2})^{-1} \boldsymbol{A}^{1/2} \right].$$

Since $\boldsymbol{Q}, \boldsymbol{M}$ are both diagonal, we first show that the optimal $\boldsymbol{A}$ is diagonal. In particular, we show that for any feasible $\boldsymbol{A}$, $F(\mathrm{Diag}(\boldsymbol{A})) \geq F(\boldsymbol{A})$, where $\mathrm{Diag}(\boldsymbol{A})$ denote the diagonal operator whose j-th diagonal entry is $[\boldsymbol{A}]_{j,j}$.

Fix a coordinate j, and let $\mathbf{e}_j$ denote the corresponding standard basis vector. Because $\boldsymbol{M}$ is diagonal and positive semidefinite,

$$\boldsymbol{M} \succeq [\boldsymbol{M}]_{j,j}\mathbf{e}_j\mathbf{e}_j^\top.$$

Congruence by $\boldsymbol{A}^{1/2}$ preserves this order, whereas inversion reverses the order of positive definite operators. Consequently,

$$\begin{aligned}(\boldsymbol{I} + \boldsymbol{A}^{1/2}\boldsymbol{M}\boldsymbol{A}^{1/2})^{-1} &\preceq (\boldsymbol{I} + [\boldsymbol{M}]_{j,j}\boldsymbol{A}^{1/2}\mathbf{e}_j\mathbf{e}_j^\top\boldsymbol{A}^{1/2})^{-1} \\ &= \boldsymbol{I} - \frac{[\boldsymbol{M}]_{j,j}\boldsymbol{A}^{1/2}\mathbf{e}_j\mathbf{e}_j^\top\boldsymbol{A}^{1/2}}{1 + [\boldsymbol{M}]_{j,j}[\boldsymbol{A}]_{j,j}},\end{aligned}$$

where the equality follows from the Sherman–Morrison identity and $\mathbf{e}_j^\top\boldsymbol{A}\mathbf{e}_j = [\boldsymbol{A}]_{j,j}$. Multiplying on both sides by $\boldsymbol{A}^{1/2}$ and taking the j-th diagonal entry gives

$$\begin{aligned}\left[\boldsymbol{A}^{1/2}(\boldsymbol{I} + \boldsymbol{A}^{1/2}\boldsymbol{M}\boldsymbol{A}^{1/2})^{-1}\boldsymbol{A}^{1/2}\right]_{j,j} &\leq [\boldsymbol{A}]_{j,j} - \frac{[\boldsymbol{M}]_{j,j}[\boldsymbol{A}]_{j,j}^2}{1 + [\boldsymbol{M}]_{j,j}[\boldsymbol{A}]_{j,j}} \\ &= \frac{[\boldsymbol{A}]_{j,j}}{1 + [\boldsymbol{M}]_{j,j}[\boldsymbol{A}]_{j,j}}.\end{aligned}$$

Since $\boldsymbol{Q}$ is diagonal with nonnegative entries, summing the preceding inequality with weights $[\boldsymbol{Q}]_{j,j}$ yields

$$\begin{aligned}F(\boldsymbol{A}) &= \sum_{j\geq 1}[\boldsymbol{Q}]_{j,j}\left[\boldsymbol{A}^{1/2}(\boldsymbol{I} + \boldsymbol{A}^{1/2}\boldsymbol{M}\boldsymbol{A}^{1/2})^{-1}\boldsymbol{A}^{1/2}\right]_{j,j} \\ &\leq \sum_{j\geq 1}\frac{[\boldsymbol{Q}]_{j,j}[\boldsymbol{A}]_{j,j}}{1 + [\boldsymbol{M}]_{j,j}[\boldsymbol{A}]_{j,j}} = F(\mathrm{diag}(\boldsymbol{A})).\end{aligned}$$

The last equality follows by evaluating F on a diagonal operator. In infinitely many coordinates, the inequality follows by first summing over $j \leq d$ and then letting $d \to \infty$; all summands are nonnegative, so this passage to the limit is justified by monotone convergence.

Moreover, $\boldsymbol{A} \succeq 0$ implies $[\boldsymbol{A}]_{j,j} \geq 0$, and taking the diagonal part preserves the trace:

$$\mathrm{Diag}(\boldsymbol{A}) \succeq 0, \qquad \mathrm{Tr}(\mathrm{Diag}(\boldsymbol{A})) = \mathrm{Tr}(\boldsymbol{A}) \leq R^2.$$

Thus, every feasible $\boldsymbol{A}$ can be replaced by a feasible diagonal operator with an objective value at least as large. Since diagonal operators are also included in the original feasible set, restricting the supremum to diagonal $\boldsymbol{A}$ leaves its value unchanged. Writing their diagonal entries as a_j, we obtain

$$\mathcal{L}(\boldsymbol{M}) = \sup_{\substack{a_j\geq 0 \\ \sum_{j\geq 1}a_j \leq R^2}} \sum_{j\geq 1}\frac{[\boldsymbol{Q}]_{j,j}a_j}{1 + [\boldsymbol{M}]_{j,j}a_j}.$$

Set $b_j = [\boldsymbol{Q}]_{j,j}a_j$ and define

$$I_j = \frac{[\boldsymbol{M}]_{j,j}}{[\boldsymbol{Q}]_{j,j}} = \frac{n_1}{\sigma_{\varepsilon_1}^2} + \frac{n_2[\boldsymbol{C}_2]_{j,j}}{\sigma_{\varepsilon_2}^2[\boldsymbol{C}_1]_{j,j}} \asymp n_1 + n_2 j^\delta.$$

Since $[\boldsymbol{Q}]_{j,j} \asymp j^{-2s}$, Theorem 1 and the preceding ellipsoid comparison give

$$\mathsf{R}^*(n_1, n_2) \asymp \sup_{\substack{b_j\geq 0 \\ \sum_{j\geq 1}j^{2s}b_j \leq R^2}} \sum_{j\geq 1}\frac{b_j}{1 + I_j b_j}. \tag{15}$$

Next, we prove a upper and lower bound on $\mathsf{R}^*(n_1, n_2)$. For $n_1 + n_2 > 0$, define

$$T_m = R^2(m+1)^{-2s}, \qquad V_m = \sum_{j=1}^{m} I_j^{-1}, \qquad m \in \mathbb{N}_0,$$

where $V_0 = 0$. We show that for any $m \in \mathbb{N}_0$,

$$\mathsf{R}^*(n_1, n_2) \lesssim (V_m + T_m) \gtrsim \min\{V_m, T_m\}, \mathsf{R}^*(n_1, n_2) \lesssim (V_m + T_m),$$

and picking the optimal m on both side gives:

$$\sum_{j \geq 1} j^{2s} b_j \leq m^{2s} \min\{V_m, T_m\} \leq R^2.$$

For the upper bound, pick any feasible sequence in (15) and any $m \geq 0$, use $\frac{b_j}{1+I_j b_j} \leq I_j^{-1}$ for $j \leq m$, and $\frac{b_j}{1+I_j b_j} \leq b_j$ for $j \geq m$. Then, this gives

$$\sum_{j \geq 1} \frac{b_j}{1 + I_j b_j} \leq \sum_{j=1}^{m} I_j^{-1} + \sum_{j > m} b_j \leq V_m + T_m.$$

For the lower bound, we simply need to find a specific feasible b_j that satisfies the lower bound. Fix $m \geq 1$ and choose

$$b_j = \begin{cases} I_j^{-1} \min\{1, T_m / V_m\}, & 1 \leq j \leq m, \\ 0, & j > m. \end{cases}$$

This choice is feasible because

$$\sum_{j \geq 1} j^{2s} b_j \leq m^{2s} \min\{V_m, T_m\} \leq R^2.$$

Moreover, $I_j b_j \leq 1$, so

$$\sum_{j \geq 1} \frac{b_j}{1 + I_j b_j} \geq \frac{1}{2} \sum_{j=1}^{m} b_j = \frac{1}{2} \min\{V_m, T_m\}.$$

We have therefore proved

$$\sup_{m \geq 1} \min\{V_m, T_m\} \lesssim \mathsf{R}^*(n_1, n_2) \lesssim \inf_{m \geq 0} (V_m + T_m). \tag{16}$$

Next, we show that the upper and lower bounds are actually matching. Let m_* be the first positive integer for which $V_{m_*} \geq T_{m_*}$. Such an integer exists since T_m is decreasing to 0 in m, V_m is increasing in m, and $T_1 = \Theta_{n_1, n_2}(1) \geq V_1 = o_{n_1, n_2}(1)$. The lower bound in (16) is at least $\min\{V_{m_*}, T_{m_*}\} = T_{m_*}$, as by definition $V_{m_*} \geq T_{m_*}$. Now we show that in fact $\inf_{m \geq 0}(V_m + T_m) \lesssim T_{m_*}$. We have

$$\inf_{m \geq 0} (V_m + T_m) \leq V_{m_* - 1} + T_{m_* - 1} \leq 2R^2 m_*^{-2s} \leq 2^{2s+1} T_{m_*},$$

where we use the definition of m_* and the fact that $T_{m_*-1}/T_{m_*} \leq 2^{2s} = \Theta(1)$. Consequently,

$$\mathsf{R}^*(n_1, n_2) \asymp \inf_{m \in \mathbb{N}_0} \left\{ R^2 (m+1)^{-2s} + \sum_{j=1}^{m} \frac{1}{n_1 + n_2 j^\delta} \right\}. \tag{17}$$

We now evaluate (17). All cutoffs below are rounded to integers, which does not affect the rates. We use $n_2 \asymp n^{\gamma_2}$ throughout.

Target-only and auxiliary-only rates. For target data alone, $V_m \asymp m/n$. Taking $m \asymp n^{1/(1+2s)}$ gives

$$V_m \asymp T_m \asymp n^{-2s/(1+2s)} = n^{-\Gamma_{\mathrm{tar}}}.$$

For auxiliary data alone,

$$V_m \asymp n_2^{-1} \sum_{j=1}^{m} j^{-\delta} \asymp \begin{cases} n_2^{-1} m^{1-\delta}, & \delta < 1, \\ n_2^{-1} \log(m+1), & \delta = 1, \\ n_2^{-1}, & \delta > 1. \end{cases}$$

The corresponding balancing cutoffs are

$$m \asymp \begin{cases} n_2^{1/(1+2s-\delta)}, & \delta < 1, \\ (n_2/\log n_2)^{1/(2s)}, & \delta = 1, \\ n_2^{1/(2s)}, & \delta > 1. \end{cases}$$

Substituting these choices gives

$$\mathsf{R}^*(0, n_2) \asymp \begin{cases} n^{-2s\gamma_2/(1+2s-\delta)} = n^{-\Gamma_{\mathrm{aux}}}, & \delta < 1, \\ n^{-\gamma_2} \log n, & \delta = 1, \\ n^{-\gamma_2}, & \delta > 1. \end{cases}$$

Mixed data with $\gamma_2 \le \gamma_{\mathrm{c}}$. Take $m \asymp n^{1/(1+2s)}$. Since $\gamma_{\mathrm{c}} = 1 - \delta/(1+2s)$,

$$n_2 m^\delta \asymp n^{\gamma_2 + \delta/(1+2s)} \lesssim n.$$

Every summand of V_m is at most a constant times $1/n$. On the block $m/2 \le j \le m$, we also have $j^\delta \asymp m^\delta$ and hence $I_j \asymp n$. Therefore

$$V_m \asymp \frac{m}{n} \asymp T_m \asymp n^{-\Gamma_{\mathrm{tar}}}.$$

This argument holds for every fixed real δ, including $\delta \le 0$, and includes the boundary $\gamma_2 = \gamma_{\mathrm{c}}$.

Mixed data with $\delta < 1$ and $\gamma_2 > \gamma_{\mathrm{c}}$. Take $m \asymp n_2^{1/(1+2s-\delta)}$. Then

$$\frac{n_2 m^\delta}{n} \asymp n^{\gamma_2(1+2s)/(1+2s-\delta)-1} \longrightarrow \infty.$$

Since $I_j \gtrsim n_2 j^\delta$, summing over all $j \le m$ gives $V_m \lesssim n_2^{-1} m^{1-\delta}$. On $m/2 \le j \le m$, we have $I_j \asymp n_2 m^\delta$, which gives the reverse bound. Consequently,

$$V_m \asymp n_2^{-1} m^{1-\delta} \asymp T_m \asymp n^{-2s\gamma_2/(1+2s-\delta)} = n^{-\Gamma_{\mathrm{aux}}}.$$

Mixed data with $\delta = 1$ and $\gamma_2 > \gamma_{\mathrm{c}}$. Take $m \asymp (n_2/\log n)^{1/(2s)}$, so $T_m \asymp n_2^{-1} \log n$. If $\gamma_2 \ge 1$, then $I_j \asymp n_2 j$ for every $j \ge 1$, and

$$V_m \asymp n_2^{-1} \log(m+1) \asymp n_2^{-1} \log n \asymp T_m.$$

If $\gamma_{\mathrm{c}} < \gamma_2 < 1$, define $j_{\mathrm{c}} = n/n_2$. Then

$$\frac{m}{j_{\mathrm{c}}} \asymp \frac{n^{\gamma_2/(2s)-(1-\gamma_2)}}{(\log n)^{1/(2s)}}, \qquad \frac{\gamma_2}{2s} - (1-\gamma_2) = \frac{1+2s}{2s}(\gamma_2 - \gamma_{\mathrm{c}}) > 0.$$

Thus $m/j_{\mathrm{c}} \to \infty$ and $\log(m/j_{\mathrm{c}}) \asymp \log n$. Splitting the sum at j_{c} gives

$$V_m \asymp \frac{j_{\mathrm{c}}}{n} + \frac{1}{n_2} \sum_{j_{\mathrm{c}} < j \le m} \frac{1}{j} \asymp \frac{1 + \log(m/j_{\mathrm{c}})}{n_2} \asymp n_2^{-1} \log n \asymp T_m.$$

Hence the risk is $\asymp n^{-\gamma_2} \log n$. At $\gamma_2 = \gamma_{\mathrm{c}}$, the preceding target-data cutoff applies and gives $n^{-\Gamma_{\mathrm{tar}}}$ without a logarithmic factor.

Mixed data with $\delta > 1$ and $\gamma_{\mathrm{c}} < \gamma_2 < 1$. Define $j_{\mathrm{c}} = (n/n_2)^{1/\delta}$ and take $m \asymp n^{(\delta-1+\gamma_2)/(2s\delta)}$. The exponent of m/j_{c} is

$$\frac{\delta - 1 + \gamma_2}{2s\delta} - \frac{1-\gamma_2}{\delta} = \frac{(1+2s)(\gamma_2 - \gamma_{\mathrm{c}})}{2s\delta} > 0.$$

Thus $m > 2j_{\mathrm{c}}$ for sufficiently large n. The terms with $j \le j_{\mathrm{c}}$ contribute order j_{c}/n. Since $\delta > 1$, the remaining terms satisfy

$$\sum_{j_{\mathrm{c}} < j \le m} I_j^{-1} \lesssim n_2^{-1} \sum_{j > j_{\mathrm{c}}} j^{-\delta} \lesssim n_2^{-1} j_{\mathrm{c}}^{1-\delta} = \frac{j_{\mathrm{c}}}{n}.$$

It follows that

$$V_m \asymp \frac{j_{\mathrm{c}}}{n} \asymp n^{-(\delta-1+\gamma_2)/\delta} \asymp T_m,$$

which proves the intermediate mixed-data rate.

Mixed data with $\delta > 1$ and $\gamma_2 \geq 1$. In this case, $I_j \asymp n_2 j^\delta$ for every $j \geq 1$. Since $\sum_{j\geq 1} j^{-\delta} < \infty$, taking $m \asymp n_2^{1/(2s)}$ gives

$$V_m \asymp n_2^{-1} \asymp T_m \asymp n^{-\gamma_2}.$$

The preceding cases cover all parameter regimes and establish the claimed rates.

D Proof of Theorem 4

For the fixed power-law signal $\boldsymbol{\theta}_*[j] = j^{-\beta}$, recall that

$$s = \frac{\alpha_1 + 2\beta - 1}{2}, \qquad j^{-\alpha_1}\boldsymbol{\theta}_*[j]^2 = j^{-1-2s}.$$

In particular, $2s - \delta = \alpha_2 + 2\beta - 1 > 0$. Recall $a = s \wedge \alpha_1$ and $b = s \wedge \alpha_2$. Write $c_{ij} = j^{-\alpha_i}$, and

$$h_j = nc_{1j} + n_2 c_{2j}, \qquad d_j(\lambda) = \mu_1 c_{1j} + \mu_2 c_{2j} + \lambda.$$

The fixed-point equations and deterministic resolvent are

$$\frac{n_i}{\mu_i} = 1 + \sum_{j\geq 1} \frac{c_{ij}}{d_j(\lambda)}, \qquad \overline{\boldsymbol{G}} = \mathrm{Diag}\big(d_j(\lambda)^{-1}\big)_{j\geq 1}. \tag{18}$$

Let $\boldsymbol{\mu} = (\mu_1, \mu_2)^\top$, $\boldsymbol{D} = \mathrm{Diag}(n_i/\mu_i^2)_{i=1}^2$, and

$$\boldsymbol{K}_{i\ell} = \sum_{j\geq 1} \frac{c_{ij}c_{\ell j}}{d_j(\lambda)^2}, \qquad \boldsymbol{L} = \boldsymbol{D} - \boldsymbol{K}.$$

Recall that the deterministic equivalence of the risk is given by

$$\overline{\mathsf{R}}_1 = \lambda^2 \left[\langle \boldsymbol{\theta}_*, \overline{\boldsymbol{G}}\boldsymbol{C}_1\overline{\boldsymbol{G}}\boldsymbol{\theta}_*\rangle + \tau_{\boldsymbol{\theta}_*}^\top \boldsymbol{L}^{-1}\tau_{\boldsymbol{C}_1}\right] + \sum_{i=1}^2 \sigma_{\varepsilon_i}^2\, \tau_{\boldsymbol{C}_1}^\top \boldsymbol{L}^{-1}\boldsymbol{e}_i.$$

General proof recipe. Define

$$B_0(\lambda) = \sum_{j\geq 1} c_{1j}\boldsymbol{\theta}_*[j]^2 \left(\frac{\lambda}{h_j + \lambda}\right)^2, \qquad V_0(\lambda) = \sum_{j\geq 1} \frac{c_{1j}h_j}{(h_j+\lambda)^2}. \tag{19}$$

Pick a positive sequence λ_n such that

$$\epsilon_n := \max_{i\in\{1,2\}} \sum_{j\geq 1} \frac{c_{ij}}{h_j + \lambda_n} = o(1).$$

We first show that

$$\min\{B_0(\lambda_n), V_0(\lambda_n)\} \lesssim \inf_{\lambda>0} \overline{\mathsf{R}}_1(\lambda) \lesssim B_0(\lambda_n) + V_0(\lambda_n) \asymp \max\{B_0(\lambda_n), V_0(\lambda_n)\}. \tag{20}$$

Then, for each regime, we construct an explicit feasible λ_n, such that $B_0(\lambda_n) \asymp V_0(\lambda_n)$ and $\varepsilon_n = o(1)$. This gives the tight scaling law of the optimal ridge risk.

Control $\boldsymbol{L}^{-1}$ terms. We first control the order of terms that depend on $\boldsymbol{L}^{-1}$ to mitigate the matrix inversion problem.

First, we have the following identities from direct computations:

$$[\boldsymbol{D}\boldsymbol{\mu}]_i = \frac{n_i}{\mu_i}, \quad [\boldsymbol{K}\boldsymbol{\mu}]_i = \sum_{j\geq 1} \frac{c_{ij}\sum_\ell \mu_\ell c_{\ell j}}{d_j(\lambda)^2} = \sum_{j\geq 1} \frac{c_{ij}(d_j(\lambda) - \lambda))}{d_j(\lambda)^2} = \frac{n_i}{\mu_i} - 1 - \lambda\tau_{\boldsymbol{I}}[i],$$

$$\tau_{\boldsymbol{I}}[i] = \sum_{j\geq 1} \frac{c_{ij}}{d_j(\lambda)^2}.$$

Together with $\boldsymbol{L} = \boldsymbol{D} - \boldsymbol{K}$ we have $\boldsymbol{L}\boldsymbol{\mu} = \mathbf{1} + \lambda\tau_{\boldsymbol{I}}$ and

$$[\boldsymbol{D}^{-1}\boldsymbol{K}\boldsymbol{\mu}]_i = \mu_i - \frac{\mu_i^2}{n_i}(1 + \lambda\tau_{\boldsymbol{I}}[i]) < \mu_i,$$

which means $\boldsymbol{D}^{-1}\boldsymbol{K}\boldsymbol{\mu} < \boldsymbol{\mu}$. Note that $\boldsymbol{D}^{-1}\boldsymbol{K}$ and $\boldsymbol{\mu}$ are both entrywise strictly positive, implying that the spectral radius of $\boldsymbol{D}^{-1}\boldsymbol{K}$ is strictly smaller than 1. To see this, denote $\boldsymbol{A} = \boldsymbol{D}^{-1}\boldsymbol{K}$ temporarily and let $(\kappa, \boldsymbol{v})$ be an arbitrary eigenvalue-eigenvector pair. Define $t = \max_{i=1,2} \frac{|v_i|}{\mu_i}$. Clearly, for $i = 1, 2$, $|v_i| \le t\mu_i$, and assume w.l.o.g. $|v_1| = t\mu_1$. Now we have:

$$|\kappa||v_1| = |[\boldsymbol{A}\boldsymbol{v}]_1| \le \sum_{\ell} \boldsymbol{A}_{1\ell}|\boldsymbol{v}_\ell| < t[\boldsymbol{A}\boldsymbol{\mu}]_1 \le t\mu_1 = |v_1|,$$

which implies that the absolute value of all the eigenvalues is strictly smaller than 1. Thus, the matrix $\boldsymbol{I} - \boldsymbol{A}$ is invertible, and the inverse can be expanded into the following Taylor series $(\boldsymbol{I} - \boldsymbol{A})^{-1} = \sum_{\ell \ge 0}(\boldsymbol{A})^\ell$, whose convergence is guaranteed by the contraction of $\boldsymbol{A}$. Now note that by definition $\boldsymbol{L} = \boldsymbol{D}(\boldsymbol{I} - \boldsymbol{D}^{-1}\boldsymbol{K})$. Consequently,

$$\boldsymbol{L}^{-1} = \sum_{k \ge 0}(\boldsymbol{D}^{-1}\boldsymbol{T})^k\boldsymbol{D}^{-1}, \qquad (\boldsymbol{L}^{-1})_{i\ell} \ge (\boldsymbol{D}^{-1})_{i\ell}, \quad \forall i, \ell \in \{1, 2\}.$$

This means that $\boldsymbol{L}^{-1}$ is entrywise non-negative and further, by the fact that $\boldsymbol{L}\boldsymbol{\mu} = \mathbf{1} + \lambda\tau_{\boldsymbol{I}}$ and the fact that $\boldsymbol{L}$ is invertible, we have $\boldsymbol{L}^{-1}\mathbf{1} \le \boldsymbol{\mu}$ entrywise.

Now recall that the variance $\overline{\mathsf{V}}_1 = \sum_{i=1}^2 \sigma_{\varepsilon_i}^2 \tau_{\boldsymbol{C}_1}^\top \boldsymbol{L}^{-1}\boldsymbol{e}_i = \tau_{\boldsymbol{C}_1}^\top \boldsymbol{L}^{-1}\boldsymbol{\sigma}$, where we define $\boldsymbol{\sigma} = [\sigma_{\varepsilon_1}^2, \sigma_{\varepsilon_2}^2]^\top$. Using that $\boldsymbol{L}^{-1} \ge \boldsymbol{D}^{-1}$, $\boldsymbol{L}^{-1}\mathbf{1} \le \mu$ entrywise and $\sigma_{\varepsilon_1}^2, \sigma_{\varepsilon_2}^2$ are constants, we have

$$C_{*,-}\sum_{i=1}^2 \frac{\mu_i^2}{n_i}\tau_{\boldsymbol{C}_1}[i] \le \overline{\mathsf{V}}_1(\lambda) \le C_{*,+}\sum_{i=1}^2 \mu_i\tau_{\boldsymbol{C}_1}[i]. \tag{21}$$

Next, consider the term $\lambda^2\tau_{\boldsymbol{\theta}_*}^\top \boldsymbol{L}^{-1}\tau_{\boldsymbol{C}_1}$. We have

$$0 \le \lambda^2\tau_{\boldsymbol{\theta}_*\boldsymbol{\theta}_*^\top}[i] = \sum_{j \ge 1} c_{ij}\boldsymbol{\theta}_*[j]^2\left(\frac{\lambda}{d_j(\lambda)}\right)^2 \le \sum_{j \ge 1} c_{ij}\boldsymbol{\theta}_*[j]^2 = \sum_{j \ge 1} j^{-\alpha_i - 2\beta} < \infty.$$

By the entrywise nonnegativity of $\boldsymbol{L}^{-1}$, this term is bounded above by a constant times $\overline{\mathsf{V}}_1$. Thus, we have

$$\overline{\mathsf{R}}_1 \asymp \lambda^2\left\langle \boldsymbol{\theta}_*, \overline{\boldsymbol{G}}\boldsymbol{C}_1\overline{\boldsymbol{G}}\boldsymbol{\theta}_*\right\rangle + \overline{\mathsf{V}}_1$$

uniformly over $\lambda > 0$.

Reduction of the optimal ridge risk to two spectral sums. Next, we study the risk under the optimal ridge regularization, and we aim to show (20).

Denote $\tilde{\mathsf{B}}(\lambda) = \lambda^2\left\langle \boldsymbol{\theta}_*, \overline{\boldsymbol{G}}\boldsymbol{C}_1\overline{\boldsymbol{G}}\boldsymbol{\theta}_*\right\rangle$ and we show that under the above choice of λ_n,

$$\overline{\mathsf{V}}_1(\lambda_n) \asymp V_0(\lambda_n), \quad \tilde{\mathsf{B}}(\lambda_n) \asymp B_0(\lambda_n).$$

To see this, it is sufficient to show $\mu_i(\lambda_n)/n_i = \Theta(1), i = 1, 2$. Let $\eta = \min_i \mu_i(\lambda_n)/n_i \in (0, 1]$. By the self-consistent equation and recalling $d_j(\lambda_n) = \mu_1 c_{1j} + \mu_2 c_{2j} + \lambda_n$, $h_j = n_1\mu_1 + n_2\mu_2 c_{2j}$, we have $d_j(\lambda_n) \ge \eta(h_j + \lambda_n)$, and (18) gives

$$\frac{\mu_i(\lambda_n)}{n_i} \ge \frac{1}{1 + \epsilon_n/\eta_n},$$

which is equivalent to $\eta_n \ge \frac{\eta_n}{\eta_n + \epsilon_n}$ and hence $1 - \epsilon_n \le \eta_n \le 1$. This always implies $d_j(\lambda_n) \asymp h_j + \lambda_n$ uniformly in j as we have the upper bound $d_j(\lambda_n) \le h_j + \lambda$ due to $\mu_i \le n_i$ by definition. This gives that $\overline{\mathsf{V}}_1 \asymp V_0(\lambda_n)$, and we obtain the upper bound

$$\inf_{\lambda > 0} \overline{\mathsf{R}}_1(\lambda) \le \mathsf{R}_1(\lambda_n) \asymp B_0(\lambda_n) + V_0(\lambda_n).$$

Next we prove the lower bound on $\inf_{\lambda>0} \overline{\mathsf{R}}_1(\lambda)$. For simplicity denote $\tilde{\mathsf{B}}(\lambda) := \lambda^2 \langle \boldsymbol{\theta}_*, \overline{\boldsymbol{G}} \boldsymbol{C}_1 \overline{\boldsymbol{G}} \boldsymbol{\theta}_* \rangle$. Fix any feasible λ_n, we aim to show that for all $\lambda \geq \lambda_n$, $\overline{\mathsf{R}}_1(\lambda) \geq \tilde{\mathsf{B}}(\lambda) \geq B_0(\lambda_n)$; and for all $0 < \lambda \leq \lambda_n$, $\mathsf{R}_1(\lambda) \geq \overline{\mathsf{V}}_1(\lambda) \geq V_0(\lambda_n)$, which implies the lower bound.

When $\lambda > \lambda_n$, it is sufficient to show that $\tilde{\mathsf{B}}(\lambda) > \tilde{\mathsf{B}}(\lambda_n)$. We prove the identity via the monotonicity of $\tilde{\mathsf{B}}(\lambda)$ w.r.t λ. In particular, define $q_j(\lambda) = \sum_{i=1}^2 \frac{c_{ij}\mu_i(\lambda)}{\lambda}$, and we have:

$$\tilde{\mathsf{B}}(\lambda) = \sum_j c_{1j} \boldsymbol{\theta}_*[j]^2 \frac{1}{(1+q_j(\lambda))^2}.$$

It is sufficient to show that $q_j(\lambda)$ is monotone decreasing in λ. Hence, we compute the derivative of $\frac{\boldsymbol{\mu}(\lambda)}{\lambda}$, and have:

$$\frac{\mathrm{d}}{\mathrm{d}\lambda} \frac{\boldsymbol{\mu}(\lambda)}{\lambda} = \frac{\lambda \boldsymbol{\mu}' - \boldsymbol{\mu}}{\lambda^2} = \frac{\lambda \boldsymbol{L}^{-1} \tau_{\boldsymbol{I}} - \boldsymbol{L}^{-1}(1 + \lambda \tau_{\boldsymbol{I}})}{\lambda^2} = -\frac{1}{\lambda^2} \boldsymbol{L}^{-1} \mathbf{1} < 0,$$

where here we use the derivative computation:

$$\boldsymbol{L}\boldsymbol{\mu}' = \tau_{\boldsymbol{I}}, \quad \boldsymbol{L}\boldsymbol{\mu} = \mathbf{1} + \lambda \pi_{\boldsymbol{I}}.$$

To obtain the identity $\boldsymbol{L}\boldsymbol{\mu}' = \tau_{\boldsymbol{I}}$, note that taking the derivative of the self-consistency equation gives

$$-\frac{n_i}{\mu_i^2}\mu_i' = -\sum_j \frac{c_{ij}}{d_j^2(\lambda)} d_j'(\lambda) = -\sum_j \frac{c_{ij}}{d_j^2(\lambda)} (1 + \sum_{\ell=1}^2 \mu_\ell' c_{\ell j}) = -\tau_{\boldsymbol{I}}[i] - [\boldsymbol{K}\boldsymbol{\mu}']_i.$$

Noting that $\frac{n_i}{\mu_i^2}\mu_i' = [\boldsymbol{D}\boldsymbol{\mu}']_i$ and $\boldsymbol{L} = \boldsymbol{D} - \boldsymbol{K}$ finishes the proof.

When $0 < \lambda < \lambda_n$, we use the monotonicity of a lower bound on $\overline{\mathsf{V}}_1(\lambda)$. By Cauchy–Schwarz, for every $j \geq 1$,

$$\begin{aligned}
\left(\sum_{i=1}^2 \mu_i c_{ij} \right)^2 &= \left(\sum_{i=1}^2 \sqrt{n_i c_{ij}}\, \mu_i \sqrt{\frac{c_{ij}}{n_i}} \right)^2 \\
&\leq \left(\sum_{i=1}^2 n_i c_{ij} \right) \left(\sum_{i=1}^2 \frac{\mu_i^2}{n_i} c_{ij} \right) = h_j \sum_{i=1}^2 \frac{\mu_i^2}{n_i} c_{ij}.
\end{aligned}$$

Using (21), we obtain

$$\begin{aligned}
\overline{\mathsf{V}}_1(\lambda) &\geq C_{*,-} \sum_{i=1}^2 \frac{\mu_i^2}{n_i} \tau_{\boldsymbol{C}_1}[i] \\
&= C_{*,-} \sum_{j \geq 1} \frac{c_{1j}}{d_j(\lambda)^2} \sum_{i=1}^2 \frac{\mu_i^2}{n_i} c_{ij} \\
&\geq C_{*,-} \sum_{j \geq 1} \frac{c_{1j}}{h_j} \left(\frac{\sum_{i=1}^2 \mu_i c_{ij}}{d_j(\lambda)} \right)^2 \\
&= C_{*,-} \sum_{j \geq 1} \frac{c_{1j}}{h_j} \left(\frac{q_j(\lambda)}{1 + q_j(\lambda)} \right)^2.
\end{aligned}$$

Since $q_j(\lambda)$ is decreasing in λ and $q \mapsto q/(1+q)$ is increasing for $q \geq 0$, we have

$$\frac{q_j(\lambda)}{1 + q_j(\lambda)} \geq \frac{q_j(\lambda_n)}{1 + q_j(\lambda_n)}, \qquad 0 < \lambda \leq \lambda_n.$$

Moreover, the previously established comparisons $\mu_i(\lambda_n) \asymp n_i$ and $d_j(\lambda_n) \asymp h_j + \lambda_n$ imply

$$\frac{q_j(\lambda_n)}{1 + q_j(\lambda_n)} = \frac{\sum_{i=1}^2 \mu_i(\lambda_n) c_{ij}}{d_j(\lambda_n)} \asymp \frac{h_j}{h_j + \lambda_n}$$

uniformly in j. Consequently, for all $0 < \lambda \le \lambda_n$,

$$\begin{aligned}\overline{\mathsf{R}}_1(\lambda) &\ge \overline{\mathsf{V}}_1(\lambda) \\ &\ge C_{*,-} \sum_{j\ge 1} \frac{c_{1j}}{h_j} \left(\frac{q_j(\lambda_n)}{1+q_j(\lambda_n)}\right)^2 \\ &\gtrsim \sum_{j\ge 1} \frac{c_{1j} h_j}{(h_j+\lambda_n)^2} = V_0(\lambda_n).\end{aligned}$$

Combining this bound with the lower bound for $\lambda \ge \lambda_n$ gives

$$\inf_{\lambda>0} \overline{\mathsf{R}}_1(\lambda) \gtrsim \min\{B_0(\lambda_n), V_0(\lambda_n)\}.$$

Together with the upper bound proved above, this establishes (20).

Evaluation of the spectral sums. We next evaluate B_0, V_0, and ϵ_n under the power-law model. Recall that:

$$B_0(\lambda) = \sum_{j\ge 1} c_{1j} \boldsymbol{\theta}_*[j]^2 \left(\frac{\lambda}{h_j+\lambda}\right)^2, \qquad V_0(\lambda) = \sum_{j\ge 1} \frac{c_{1j} h_j}{(h_j+\lambda)^2}.$$

For the fixed power-law signal $\boldsymbol{\theta}_*[j] = j^{-\beta}$, recall that

$$s = \frac{\alpha_1 + 2\beta - 1}{2}, \qquad j^{-\alpha_1} \boldsymbol{\theta}_*[j]^2 = j^{-1-2s}.$$

In particular, $2s - \delta = \alpha_2 + 2\beta - 1 > 0$. Recall $a = s \wedge \alpha_1$ and $b = s \wedge \alpha_2$. Define $c_{ij} = j^{-\alpha_i}$, and

$$h_j = n c_{1j} + n_2 c_{2j}, \qquad d_j(\lambda) = \mu_1 c_{1j} + \mu_2 c_{2j} + \lambda.$$

For $m \ge 1$, define

$$h_m = n m^{-\alpha_1} + n_2 m^{-\alpha_2}, \qquad \lambda(m) = h_m.$$

Thus, choosing a cutoff m specifies the regularization $\lambda_n = \lambda(m)$. All comparison constants below are independent of n and m, with the other model parameters held fixed.

Since h_j is decreasing in j, splitting the bias at $m \ge 1$ gives

$$\begin{aligned}B_0(h_m) &\asymp h_m^2 \sum_{j\le m} \frac{j^{-1-2s}}{h_j^2} + \sum_{j>m} j^{-1-2s} \\ &\asymp h_m^2 \sum_{j\le m} \frac{j^{-1-2s}}{h_j^2} + m^{-2s}.\end{aligned} \tag{22}$$

Similarly,

$$\begin{aligned}V_0(h_m) &\asymp \sum_{j\le m} \frac{c_{1j}}{h_j} + h_m^{-2} \sum_{j>m} c_{1j} h_j, \\ h_m^{-2} \sum_{j>m} c_{1j} h_j &\asymp \frac{n m^{1-2\alpha_1} + n_2 m^{1-\alpha_1-\alpha_2}}{h_m^2} = \frac{m}{n + n_2 m^{\delta}}.\end{aligned}$$

Note that $\sum_{j=m/2}^{m} \frac{c_{1j}}{h_j} = \sum_{j=m/2}^{m} \frac{j^{-\alpha_1}}{n j^{-\alpha_1} + n_2 j^{-\alpha_2}} \asymp \frac{m}{n+n_2 m^{\delta}}$.

Consequently,

$$V_0(h_m) \asymp \sum_{j\le m} \frac{1}{n + n_2 j^{\delta}}. \tag{23}$$

For the feasibility condition, define

$$\epsilon(m) := \max_{i\in\{1,2\}} \sum_{j\ge 1} \frac{c_{ij}}{h_j + h_m}.$$

Let $\alpha_{\min} = \min\{\alpha_1, \alpha_2\}$ and $\delta_+ = \max\{\delta, 0\}$. Since $\max_i c_{ij} = j^{-\alpha_{\min}}$ for every j, we have

$$\begin{aligned}\epsilon(m) &= \sum_{j\geq 1} \frac{j^{-\alpha_{\min}}}{h_j + h_m} \\ &\asymp \sum_{j\leq m} \frac{j^{\delta_+}}{n + n_2 j^{\delta}} + \frac{m^{1-\alpha_{\min}}}{h_m} \\ &\asymp \frac{m^{1+\delta_+}}{n + n_2 m^{\delta}}.\end{aligned} \tag{24}$$

For the last comparison, note that $\frac{m^{1-\alpha_{\min}}}{h_m} \asymp \frac{m^{1+\delta_+}}{n+n_2 m^{\delta}}$, and $\frac{j^{\delta_+}}{n+n_2 j^{\delta}}$ is non-decreasing in j which implies $\sum_{j\leq m} \frac{j^{\delta_+}}{n+n_2 j^{\delta}} \leq \frac{m^{1+\delta_+}}{n+n_2 m^{\delta}}$.

We will retain the logarithmic factors arising at the critical source exponents. Write

$$\ell(t) = 1 + \log t, \qquad \ell_i(t) = 1 + \mathbf{1}_{\{s=\alpha_i\}} \log t, \qquad t \geq 1.$$

Next, we discuss how different choices of m give different scalings in each regime.

Spectral sums when $\delta > 0$. In this case $\alpha_1 > \alpha_2$ and $a \geq b$. Define

$$J = \max\left\{1, \left(\frac{n}{n_2}\right)^{1/\delta}\right\}.$$

Recall that

$$h_j = n j^{-\alpha_1} + n_2 j^{-\alpha_2}, \qquad \ell(t) = 1 + \log t, \qquad \ell_i(t) = 1 + \mathbf{1}_{\{s=\alpha_i\}} \log t.$$

We will repeatedly use the following consequence of Lemma 2: for $i \in \{1, 2\}$ and $x \geq 1$,

$$\sum_{j\leq x} j^{2\alpha_i - 1 - 2s} \asymp \begin{cases} x^{2\alpha_i - 2s}, & s < \alpha_i, \\ \ell(x), & s = \alpha_i, \\ 1, & s > \alpha_i. \end{cases}$$

When $n_2 \leq n$, the target dataset dominates h_j for $j \leq J$, whereas the auxiliary dataset dominates for $j \geq J$. When $n_2 > n$, we have $J = 1$ and $h_j \asymp n_2 j^{-\alpha_2}$ for every $j \geq 1$.

For $1 \leq m \leq J$ with $n_2 \leq n$, $h_m \asymp n m^{-\alpha_1}$, and (22) gives

$$B_0(h_m) \asymp m^{-2s} + m^{-2\alpha_1} \sum_{j\leq m} j^{2\alpha_1 - 1 - 2s}.$$

The second term has order

$$m^{-2\alpha_1} \sum_{j\leq m} j^{2\alpha_1 - 1 - 2s} \asymp \begin{cases} m^{-2s}, & s < \alpha_1, \\ m^{-2\alpha_1}\ell(m), & s = \alpha_1, \\ m^{-2\alpha_1}, & s > \alpha_1. \end{cases}$$

In each case, m^{-2s} is bounded above by a constant times this contribution. Consequently,

$$B_0(h_m) \asymp m^{-2a} \ell_1(m).$$

For $m \geq J$, splitting the first sum in (22) at J gives

$$B_0(h_m) \asymp m^{-2s} + m^{-2\alpha_2} \left[\underbrace{J^{-2\delta} \sum_{j\leq J} j^{2\alpha_1 - 1 - 2s}}_{=:S_+(J)} + \underbrace{\sum_{J<j\leq m} j^{2\alpha_2 - 1 - 2s}}_{=:T_+(m,J)} \right].$$

When $n_2 \le n$, the factor $J^{-2\delta}$ follows from $(n_2/n)^2 = J^{-2\delta}$. When $n_2 > n$, we have $J = 1$, and the same display follows by separating the term $j = 1$ and using $h_j \asymp n_2 j^{-\alpha_2}$ throughout.

Applying Lemma 2 to the two sums separately, we obtain

$$S_+(J) \asymp \begin{cases} J^{2\alpha_2 - 2s}, & s < \alpha_1, \\ J^{-2\delta}\ell(J), & s = \alpha_1, \\ J^{-2\delta}, & s > \alpha_1, \end{cases}$$

and

$$T_+(m, J) \asymp \begin{cases} m^{2\alpha_2 - 2s} - J^{2\alpha_2 - 2s}, & s < \alpha_2, \\ \log(m/J), & s = \alpha_2, \\ J^{2\alpha_2 - 2s} - m^{2\alpha_2 - 2s}, & s > \alpha_2. \end{cases}$$

We now compare these contributions. If $s < \alpha_2$, then

$$S_+(J) + T_+(m, J) \lesssim m^{2\alpha_2 - 2s},$$

so the contribution multiplied by $m^{-2\alpha_2}$ is bounded above by a constant times m^{-2s}. Together with the tail term, this gives $B_0(h_m) \asymp m^{-2s}$.

If $s = \alpha_2$, then $S_+(J) \asymp 1$, and hence

$$B_0(h_m) \asymp m^{-2\alpha_2}\big[1 + T_+(m, J)\big] \asymp m^{-2\alpha_2}\ell(m/J).$$

Here the constant contribution is retained even when $m = J$.

If $s > \alpha_2$, then

$$S_+(J) \asymp J^{-2(a-\alpha_2)}\ell_1(J).$$

Moreover,

$$\begin{aligned} T_+(m, J) &\lesssim J^{2\alpha_2 - 2s} \lesssim S_+(J), \\ m^{-2s} &\le m^{-2\alpha_2} J^{2\alpha_2 - 2s} \lesssim m^{-2\alpha_2} S_+(J). \end{aligned}$$

Thus the contribution from $S_+(J)$ controls the bias:

$$B_0(h_m) \asymp m^{-2\alpha_2} J^{-2(a-\alpha_2)}\ell_1(J).$$

Define

$$\ell_+(m, J) = 1 + \mathbf{1}_{\{s=\alpha_2\}} \log(m/J) + \mathbf{1}_{\{s=\alpha_1\}} \log J.$$

Combining the preceding cases gives

$$B_0(h_m) \asymp \begin{cases} m^{-2a}\ell_1(m), & 1 \le m \le J,\ n_2 \le n, \\ m^{-2b} J^{-2(a-b)}\ell_+(m, J), & m \ge J. \end{cases} \tag{25}$$

For the variance, recall from (23) that

$$V_0(h_m) \asymp \sum_{j \le m} \frac{1}{n + n_2 j^\delta}.$$

If $1 \le m \le J$ and $n_2 \le n$, then $n_2 j^\delta \le n$ for every $j \le m$, so

$$V_0(h_m) \asymp \frac{1}{n} \sum_{j \le m} 1 \asymp \frac{m}{n}.$$

If $m \ge J$, we split the sum at J. The first contribution satisfies

$$\sum_{j \le J} \frac{1}{n + n_2 j^\delta} \asymp n_2^{-1} J^{1-\delta}.$$

Indeed, when $n_2 \le n$, it has order $J/n = n_2^{-1} J^{1-\delta}$. When $n_2 > n$, we have $J = 1$, and this contribution equals $(n + n_2)^{-1} \asymp n_2^{-1}$.

For the second contribution, $n_2 j^\delta \geq n$ whenever $j > J$, and therefore

$$\sum_{J<j\leq m} \frac{1}{n + n_2 j^\delta} \asymp n_2^{-1} \sum_{J<j\leq m} j^{-\delta}.$$

Lemma 2 gives

$$\sum_{J<j\leq m} j^{-\delta} \asymp \begin{cases} m^{1-\delta} - J^{1-\delta}, & 0 < \delta < 1, \\ \log(m/J), & \delta = 1, \\ J^{1-\delta} - m^{1-\delta}, & \delta > 1. \end{cases}$$

Consequently,

$$V_0(h_m) \asymp n_2^{-1} \left[J^{1-\delta} + \sum_{J<j\leq m} j^{-\delta} \right].$$

For $0 < \delta < 1$, adding the first contribution yields

$$J^{1-\delta} + \sum_{J<j\leq m} j^{-\delta} \asymp J^{1-\delta} + \left(m^{1-\delta} - J^{1-\delta}\right) \asymp m^{1-\delta}.$$

For $\delta = 1$, the two contributions together have order

$$1 + \sum_{J<j\leq m} j^{-1} \asymp 1 + \log(m/J) = \ell(m/J).$$

For $\delta > 1$, the second contribution satisfies

$$0 \leq \sum_{J<j\leq m} j^{-\delta} \lesssim J^{1-\delta},$$

so it is bounded above by a constant times the first contribution. We conclude that

$$V_0(h_m) \asymp \begin{cases} m/n, & 1 \leq m \leq J,\ n_2 \leq n, \\ n_2^{-1} m^{1-\delta}, & m \geq J,\ 0 < \delta < 1, \\ n_2^{-1} \ell(m/J), & m \geq J,\ \delta = 1, \\ n_2^{-1} J^{1-\delta}, & m \geq J,\ \delta > 1. \end{cases} \tag{26}$$

When $n_2 \leq n$, the applicable expressions agree at $m = J$, since $J/n = n_2^{-1} J^{1-\delta}$ and $\ell(1) = 1$.

Finally, (24) gives

$$\epsilon(m) \asymp \frac{m^{1+\delta}}{n + n_2 m^\delta} \asymp \begin{cases} m^{1+\delta}/n, & 1 \leq m \leq J,\ n_2 \leq n, \\ m/n_2, & m \geq J. \end{cases} \tag{27}$$

Spectral sums when $\delta \leq 0$. If $\delta = 0$, then $\alpha_1 = \alpha_2$, $a = b$, and $h_j = (n + n_2) j^{-\alpha_1}$. The cutoff formulas give

$$B_0(h_m) \asymp m^{-2s} + m^{-2\alpha_1} \sum_{j\leq m} j^{2\alpha_1 - 1 - 2s},$$

$$V_0(h_m) \asymp \frac{1}{n + n_2} \sum_{j\leq m} 1.$$

The prefix-sum estimates above show that the second contribution to the bias has order $m^{-2a}\ell_1(m)$ and controls m^{-2s}. The variance sum has order $m/(n + n_2)$. Together with (24), this gives

$$B_0(h_m) \asymp m^{-2a}\ell_1(m), \qquad V_0(h_m) \asymp \epsilon(m) \asymp \frac{m}{n + n_2}. \tag{28}$$

Suppose next that $\delta < 0$. Here $\alpha_1 < \alpha_2$ and $a \leq b$. Define

$$J = \max\left\{1, \left(\frac{n_2}{n}\right)^{1/(-\delta)}\right\}.$$

Again write $m = \lfloor m \rfloor$ and $J = \lfloor J \rfloor$ when evaluating interval sums. When $n_2 \geq n$, the auxiliary dataset dominates below J and the target dataset dominates above J. When $n_2 < n$, we have $J = 1$ and $h_j \asymp nj^{-\alpha_1}$ for every $j \geq 1$.

For $1 \leq m \leq J$ with $n_2 \geq n$, $h_m \asymp n_2 m^{-\alpha_2}$, and (22) gives

$$B_0(h_m) \asymp m^{-2s} + m^{-2\alpha_2} \sum_{j \leq m} j^{2\alpha_2 - 1 - 2s}.$$

The second contribution satisfies

$$m^{-2\alpha_2} \sum_{j \leq m} j^{2\alpha_2 - 1 - 2s} \asymp \begin{cases} m^{-2s}, & s < \alpha_2, \\ m^{-2\alpha_2} \ell(m), & s = \alpha_2, \\ m^{-2\alpha_2}, & s > \alpha_2. \end{cases}$$

It controls the tail term m^{-2s} in every case, so

$$B_0(h_m) \asymp m^{-2b} \ell_2(m).$$

For $m \geq J$, splitting at J gives

$$B_0(h_m) \asymp m^{-2s} + m^{-2\alpha_1} \left[\underbrace{J^{2\delta} \sum_{j \leq J} j^{2\alpha_2 - 1 - 2s}}_{=: S_-(J)} + \underbrace{\sum_{J < j \leq m} j^{2\alpha_1 - 1 - 2s}}_{=: T_-(m, J)} \right].$$

When $n_2 \geq n$, the factor $J^{2\delta}$ follows from $(n/n_2)^2 = J^{2\delta}$. When $n_2 < n$, the same display follows by taking $J = 1$, separating the term $j = 1$, and using $h_j \asymp nj^{-\alpha_1}$ throughout.

Applying Lemma 2 to each sum gives

$$S_-(J) \asymp \begin{cases} J^{2\alpha_1 - 2s}, & s < \alpha_2, \\ J^{2\delta} \ell(J), & s = \alpha_2, \\ J^{2\delta}, & s > \alpha_2, \end{cases}$$

and

$$T_-(m, J) \asymp \begin{cases} m^{2\alpha_1 - 2s} - J^{2\alpha_1 - 2s}, & s < \alpha_1, \\ \log(m/J), & s = \alpha_1, \\ J^{2\alpha_1 - 2s} - m^{2\alpha_1 - 2s}, & s > \alpha_1. \end{cases}$$

If $s < \alpha_1$, then

$$S_-(J) + T_-(m, J) \lesssim m^{2\alpha_1 - 2s}.$$

Thus both contributions to the bias are bounded above by a constant times m^{-2s}, and the tail term gives $B_0(h_m) \asymp m^{-2s}$.

If $s = \alpha_1$, then $S_-(J) \asymp 1$, and

$$B_0(h_m) \asymp m^{-2\alpha_1} \big[1 + T_-(m, J) \big] \asymp m^{-2\alpha_1} \ell(m/J).$$

If $s > \alpha_1$, then

$$S_-(J) \asymp J^{-2(b - \alpha_1)} \ell_2(J).$$

Furthermore,

$$T_-(m, J) \lesssim J^{2\alpha_1 - 2s} \lesssim S_-(J),$$
$$m^{-2s} \leq m^{-2\alpha_1} J^{2\alpha_1 - 2s} \lesssim m^{-2\alpha_1} S_-(J).$$

Consequently,

$$B_0(h_m) \asymp m^{-2\alpha_1} J^{-2(b - \alpha_1)} \ell_2(J).$$

Define

$$\ell_-(m, J) = 1 + \mathbf{1}_{\{s = \alpha_1\}} \log(m/J) + \mathbf{1}_{\{s = \alpha_2\}} \log J.$$

Combining these estimates yields

$$B_0(h_m) \asymp \begin{cases} m^{-2b}\ell_2(m), & 1 \le m \le J,\ n_2 \ge n, \\ m^{-2a}J^{-2(b-a)}\ell_-(m,J), & m \ge J. \end{cases} \tag{29}$$

For the variance, first suppose that $1 \le m \le J$ and $n_2 \ge n$. Since $n_2 j^\delta \ge n$ for $j \le m$,

$$V_0(h_m) \asymp n_2^{-1} \sum_{j \le m} j^{-\delta}.$$

Here $1 - \delta > 0$, so Lemma 2 gives

$$\sum_{j \le m} j^{-\delta} \asymp m^{1-\delta}, \qquad V_0(h_m) \asymp n_2^{-1} m^{1-\delta}.$$

For $m \ge J$, split the variance sum as

$$V_0(h_m) \asymp \sum_{j \le J} \frac{1}{n + n_2 j^\delta} + \sum_{J < j \le m} \frac{1}{n + n_2 j^\delta}.$$

If $n_2 \ge n$, the first contribution satisfies

$$\sum_{j \le J} \frac{1}{n + n_2 j^\delta} \asymp n_2^{-1} \sum_{j \le J} j^{-\delta} \asymp n_2^{-1} J^{1-\delta} = \frac{J}{n},$$

where we used $n_2 = nJ^{-\delta}$. If $n_2 < n$, then $J = 1$, and the same contribution equals $(n + n_2)^{-1} \asymp 1/n = J/n$.

For the second contribution, $n_2 j^\delta \le n$ whenever $j > J$. Therefore,

$$\sum_{J < j \le m} \frac{1}{n + n_2 j^\delta} \asymp \frac{1}{n} \sum_{J < j \le m} 1 = \frac{m - J}{n}.$$

Adding the two contributions gives

$$V_0(h_m) \asymp \frac{J + m - J}{n} \asymp \frac{m}{n}.$$

This estimate also holds at $m = J$, when the second sum is empty.

Finally, since $\delta_+ = 0$, (24) gives

$$\epsilon(m) \asymp \frac{m}{n + n_2 m^\delta},$$

which has the same two orders as the variance. Hence

$$V_0(h_m) \asymp \epsilon(m) \asymp \begin{cases} n_2^{-1} m^{1-\delta}, & 1 \le m \le J,\ n_2 \ge n, \\ m/n, & m \ge J. \end{cases} \tag{30}$$

Feasibility of a balanced cutoff. Before choosing m, we show that any sequence satisfying

$$m \longrightarrow \infty, \qquad B_0(h_m) \asymp V_0(h_m)$$

also satisfies $\epsilon(m) = o(1)$. This verifies the feasibility condition for the constructions below.

First note that

$$2a - \delta = \min\{2s - \delta, \alpha_1 + \alpha_2\} > 0.$$

We also have $2b > \delta$ when $0 < \delta \le 1$, and $2b > 1$ when $\delta > 1$.

If $\delta \le 0$, the preceding estimates give, with $m = m$,

$$\epsilon(m) \asymp V_0(h_m) \asymp B_0(h_m) \lesssim m^{-2a}\ell(m) = o(1).$$

Here we used $b \ge a$ and $J \ge 1$ when $\delta < 0$.

Suppose $\delta > 0$. If $m \le J$, then

$$\epsilon(m) \asymp m^{\delta} V_0(h_m) \asymp m^{\delta} B_0(h_m) \lesssim m^{\delta-2a}\ell(m) = o(1).$$

If $m \ge J$ and $0 < \delta \le 1$, then

$$\begin{aligned}\epsilon(m) &\asymp m/n_2 \lesssim m^{\delta} V_0(h_m) \asymp m^{\delta} B_0(h_m) \\ &\lesssim m^{\delta-2b}\ell(m) = o(1).\end{aligned}$$

Finally, if $m \ge J$ and $\delta > 1$, then

$$\begin{aligned}\epsilon(m) &\asymp mJ^{\delta-1} V_0(h_m) \asymp mJ^{\delta-1} B_0(h_m) \\ &\lesssim m^{1-2b} J^{\delta-1-2(a-b)}\ell(m) \\ &\le m^{-\min\{2b-1,\, 2a-\delta\}}\ell(m) = o(1).\end{aligned}$$

The last inequality follows from $1 \le J \le m$.

Consequently, for every balanced sequence constructed below, $\lambda_n = h_m$ is feasible, and (20) gives

$$\overline{\mathsf{R}}_1^{*} := \inf_{\lambda > 0} \overline{\mathsf{R}}_1(\lambda) \asymp B_0(h_m) \asymp V_0(h_m). \tag{31}$$

Optimal scaling when $\delta > 0$. Define the threshold

$$\gamma_{\mathrm{c}}^{+} = 1 - \frac{\delta}{1+2a}.$$

Since $2a > \delta$, we have $0 < \gamma_{\mathrm{c}}^{+} < 1$.

First suppose $\gamma_2 < \gamma_{\mathrm{c}}^{+}$. Let

$$u_n = n^{1/(1+2a)}, \qquad m = u_n \ell_1(u_n)^{1/(1+2a)}.$$

Since $\gamma_2 < \gamma_{\mathrm{c}}^{+} < 1$, we have $J = n^{\frac{1-\gamma_2}{\delta}}$, which implies $u_n/J = n^{\frac{\gamma_2-1}{\delta} - \frac{1}{1+2a}} = o(1)$. Thus $m \le J$ for all sufficiently large n, and $\ell_1(m) \asymp \ell_1(u_n)$. Therefore,

$$\begin{aligned}B_0(h_m) &\asymp m^{-2a}\ell_1(m) \\ &\asymp n^{-2a/(1+2a)}\ell_1(u_n)^{1/(1+2a)} \asymp \frac{m}{n} \asymp V_0(h_m).\end{aligned}$$

Applying (31) gives

$$\overline{\mathsf{R}}_1^{*} \asymp n^{-2a/(1+2a)}\ell_1(u_n)^{1/(1+2a)}.$$

In particular, the polynomial exponent is

$$\Gamma = \frac{2a}{1+2a}.$$

Next suppose $\gamma_2 = \gamma_{\mathrm{c}}^{+}$. Here

$$J \asymp n^{1/(1+2a)}, \qquad \frac{J}{n} \asymp J^{-2a}.$$

If $s \ne \alpha_1$, choosing $m = J$ immediately gives

$$B_0(h_m) \asymp V_0(h_m) \asymp \frac{J}{n} \asymp n^{-2a/(1+2a)}.$$

If $s = \alpha_1$, then $a = \alpha_1$, $b = \alpha_2$, and $a - b = \delta$. Put $L_n = \ell(J)$ and write $m = Jt$ with $t \ge 1$. Equations (25) and (26) give

$$B_0(h_{Jt}) \asymp \frac{J}{n} t^{-2b} L_n, \qquad V_0(h_{Jt}) \asymp \frac{J}{n} \begin{cases} t^{1-\delta}, & 0 < \delta < 1, \\ \ell(t), & \delta = 1, \\ 1, & \delta > 1. \end{cases}$$

We therefore choose $m = Jt_n$, where

$$t_n = \begin{cases} L_n^{1/(1+2b-\delta)}, & 0 < \delta < 1, \\ \big(L_n/\ell(L_n)\big)^{1/(2b)}, & \delta = 1, \\ L_n^{1/(2b)}, & \delta > 1. \end{cases}$$

For $\delta = 1$, we used $\ell(t_n) \asymp \ell(L_n)$. These choices yield

$$\overline{\mathsf{R}}_1^* \asymp B_0(h_m) \asymp V_0(h_m) \asymp \frac{J}{n} \begin{cases} L_n^{(1-\delta)/(1+2b-\delta)}, & 0 < \delta < 1, \\ \ell(L_n), & \delta = 1, \\ 1, & \delta > 1. \end{cases}$$

Thus the boundary $\gamma_2 = \gamma_{\mathrm{c}}^+$ has the same polynomial exponent $\Gamma = 2a/(1+2a)$. In particular, the fixed-signal critical case $s = \alpha_1$, $\delta = 1$, and $\gamma_2 = \gamma_{\mathrm{c}}^+$ has an additional factor of order $\log\log n$.

It remains to consider $\gamma_2 > \gamma_{\mathrm{c}}^+$. We treat the three variance regimes separately.

If $0 < \delta < 1$, define

$$u_n = \Big(n_2 J^{-2(a-b)}\Big)^{1/(1+2b-\delta)}, \qquad m = u_n \ell_+(u_n, J)^{1/(1+2b-\delta)}.$$

The denominator is positive because $2b > \delta$. Moreover,

$$\frac{u_n}{J} = \Big(n_2 J^{-(1+2a-\delta)}\Big)^{1/(1+2b-\delta)} \longrightarrow \infty.$$

Hence $m \geq J$ and $\ell_+(m, J) \asymp \ell_+(u_n, J)$. By construction,

$$u_n^{-2b} J^{-2(a-b)} = n_2^{-1} u_n^{1-\delta}.$$

Consequently,

$$\begin{aligned} B_0(h_m) &\asymp m^{-2b} J^{-2(a-b)} \ell_+(m, J) \\ &\asymp n_2^{-1} u_n^{1-\delta} \ell_+(u_n, J)^{(1-\delta)/(1+2b-\delta)} \\ &\asymp n_2^{-1} m^{1-\delta} \asymp V_0(h_m). \end{aligned}$$

Since

$$n_2^{-1} u_n^{1-\delta} = n_2^{-2b/(1+2b-\delta)} J^{-2(a-b)(1-\delta)/(1+2b-\delta)}$$

and $J \asymp n^{(1-\gamma_2)_+/\delta}$, the polynomial exponent is

$$\Gamma = \frac{2b\gamma_2 + \frac{2(1-\delta)}{\delta}(a-b)(1-\gamma_2)_+}{1+2b-\delta}.$$

If $\delta = 1$, define

$$u_n = \Big(n_2 J^{-2(a-b)}\Big)^{1/(2b)}, \qquad m = u_n \left(\frac{\ell_+(u_n, J)}{\ell(u_n/J)}\right)^{1/(2b)}.$$

Again, u_n/J grows as a positive power of n. The logarithmic correction preserves $m/J \to \infty$, and

$$\ell_+(m, J) \asymp \ell_+(u_n, J), \qquad \ell(m/J) \asymp \ell(u_n/J) \asymp \log n.$$

It follows that

$$\begin{aligned} B_0(h_m) &\asymp m^{-2b} J^{-2(a-b)} \ell_+(m, J) \\ &\asymp n_2^{-1} \ell(u_n/J) \asymp V_0(h_m). \end{aligned}$$

Therefore,

$$\overline{\mathsf{R}}_1^* \asymp n^{-\gamma_2} \log n, \qquad \Gamma = \gamma_2.$$

If $\delta > 1$, define

$$u_n = \Big(n_2 J^{\delta-1-2(a-b)}\Big)^{1/(2b)}, \qquad m = u_n \ell_+(u_n, J)^{1/(2b)}.$$

Here

$$\frac{u_n}{J} = \left(n_2 J^{-(1+2a-\delta)}\right)^{1/(2b)} \longrightarrow \infty$$

at a polynomial rate. Thus $m \geq J$ and $\ell_+(m, J) \asymp \ell_+(u_n, J)$. We obtain

$$\begin{aligned} B_0(h_m) &\asymp m^{-2b} J^{-2(a-b)} \ell_+(m, J) \\ &\asymp u_n^{-2b} J^{-2(a-b)} = n_2^{-1} J^{1-\delta} \asymp V_0(h_m). \end{aligned}$$

Hence

$$\overline{\mathsf{R}}_1^* \asymp n_2^{-1} J^{1-\delta} \asymp \begin{cases} n^{-(\delta-1+\gamma_2)/\delta}, & \gamma_{\mathrm{c}}^+ < \gamma_2 < 1, \\ n^{-\gamma_2}, & \gamma_2 \geq 1. \end{cases}$$

Optimal scaling when $\delta = 0$. In this case the two datasets have the same spectral decay. Let

$$u_n = (n + n_2)^{1/(1+2a)}, \qquad m = u_n \ell_1(u_n)^{1/(1+2a)}.$$

Using (28), we obtain

$$\begin{aligned} B_0(h_m) &\asymp m^{-2a} \ell_1(m) \\ &\asymp (n + n_2)^{-2a/(1+2a)} \ell_1(u_n)^{1/(1+2a)} \\ &\asymp \frac{m}{n + n_2} \asymp V_0(h_m). \end{aligned}$$

Therefore the polynomial exponent is

$$\Gamma = \frac{2a \max\{1, \gamma_2\}}{1 + 2a}.$$

Optimal scaling when $\delta < 0$. Define

$$\gamma_{\mathrm{c}}^- = 1 - \frac{\delta}{1 + 2b} > 1.$$

First suppose $\gamma_2 \leq \gamma_{\mathrm{c}}^-$. Let

$$u_n = \left(n J^{-2(b-a)}\right)^{1/(1+2a)}, \qquad m = u_n \ell_-(u_n, J)^{1/(1+2a)}.$$

We have $u_n \geq J$ for all sufficiently large n. Indeed, when $\gamma_2 \leq 1$, $J = 1$, whereas for $1 < \gamma_2 \leq \gamma_{\mathrm{c}}^-$,

$$\frac{u_n}{J} = \left(n J^{-(1+2b)}\right)^{1/(1+2a)} \geq 1.$$

The last inequality follows from $J \leq n^{(\gamma_2-1)/(-\delta)}$ and the definition of γ_{c}^-. Moreover, $u_n \to \infty$ at a polynomial rate, including at $\gamma_2 = \gamma_{\mathrm{c}}^-$.

Since $m \geq u_n \geq J$, we can use the second cases of (29) and (30). The definition of ℓ_- gives $\ell_-(m, J) \asymp \ell_-(u_n, J)$, including when $u_n \asymp J$. Thus

$$\begin{aligned} B_0(h_m) &\asymp m^{-2a} J^{-2(b-a)} \ell_-(m, J) \\ &\asymp \frac{u_n}{n} \ell_-(u_n, J)^{1/(1+2a)} \\ &= \frac{m}{n} \asymp V_0(h_m). \end{aligned}$$

Using

$$\frac{u_n}{n} = n^{-2a/(1+2a)} J^{-2(b-a)/(1+2a)}, \qquad J \asymp n^{(\gamma_2-1)_+/(-\delta)},$$

we obtain

$$\Gamma = \frac{2a + \frac{2(b-a)}{-\delta}(\gamma_2 - 1)_+}{1 + 2a}, \qquad \gamma_2 \leq \gamma_{\mathrm{c}}^-.$$

Finally, suppose $\gamma_2 > \gamma_{\mathrm{c}}^-$. Define

$$u_n = n_2^{1/(1+2b-\delta)}, \qquad m = u_n \ell_2(u_n)^{1/(1+2b-\delta)}.$$

The strict inequality $\gamma_2 > \gamma_c^-$ implies that $u_n/J \to 0$ at a polynomial rate. Therefore $m \le J$ for all sufficiently large n and $\ell_2(m) \asymp \ell_2(u_n)$. Since $u_n^{1+2b-\delta} = n_2$, we have

$$\begin{aligned} B_0(h_m) &\asymp m^{-2b}\ell_2(m) \\ &\asymp u_n^{-2b}\ell_2(u_n)^{(1-\delta)/(1+2b-\delta)} \\ &\asymp n_2^{-1}m^{1-\delta} \asymp V_0(h_m). \end{aligned}$$

Consequently,

$$\Gamma = \frac{2b\gamma_2}{1+2b-\delta}, \qquad \gamma_2 > \gamma_c^-.$$

Every cutoff constructed above satisfies $m \to \infty$ and $B_0(h_m) \asymp V_0(h_m)$. The feasibility argument therefore gives $\epsilon_n = \epsilon(m) = o(1)$ in every regime, including the threshold cases. Applying (20) with $\lambda_n = h_m$ proves the stated polynomial exponents. More precisely, all the displayed logarithmic factors lie between a positive constant and a constant times $1 + \log n$, so

$$n^{-\Gamma} \lesssim \overline{\mathsf{R}}_1^* \lesssim n^{-\Gamma}(1 + \log n).$$

This completes the proof.

D.1 Technical Lemma

Lemma 2 (Power-law sums)**.** *For every fixed $u \in \mathbb{R}$ and all $m \ge 1$,*

$$\sum_{j \le m} j^{u-1} \asymp \begin{cases} m^u, & u > 0, \\ 1 + \log m, & u = 0, \\ 1, & u < 0. \end{cases}$$

Moreover, for $u < 0$,

$$\sum_{j \ge m} j^{u-1} \asymp m^u,$$

whereas the tail sum diverges for $u \ge 0$.

For integers $1 \le m_- < m_+$,

$$\sum_{m_- < j \le m_+} j^{u-1} \asymp \begin{cases} m_+^u - m_-^u, & u > 0, \\ \log(m_+/m_-), & u = 0, \\ m_-^u - m_+^u, & u < 0. \end{cases}$$

For real endpoints $1 \le m_- < m_+$, the same interval estimate holds with $m_\pm$ on the right-hand side replaced by $\lfloor m_\pm \rfloor$. The sum is zero if these two integers coincide.

All comparison constants may depend on u but are independent of the summation endpoints.

Proof. We first consider integer endpoints. For every $j \ge 2$ and $x \in [j-1, j]$, we have $j/2 \le x \le j$. Consequently,

$$2^{-|u-1|}j^{u-1} \le \int_{j-1}^{j} x^{u-1}\,\mathrm{d}x \le 2^{|u-1|}j^{u-1}.$$

Summing these inequalities gives

$$\sum_{j \le m} j^{u-1} \asymp 1 + \int_1^m x^{u-1}\,\mathrm{d}x.$$

If $u \ne 0$, the integral equals $(m^u - 1)/u$, whereas for $u = 0$ it equals $\log m$. Hence

$$1 + \int_1^m x^{u-1}\,\mathrm{d}x \asymp \begin{cases} m^u, & u > 0, \\ 1 + \log m, & u = 0, \\ 1, & u < 0, \end{cases}$$

which proves the prefix-sum estimate.

Similarly, since $j \le x \le 2j$ for $x \in [j, j+1]$ and $j \ge 1$,

$$j^{u-1} \asymp \int_j^{j+1} x^{u-1}\,\mathrm{d}x.$$

For $u < 0$, summing over $j \ge m$ yields

$$\sum_{j \ge m} j^{u-1} \asymp \int_m^{\infty} x^{u-1}\,\mathrm{d}x = \frac{m^u}{-u} \asymp m^u.$$

For $u \ge 0$, the integral diverges, and the same comparison proves that the tail sum diverges.

For the interval sum, we obtain

$$\sum_{m_- < j \le m_+} j^{u-1} \asymp \int_{m_-}^{m_+} x^{u-1}\,\mathrm{d}x = \begin{cases} \dfrac{m_+^u - m_-^u}{u}, & u \ne 0, \\ \log(m_+/m_-), & u = 0. \end{cases}$$

Separating the cases $u > 0$ and $u < 0$, and absorbing the fixed factor $1/|u|$ into the comparison constants, proves the stated estimate.

Finally, for real endpoints,

$$\sum_{j \le m} j^{u-1} = \sum_{j \le \lfloor m \rfloor} j^{u-1}, \qquad \sum_{j \ge m} j^{u-1} = \sum_{j \ge \lceil m \rceil} j^{u-1}.$$

Since $\lfloor m \rfloor \asymp m \asymp \lceil m \rceil$ for $m \ge 1$, the prefix and tail estimates remain valid. For the interval sum, the exact identity

$$\sum_{m_- < j \le m_+} j^{u-1} = \sum_{\lfloor m_- \rfloor < j \le \lfloor m_+ \rfloor} j^{u-1}$$

gives the asserted extension. □

E Proof of Lemma 1

Our goal is to prove that the approximation error is small relative to $\overline{\mathsf{R}}_1(\lambda_n)$. We compare both $\mathsf{R}(\widehat{\boldsymbol{\theta}})$ and $\overline{\mathsf{R}}_1(\lambda_n)$ with the same population ridge risk. Specifically, define

$$\boldsymbol{G}_0 = (n\boldsymbol{C}_1 + n_2\boldsymbol{C}_2 + \lambda_n \boldsymbol{I})^{-1}, \qquad B_0 = \lambda_n^2 \boldsymbol{\theta}_*^\top \boldsymbol{G}_0 \boldsymbol{C}_1 \boldsymbol{G}_0 \boldsymbol{\theta}_*, \qquad V_\sigma = \sum_{i=1}^{2} \sigma_{\varepsilon_i}^2 n_i \mathrm{Tr}(\boldsymbol{C}_1 \boldsymbol{G}_0 \boldsymbol{C}_i \boldsymbol{G}_0).$$

Here B_0 is defined in (19), and $V_\sigma \asymp V_0$ in (19) because the noise variances are bounded above and below. Recall that the proof of Theorem 4 constructs a sequence $\lambda_n = h_m$ such that

$$\epsilon_n \lesssim -c_-, \quad m \gtrsim n^{1/(1+2s)}.$$

for some fixed $c_- > 0$. Throughout the proof, we suppress the subscript n when no confusion can arise.

To bound the approximation error, the triangle inequality gives

$$|\mathsf{R}(\widehat{\boldsymbol{\theta}}) - \overline{\mathsf{R}}_1(\lambda_n)| \le |\mathsf{R}(\widehat{\boldsymbol{\theta}}) - (B_0 + V_\sigma)| + |\overline{\mathsf{R}}_1(\lambda_n) - (B_0 + V_\sigma)|.$$

Accordingly, we will prove the following two estimates:

$$|\mathsf{R}(\widehat{\boldsymbol{\theta}}) - (B_0 + V_\sigma)| \le C_D (\log N)^c \left(\sqrt{\epsilon} + \sqrt{\lambda_n \epsilon} + \lambda_n \epsilon\right)(B_0 + V_\sigma), \tag{32}$$

$$|\overline{\mathsf{R}}_1(\lambda_n) - (B_0 + V_\sigma)| \le C(\epsilon + \lambda_n \epsilon)(B_0 + V_\sigma). \tag{33}$$

The first estimate holds with probability at least $1 - N^{-D}$, whereas the second is deterministic.

We first explain why these two estimates imply the desired conclusion.

We note that $\lambda_n \epsilon \asymp m^{1-\alpha_{\min}}$. Since the same domain maximizes $j^{-\alpha_i}$ for every j,

$$\epsilon = \sum_{j \geq 1} \frac{j^{-\alpha_{\min}}}{h_j + \lambda_n}.$$

For $j \leq m$, the inequality $h_j \geq \lambda_n (m/j)^{\alpha_{\min}}$ gives

$$\sum_{j \leq m} \frac{j^{-\alpha_{\min}}}{h_j + \lambda_n} \leq \frac{m^{1-\alpha_{\min}}}{\lambda_n}.$$

For $j > m$, we have $\lambda_n \leq h_j + \lambda_n \leq 2\lambda_n$, so

$$\sum_{j > m} \frac{j^{-\alpha_{\min}}}{h_j + \lambda_n} \asymp \frac{1}{\lambda_n} \sum_{j > m} j^{-\alpha_{\min}} \asymp \frac{m^{1-\alpha_{\min}}}{\lambda_n}.$$

Dividing the triangle inequality by $\overline{\mathsf{R}}_1(\lambda_n)$ and applying (32)–(33), we obtain

$$\begin{aligned} \frac{|\mathsf{R}(\widehat{\boldsymbol{\theta}}) - \overline{\mathsf{R}}_1(\lambda_n)|}{\overline{\mathsf{R}}_1(\lambda_n)} &\leq C_D (\log N)^c \left(\sqrt{\epsilon} + \sqrt{\lambda_n \epsilon} + \lambda_n \epsilon \right) + C(\epsilon + \lambda_n \epsilon) \\ &\leq C_D (\log N)^c \left(\sqrt{\epsilon} + m^{-(\alpha_{\min}-1)/2} \right), \end{aligned}$$

where in the last step, we use $\varepsilon \leq \sqrt{\varepsilon}, \lambda_n \varepsilon \leq \sqrt{\lambda_n \varepsilon}$, since $\alpha_{\min} > 1$, $m \geq n^{c_-}$, and $\epsilon \leq n^{-c_-}$, so $\epsilon, \lambda_n \epsilon = o(1)$. And this finishes the proof.

It remains to establish the two comparison estimates. For (32), we control the normalized empirical covariance and then bound the resulting perturbations of the population bias and variance. For (33), we use the fixed-point equations to compare μ_i with n_i, and then compare the deterministic bias and variance with B_0 and V_σ.

Proof of (32)**.** We start from

$$|\mathsf{R}(\widehat{\boldsymbol{\theta}}) - (B_0 + V_\sigma)| \leq |\mathsf{B}_1 - B_0| + |\mathsf{V}_1 - V_\sigma|.$$

Define the normalized covariance errors

$$\boldsymbol{E} = \boldsymbol{G}_0^{1/2} \left(\sum_{i=1}^{2} (\boldsymbol{X}_i^\top \boldsymbol{X}_i - n_i \boldsymbol{C}_i) \right) \boldsymbol{G}_0^{1/2},$$

$$\boldsymbol{E}_\sigma = \boldsymbol{G}_0^{1/2} \left(\sum_{i=1}^{2} \sigma_{\varepsilon_i}^2 (\boldsymbol{X}_i^\top \boldsymbol{X}_i - n_i \boldsymbol{C}_i) \right) \boldsymbol{G}_0^{1/2}.$$

We first derive deterministic bounds for the bias and variance differences under $\|\boldsymbol{E}\|_{\mathrm{op}} \leq 1/2$. We then establish, independently of this condition, a high-probability bound on $\|\boldsymbol{E}\|_{\mathrm{op}}$ and $\|\boldsymbol{E}_\sigma\|_{\mathrm{op}}$.

Bounding $\mathsf{B}_1 - B_0$. The definition of $\boldsymbol{E}$ gives

$$\boldsymbol{G}^{-1} = \boldsymbol{G}_0^{-1/2} (\boldsymbol{I} + \boldsymbol{E}) \boldsymbol{G}_0^{-1/2}, \qquad \boldsymbol{G} = \boldsymbol{G}_0^{1/2} (\boldsymbol{I} + \boldsymbol{E})^{-1} \boldsymbol{G}_0^{1/2}.$$

When $\|\boldsymbol{E}\|_{\mathrm{op}} \leq 1/2$, we have $\boldsymbol{I} + \boldsymbol{E} \succeq \boldsymbol{I}/2$, and hence

$$\|(\boldsymbol{I} + \boldsymbol{E})^{-1}\|_{\mathrm{op}} \leq 2, \qquad \|(\boldsymbol{I} + \boldsymbol{E})^{-1} - \boldsymbol{I}\|_{\mathrm{op}} = \|(\boldsymbol{I} + \boldsymbol{E})^{-1} \boldsymbol{E}\|_{\mathrm{op}} \leq 2\|\boldsymbol{E}\|_{\mathrm{op}}.$$

Here we used $(\boldsymbol{I} + \boldsymbol{E})^{-1} - \boldsymbol{I} = -(\boldsymbol{I} + \boldsymbol{E})^{-1} \boldsymbol{E}$.

Since

$$\mathsf{B}_1 = \|\lambda_n \boldsymbol{C}_1^{1/2} \boldsymbol{G} \boldsymbol{\theta}_*\|^2, \qquad B_0 = \|\lambda_n \boldsymbol{C}_1^{1/2} \boldsymbol{G}_0 \boldsymbol{\theta}_*\|^2,$$

the triangle inequality implies

$$\begin{aligned} |\sqrt{\mathsf{B}_1} - \sqrt{B_0}| &\leq \lambda_n \|\boldsymbol{C}_1^{1/2} (\boldsymbol{G} - \boldsymbol{G}_0) \boldsymbol{\theta}_*\| \\ &= \lambda_n \|\boldsymbol{C}_1^{1/2} \boldsymbol{G}_0^{1/2} ((\boldsymbol{I} + \boldsymbol{E})^{-1} - \boldsymbol{I}) \boldsymbol{G}_0^{1/2} \boldsymbol{\theta}_*\| \\ &\leq 2\lambda_n \|\boldsymbol{E}\|_{\mathrm{op}} \|\boldsymbol{C}_1^{1/2} \boldsymbol{G}_0^{1/2}\|_{\mathrm{op}} \|\boldsymbol{G}_0^{1/2} \boldsymbol{\theta}_*\|. \end{aligned}$$

The remaining factors satisfy

$$\|\boldsymbol{C}_1^{1/2}\boldsymbol{G}_0^{1/2}\|_{\mathrm{op}}^2 \le \mathrm{Tr}(\boldsymbol{C}_1\boldsymbol{G}_0), \qquad \|\boldsymbol{G}_0^{1/2}\boldsymbol{\theta}_*\|^2 \le \lambda_n^{-1}\|\boldsymbol{\theta}_*\|^2.$$

Since $\beta > 1/2$, $\|\boldsymbol{\theta}_*\|^2 = \sum_{j\ge 1} j^{-2\beta} < \infty$. Therefore,

$$|\sqrt{\mathsf{B}_1} - \sqrt{B_0}|^2 \le C\lambda_n\|\boldsymbol{E}\|_{\mathrm{op}}^2\mathrm{Tr}(\boldsymbol{C}_1\boldsymbol{G}_0).$$

Expanding the difference of squares yields

$$\begin{aligned}|\mathsf{B}_1 - B_0| &\le 2\sqrt{B_0}\,|\sqrt{\mathsf{B}_1} - \sqrt{B_0}| + |\sqrt{\mathsf{B}_1} - \sqrt{B_0}|^2 \\ &\le C\|\boldsymbol{E}\|_{\mathrm{op}}\sqrt{\lambda_n B_0 \mathrm{Tr}(\boldsymbol{C}_1\boldsymbol{G}_0)} + C\lambda_n\|\boldsymbol{E}\|_{\mathrm{op}}^2\mathrm{Tr}(\boldsymbol{C}_1\boldsymbol{G}_0).\end{aligned}$$

Bounding $\mathsf{V}_1 - V_\sigma$. By the definitions of V_1 and V_σ,

$$\begin{aligned}\mathsf{V}_1 - V_\sigma = \mathrm{Tr}&\left[\boldsymbol{C}_1\boldsymbol{G}\left(\sum_{i=1}^2 \sigma_{\varepsilon_i}^2(\boldsymbol{X}_i^\top\boldsymbol{X}_i - n_i\boldsymbol{C}_i)\right)\boldsymbol{G}\right] \\ &+ \sum_{i=1}^2 \sigma_{\varepsilon_i}^2 n_i \mathrm{Tr}\left[\boldsymbol{C}_1(\boldsymbol{G}-\boldsymbol{G}_0)\boldsymbol{C}_i\boldsymbol{G}\right] \\ &+ \sum_{i=1}^2 \sigma_{\varepsilon_i}^2 n_i \mathrm{Tr}\left[\boldsymbol{C}_1\boldsymbol{G}_0\boldsymbol{C}_i(\boldsymbol{G}-\boldsymbol{G}_0)\right].\end{aligned}$$

For the first term, by the definition of $\boldsymbol{E}_\sigma$,

$$\begin{aligned}&\mathrm{Tr}\left[\boldsymbol{C}_1\boldsymbol{G}\left(\sum_{i=1}^2 \sigma_{\varepsilon_i}^2(\boldsymbol{X}_i^\top\boldsymbol{X}_i - n_i\boldsymbol{C}_i)\right)\boldsymbol{G}\right] \\ &\quad= \mathrm{Tr}\Big[\boldsymbol{C}_1\big(\boldsymbol{G}_0^{1/2}(\boldsymbol{I}+\boldsymbol{E})^{-1}\boldsymbol{G}_0^{1/2}\big)\big(\boldsymbol{G}_0^{-1/2}\boldsymbol{E}_\sigma\boldsymbol{G}_0^{-1/2}\big)\big(\boldsymbol{G}_0^{1/2}(\boldsymbol{I}+\boldsymbol{E})^{-1}\boldsymbol{G}_0^{1/2}\big)\Big] \\ &\quad= \mathrm{Tr}\left[\boldsymbol{C}_1\boldsymbol{G}_0(\boldsymbol{I}+\boldsymbol{E})^{-1}\boldsymbol{E}_\sigma(\boldsymbol{I}+\boldsymbol{E})^{-1}\right] \\ &\quad\le 4\|\boldsymbol{E}_\sigma\|_{\mathrm{op}}\mathrm{Tr}(\boldsymbol{C}_1\boldsymbol{G}_0).\end{aligned}$$

For the other two terms, observe that

$$0 \preceq \boldsymbol{G}_0^{1/2}\left(\sum_{i=1}^2 \sigma_{\varepsilon_i}^2 n_i\boldsymbol{C}_i\right)\boldsymbol{G}_0^{1/2} \preceq \left(\max_i \sigma_{\varepsilon_i}^2\right)(\boldsymbol{I} - \lambda_n\boldsymbol{G}_0) \preceq \left(\max_i \sigma_{\varepsilon_i}^2\right)\boldsymbol{I}.$$

Substituting

$$\boldsymbol{G} - \boldsymbol{G}_0 = \boldsymbol{G}_0^{1/2}\left((\boldsymbol{I}+\boldsymbol{E})^{-1} - \boldsymbol{I}\right)\boldsymbol{G}_0^{1/2}$$

into the second term, and using cyclicity of the trace, gives

$$\begin{aligned}&\left|\sum_{i=1}^2 \sigma_{\varepsilon_i}^2 n_i \mathrm{Tr}\left[\boldsymbol{C}_1(\boldsymbol{G}-\boldsymbol{G}_0)\boldsymbol{C}_i\boldsymbol{G}\right]\right| \\ &\quad\le \mathrm{Tr}(\boldsymbol{C}_1\boldsymbol{G}_0)\|(\boldsymbol{I}+\boldsymbol{E})^{-1} - \boldsymbol{I}\|_{\mathrm{op}}\left\|\boldsymbol{G}_0^{1/2}\left(\sum_{i=1}^2 \sigma_{\varepsilon_i}^2 n_i\boldsymbol{C}_i\right)\boldsymbol{G}_0^{1/2}\right\|_{\mathrm{op}}\|(\boldsymbol{I}+\boldsymbol{E})^{-1}\|_{\mathrm{op}} \\ &\quad\le 4\left(\max_i \sigma_{\varepsilon_i}^2\right)\|\boldsymbol{E}\|_{\mathrm{op}}\mathrm{Tr}(\boldsymbol{C}_1\boldsymbol{G}_0).\end{aligned}$$

The third term is bounded in the same way, with no final inverse factor:

$$\left|\sum_{i=1}^2 \sigma_{\varepsilon_i}^2 n_i \mathrm{Tr}\left[\boldsymbol{C}_1\boldsymbol{G}_0\boldsymbol{C}_i(\boldsymbol{G}-\boldsymbol{G}_0)\right]\right| \le 2\left(\max_i \sigma_{\varepsilon_i}^2\right)\|\boldsymbol{E}\|_{\mathrm{op}}\mathrm{Tr}(\boldsymbol{C}_1\boldsymbol{G}_0).$$

Combining the three terms and using the bounded noise variances, we obtain

$$|\mathsf{V}_1 - V_\sigma| \le C\left(\|\boldsymbol{E}\|_{\mathrm{op}} + \|\boldsymbol{E}_\sigma\|_{\mathrm{op}}\right)\mathrm{Tr}(\boldsymbol{C}_1\boldsymbol{G}_0).$$

Bounding the common trace factor. We next show that $\mathrm{Tr}(\boldsymbol{C}_1\boldsymbol{G}_0) \lesssim V_0$. Recall

$$\mathrm{Tr}(\boldsymbol{C}_1\boldsymbol{G}_0) = \sum_{j\geq 1} \frac{j^{-\alpha_1}}{h_j+\lambda_n}, \qquad V_0 = \sum_{j\geq 1} \frac{j^{-\alpha_1} h_j}{(h_j+\lambda_n)^2}.$$

For $j \leq m$, we have $h_j \geq \lambda_n$, so

$$\sum_{j\leq m} \frac{j^{-\alpha_1}}{h_j+\lambda_n} \leq 2 \sum_{j\leq m} \frac{j^{-\alpha_1} h_j}{(h_j+\lambda_n)^2} \leq 2V_0.$$

For the tail,

$$\sum_{j>m} \frac{j^{-\alpha_1}}{h_j+\lambda_n} \leq \frac{1}{\lambda_n} \sum_{j>m} j^{-\alpha_1} \lesssim \frac{m^{1-\alpha_1}}{\lambda_n}.$$

On the block $m/2 \leq j \leq m$, we have $h_j \asymp \lambda_n$ and $j^{-\alpha_1} \asymp m^{-\alpha_1}$. Hence

$$V_0 \geq \sum_{m/2\leq j\leq m} \frac{j^{-\alpha_1} h_j}{(h_j+\lambda_n)^2} \gtrsim \frac{m^{1-\alpha_1}}{\lambda_n}.$$

Combining the head and tail estimates proves $\mathrm{Tr}(\boldsymbol{C}_1\boldsymbol{G}_0) \lesssim V_0$.

Substituting this comparison into the preceding bias and variance bounds gives

$$\begin{aligned} |\mathsf{B}_1 - B_0| &\leq C\|\boldsymbol{E}\|_{\mathrm{op}} \sqrt{\lambda_n B_0 V_0} + C\lambda_n \|\boldsymbol{E}\|_{\mathrm{op}}^2 V_0 \\ &\leq C\left(\sqrt{\lambda_n}\|\boldsymbol{E}\|_{\mathrm{op}} + \lambda_n \|\boldsymbol{E}\|_{\mathrm{op}}^2\right)(B_0+V_0), \\ |\mathsf{V}_1 - V_\sigma| &\leq C\left(\|\boldsymbol{E}\|_{\mathrm{op}} + \|\boldsymbol{E}_\sigma\|_{\mathrm{op}}\right) V_0. \end{aligned}$$

Here we used $2\sqrt{B_0V_0} \leq B_0 + V_0$. Since $V_\sigma \asymp V_0$, we conclude that, whenever $\|\boldsymbol{E}\|_{\mathrm{op}} \leq 1/2$,

$$\frac{|\mathsf{R}(\widehat{\boldsymbol{\theta}}) - (B_0+V_\sigma)|}{B_0+V_\sigma} \leq C\left(\|\boldsymbol{E}\|_{\mathrm{op}} + \|\boldsymbol{E}_\sigma\|_{\mathrm{op}} + \sqrt{\lambda_n}\|\boldsymbol{E}\|_{\mathrm{op}} + \lambda_n\|\boldsymbol{E}\|_{\mathrm{op}}^2\right).$$

Bounding the normalized covariance errors. It remains to prove, directly from Assumption 1, that

$$\max\{\|\boldsymbol{E}\|_{\mathrm{op}}, \|\boldsymbol{E}_\sigma\|_{\mathrm{op}}\} \leq C_D(\log N)^c\sqrt{\epsilon}$$

with probability at least $1 - N^{-D}$. We first define

$$\boldsymbol{M}_{i,r} = \boldsymbol{G}_0^{1/2}\boldsymbol{x}_{i,r}\boldsymbol{x}_{i,r}^\top \boldsymbol{G}_0^{1/2}, \qquad \boldsymbol{Z}_{i,r} = \boldsymbol{M}_{i,r}\mathbf{1}_{\{\boldsymbol{x}_{i,r}^\top \boldsymbol{G}_0 \boldsymbol{x}_{i,r} \leq C_A\epsilon(\log N)^{c_1}\}}.$$

and decompose

$$\boldsymbol{E} = \sum_{i,r}(\boldsymbol{Z}_{i,r} - \mathbb{E}\boldsymbol{Z}_{i,r}) + \sum_{i,r}(\boldsymbol{M}_{i,r} - \boldsymbol{Z}_{i,r}) - \sum_{i,r}\mathbb{E}[\boldsymbol{M}_{i,r} - \boldsymbol{Z}_{i,r}].$$

We control the first term via matrix Bernstein inequality. By the definition of ϵ, $\mathrm{Tr}(\boldsymbol{C}_i\boldsymbol{G}_0) = \sum_{j\geq 1}\frac{j^{-\alpha_i}}{h_j+\lambda_n} \leq \epsilon$. Since each $\boldsymbol{Z}_{i,r}$ depends only on $\boldsymbol{x}_{i,r}$, they are still independent with each other. By Assumption 1, using $\|\boldsymbol{C}_i^{1/2}\boldsymbol{G}_0\boldsymbol{C}_i^{1/2}\|_F \leq \mathrm{Tr}(\boldsymbol{C}_i\boldsymbol{G}_0) \leq \epsilon$ gives

$$\mathbb{P}\left(\boldsymbol{x}_{i,r}^\top \boldsymbol{G}_0 \boldsymbol{x}_{i,r} > C_A\epsilon(\log N)^{c_1}\right) \leq N^{-A},$$

for any $A > 0$ with some constants C_A. Thus, we have$\|\boldsymbol{M}_{i,r}\|_{\mathrm{op}} = \boldsymbol{x}_{i,r}^\top \boldsymbol{G}_0 \boldsymbol{x}_{i,r}$, meaning that with probability $1 - N^{-A}$,

$$\|\boldsymbol{Z}_{i,r} - \mathbb{E}\boldsymbol{Z}_{i,r}\|_{\mathrm{op}} \leq C_A\epsilon(\log N)^{c_1}.$$

For the second moment, note that

$$\boldsymbol{M}_{i,r}^2 = (\boldsymbol{x}_{i,r}^\top \boldsymbol{G}_0 \boldsymbol{x}_{i,r})\boldsymbol{M}_{i,r}.$$

Consequently,

$$\begin{aligned}\boldsymbol{Z}_{i,r}^2 &= (\boldsymbol{x}_{i,r}^\top \boldsymbol{G}_0 \boldsymbol{x}_{i,r}) \boldsymbol{M}_{i,r} \mathbf{1}_{\{\boldsymbol{x}_{i,r}^\top \boldsymbol{G}_0 \boldsymbol{x}_{i,r} \le C_A \epsilon (\log N)^{c_1}\}} \\ &\preceq C_A \epsilon (\log N)^{c_1} \boldsymbol{M}_{i,r}.\end{aligned}$$

Using $\mathbb{E}\boldsymbol{M}_{i,r} = \boldsymbol{G}_0^{1/2} \boldsymbol{C}_i \boldsymbol{G}_0^{1/2}$, we obtain

$$\sum_{i,r} \mathbb{E}[(\boldsymbol{Z}_{i,r} - \mathbb{E}\boldsymbol{Z}_{i,r})^2] \preceq \sum_{i,r} \mathbb{E}[\boldsymbol{Z}_{i,r}^2] \preceq C_A \epsilon (\log N)^{c_1} \boldsymbol{G}_0^{1/2} \left(\sum_{i=1}^{2} n_i \boldsymbol{C}_i \right) \boldsymbol{G}_0^{1/2}.$$

We show that

$$\left\| \boldsymbol{G}_0^{1/2} \left(\sum_i n_i \boldsymbol{C}_i \right) \boldsymbol{G}_0^{1/2} \right\|_{\mathrm{op}} \in [1/2, 1], \qquad \mathrm{Tr}\left[\boldsymbol{G}_0^{1/2} \left(\sum_i n_i \boldsymbol{C}_i \right) \boldsymbol{G}_0^{1/2} \right] \asymp m.$$

To see this, by definition,

$$\boldsymbol{G}_0^{1/2} \left(\sum_{i=1}^{2} n_i \boldsymbol{C}_i \right) \boldsymbol{G}_0^{1/2} = \boldsymbol{I} - \lambda_n \boldsymbol{G}_0 = \mathrm{Diag}\left(\frac{h_j}{h_j + \lambda_n} \right)_{j \ge 1}.$$

The operator norm of the above lies in $[1/2, 1]$ as each of its diagonal entries is at most one, and the first one is at least $1/2$ because $h_1 \ge h_m = \lambda_n$, and thus its trace is of order m. Indeed, the first $\lfloor m \rfloor$ entries are at least $1/2$, while

$$\sum_{j>m} \frac{h_j}{h_j + \lambda_n} \le \sum_{j>m} (m/j)^{\alpha_{\min}} \lesssim m,$$

where we used $h_j \le \lambda_n (m/j)^{\alpha_{\min}}$ for $j > m$ and $\alpha_{\min} > 1$.

Thus, $\boldsymbol{G}_0^{1/2} \left(\sum_i n_i \boldsymbol{C}_i\right) \boldsymbol{G}_0^{1/2}$ has intrinsic dimension $O(m)$, and by applying the intrinsic-dimension matrix Bernstein inequality (Tropp, 2015, Theorem 7.3.1) and recalling $\log m = O(\log N)$,

$$\left\| \sum_{i,r} (\boldsymbol{Z}_{i,r} - \mathbb{E}\boldsymbol{Z}_{i,r}) \right\|_{\mathrm{op}} \le C_D \left(\sqrt{\epsilon} (\log N)^{(c_1+1)/2} + \epsilon (\log N)^{c_1+1} \right),$$

with probability at least $1 - N^{-(D+2)}$.

Next, we control the remaining two terms. A union bound gives

$$\mathbb{P}\left(\boldsymbol{M}_{i,r} \ne \boldsymbol{Z}_{i,r} \text{ for some } i, r \right) \le N^{1-A}.$$

On the complementary event, the decomposition reduces to

$$\boldsymbol{E} = \sum_{i,r} (\boldsymbol{Z}_{i,r} - \mathbb{E}\boldsymbol{Z}_{i,r}) - \sum_{i,r} \mathbb{E}[\boldsymbol{M}_{i,r} - \boldsymbol{Z}_{i,r}].$$

To bound the remaining deterministic term, fix a deterministic vector $\boldsymbol{v}$ and write $q = \boldsymbol{v}^\top \boldsymbol{G}_0^{1/2} \boldsymbol{C}_i \boldsymbol{G}_0^{1/2} \boldsymbol{v}$. Assumption 1 gives

$$\mathbb{P}\left(\left| (\boldsymbol{v}^\top \boldsymbol{G}_0^{1/2} \boldsymbol{x}_{i,r})^2 - q \right| \ge tq \right) \le C \exp\left(-ct^{1/\beta} \right), \qquad t > 0,$$

where $\beta > 0$ is the fixed tail exponent supplied by the assumption. Since $\mathbb{E}[(\boldsymbol{v}^\top \boldsymbol{G}_0^{1/2} \boldsymbol{x}_{i,r})^2] = q$, integrating this tail bound yields

$$\begin{aligned}\mathbb{E}[(\boldsymbol{v}^\top \boldsymbol{G}_0^{1/2} \boldsymbol{x}_{i,r})^4] &= q^2 + \mathbb{E}\left[\left((\boldsymbol{v}^\top \boldsymbol{G}_0^{1/2} \boldsymbol{x}_{i,r})^2 - q \right)^2 \right] \\ &\le q^2 + 2Cq^2 \int_0^\infty t \exp(-ct^\eta)\, dt \\ &\le C' (\boldsymbol{v}^\top \boldsymbol{G}_0^{1/2} \boldsymbol{C}_i \boldsymbol{G}_0^{1/2} \boldsymbol{v})^2.\end{aligned}$$

Cauchy–Schwarz inequality gives

$$\begin{aligned}\boldsymbol{v}^\top \mathbb{E}[\boldsymbol{M}_{i,r} - \boldsymbol{Z}_{i,r}]\boldsymbol{v} &= \mathbb{E}\left[(\boldsymbol{v}^\top \boldsymbol{G}_0^{1/2}\boldsymbol{x}_{i,r})^2 \mathbf{1}_{\{\boldsymbol{x}_{i,r}^\top \boldsymbol{G}_0 \boldsymbol{x}_{i,r} > C_A \epsilon (\log N)^{c_1}\}}\right] \\ &\le \left(\mathbb{E}[(\boldsymbol{v}^\top \boldsymbol{G}_0^{1/2}\boldsymbol{x}_{i,r})^4]\right)^{1/2} \mathbb{P}\left(\boldsymbol{x}_{i,r}^\top \boldsymbol{G}_0 \boldsymbol{x}_{i,r} > C_A \epsilon (\log N)^{c_1}\right)^{1/2} \\ &\le C N^{-A/2} \boldsymbol{v}^\top \boldsymbol{G}_0^{1/2} \boldsymbol{C}_i \boldsymbol{G}_0^{1/2} \boldsymbol{v}.\end{aligned}$$

Since this holds for every $\boldsymbol{v}$, summing over the samples yields

$$\mathbf{0} \preceq \sum_{i,r} \mathbb{E}[\boldsymbol{M}_{i,r} - \boldsymbol{Z}_{i,r}] \preceq C N^{-A/2} \boldsymbol{G}_0^{1/2} \left(\sum_i n_i \boldsymbol{C}_i\right) \boldsymbol{G}_0^{1/2} \preceq C N^{-A/2} \boldsymbol{I}.$$

Combining this estimate with the Bernstein bound, we conclude that

$$\|\boldsymbol{E}\|_{\mathrm{op}} \le C_D \left(\sqrt{\epsilon}(\log N)^{(c_1+1)/2} + \epsilon (\log N)^{c_1+1}\right) + C N^{-A/2},$$

with probability at least $1 - N^{-(D+2)} - N^{1-A}$.

Finally, since

$$N\epsilon \ge \sum_i n_i \mathrm{Tr}(\boldsymbol{C}_i \boldsymbol{G}_0) = \mathrm{Tr}\left[\boldsymbol{G}_0^{1/2}\left(\sum_i n_i \boldsymbol{C}_i\right)\boldsymbol{G}_0^{1/2}\right] \ge \frac{1}{2},$$

we have $N^{-A/2} = o(\sqrt{\epsilon})$. Also, $\epsilon \le n^{-c_-}$ implies $\epsilon \le 1$ for sufficiently large n, so increasing the fixed logarithmic exponent gives

$$\|\boldsymbol{E}\|_{\mathrm{op}} \le C_D (\log N)^c \sqrt{\epsilon}.$$

The same argument applies to

$$\boldsymbol{E}_\sigma = \sum_{i,r} \sigma_{\varepsilon_i}^2 (\boldsymbol{M}_{i,r} - \mathbb{E}\boldsymbol{M}_{i,r}),$$

using the truncated matrices $\sigma_{\varepsilon_i}^2 \boldsymbol{Z}_{i,r}$. Since the noise variances are uniformly bounded, the single-matrix norm bound, the matrix variance bound, and the discarded-mean bound change only by constant factors. Taking a union bound and using $A > D + 4$, we obtain

$$\max\{\|\boldsymbol{E}\|_{\mathrm{op}}, \|\boldsymbol{E}_\sigma\|_{\mathrm{op}}\} \le C_D (\log N)^c \sqrt{\epsilon},$$

with probability at least $1 - N^{-D}$ for sufficiently large n.

Proof of (33)**.** For $\boldsymbol{A}_* = \boldsymbol{\theta}_* \boldsymbol{\theta}_*^\top$, the definition of the deterministic equivalent gives

$$\begin{aligned}\overline{\mathsf{R}}_1(\lambda_n) - (B_0 + V_\sigma) = \left(\lambda_n^2 \tau_{\boldsymbol{A}_*}[1] - B_0\right) + \left(\overline{\mathsf{V}}_1 - V_\sigma\right) \\ + \lambda_n^2 \tau_{\boldsymbol{A}_*}^\top \boldsymbol{L}^{-1} \tau_{\boldsymbol{C}_1}.\end{aligned}$$

We will show that the first two terms are bounded by $C\epsilon B_0$ and $C\epsilon V_\sigma$, respectively, while the last term is bounded by $C\lambda_n \epsilon V_\sigma$.

For the first term, recall that

$$B_0 = \lambda_n^2 \boldsymbol{\theta}_*^\top \boldsymbol{G}_0 \boldsymbol{C}_1 \boldsymbol{G}_0 \boldsymbol{\theta}_*, \quad \tau_{\boldsymbol{A}_*}[1] = \mathrm{Tr}(\boldsymbol{A}_* \overline{\boldsymbol{G}} \boldsymbol{C}_1 \overline{\boldsymbol{G}}) = \boldsymbol{\theta}_*^\top \overline{\boldsymbol{G}} \boldsymbol{C}_1 \overline{\boldsymbol{G}} \boldsymbol{\theta}_*.$$

We first compare $\boldsymbol{G}_0$ and $\boldsymbol{G}$. Let $\eta = \min_{i=1,2} \frac{\mu_i}{n_i} \in (0, 1]$. Since

$$\sum_i \mu_i \boldsymbol{C}_i + \lambda_n \boldsymbol{I} \succeq \eta (n \boldsymbol{C}_1 + n_2 \boldsymbol{C}_2 + \lambda_n \boldsymbol{I}),$$

we have $\overline{\boldsymbol{G}} \preceq \eta^{-1} \boldsymbol{G}_0$. Consequently, the fixed-point equations imply

$$\frac{\mu_i}{n_i} = \frac{1}{1 + \mathrm{Tr}(\boldsymbol{C}_i \overline{\boldsymbol{G}})} \ge \frac{1}{1 + \epsilon/\eta}.$$

Taking the minimum over i gives $\eta \geq \eta/(\eta+\epsilon)$. Since $\eta > 0$, this implies $\eta + \epsilon \geq 1$, and therefore

$$1 - \epsilon \leq \frac{\mu_i}{n_i} \leq 1,$$

which implies

$$\boldsymbol{G}_0 \preceq \overline{\boldsymbol{G}} \preceq (1-\epsilon)^{-1}\boldsymbol{G}_0, \qquad \mathrm{Tr}(\boldsymbol{C}_i\overline{\boldsymbol{G}}) \leq \frac{\epsilon}{1-\epsilon}.$$

Since $\boldsymbol{C}_1$, $\boldsymbol{G}_0$, and $\overline{\boldsymbol{G}}$ are diagonal in the same basis, the comparison $\boldsymbol{G}_0 \preceq \overline{\boldsymbol{G}} \preceq (1-\epsilon)^{-1}\boldsymbol{G}_0$ gives

$$B_0 \leq \lambda_n^2 \tau_{\boldsymbol{A}_*}[1] \leq (1-\epsilon)^{-2} B_0.$$

For sufficiently large n, $\epsilon \leq 1/2$, so

$$0 \leq \lambda_n^2 \tau_{\boldsymbol{A}_*}[1] - B_0 \leq \left((1-\epsilon)^{-2} - 1\right) B_0 = \frac{2\epsilon - \epsilon^2}{(1-\epsilon)^2} B_0 \leq 8\epsilon B_0.$$

To control the variance term, recall that $\boldsymbol{L} = \boldsymbol{D} - \boldsymbol{K}$, where

$$\boldsymbol{D}_{ii} = \frac{n_i}{\mu_i^2}, \qquad \boldsymbol{K}_{ij} = \mathrm{Tr}(\boldsymbol{C}_i\overline{\boldsymbol{G}}\boldsymbol{C}_j\overline{\boldsymbol{G}}).$$

Define

$$\boldsymbol{\sigma} = (\sigma_{\varepsilon_1}^2, \sigma_{\varepsilon_2}^2)^\top, \qquad w_i = \frac{(\boldsymbol{L}^{-1}\boldsymbol{\sigma})_i}{\mu_i}.$$

The equation $\boldsymbol{L}(\boldsymbol{L}^{-1}\boldsymbol{\sigma}) = \boldsymbol{\sigma}$ becomes

$$w_i = \frac{\mu_i}{n_i}\sigma_{\varepsilon_i}^2 + \sum_{j=1}^{2} \frac{\mu_i\mu_j}{n_i} \boldsymbol{K}_{ij} w_j.$$

The coefficients in the sum are nonnegative, and their row sums satisfy

$$\begin{aligned}
\sum_{j=1}^{2} \frac{\mu_i\mu_j}{n_i}\boldsymbol{K}_{ij} &= \frac{\mu_i}{n_i}\mathrm{Tr}\left[\boldsymbol{C}_i\overline{\boldsymbol{G}}\left(\sum_{j=1}^{2}\mu_j\boldsymbol{C}_j\right)\overline{\boldsymbol{G}}\right] \\
&\leq \mathrm{Tr}(\boldsymbol{C}_i\overline{\boldsymbol{G}}) \\
&\leq C\epsilon.
\end{aligned}$$

Here we used $\sum_j \mu_j \boldsymbol{C}_j \preceq \overline{\boldsymbol{G}}^{-1}$ and $\mu_i/n_i \leq 1$. It follows that

$$\max_i |w_i| \leq \max_i \sigma_{\varepsilon_i}^2 + C\epsilon \max_i |w_i|,$$

and hence $\max_i |w_i| = O(1)$. Substituting this bound back into the equation for w_i gives

$$w_i = \frac{\mu_i}{n_i}\sigma_{\varepsilon_i}^2 + O(\epsilon).$$

Multiplying by μ_i, and using $\mu_i/n_i = 1 + O(\epsilon)$, yields

$$(\boldsymbol{L}^{-1}\boldsymbol{\sigma})_i = \frac{\mu_i^2}{n_i}\sigma_{\varepsilon_i}^2 + O(\mu_i\epsilon) = n_i\sigma_{\varepsilon_i}^2(1 + O(\epsilon)).$$

The last equality uses the positive lower bound on the noise variances.

Since $\boldsymbol{G}_0, \boldsymbol{G}$ are both diagonal, we have

$$\tau_{\boldsymbol{C}_1}[i] = (1 + O(\epsilon))\mathrm{Tr}(\boldsymbol{C}_1\boldsymbol{G}_0\boldsymbol{C}_i\boldsymbol{G}_0).$$

Consequently,

$$\begin{aligned}
\overline{\mathsf{V}}_1 &= \tau_{\boldsymbol{C}_1}^\top \boldsymbol{L}^{-1}\boldsymbol{\sigma} \\
&= (1 + O(\epsilon)) \sum_{i=1}^{2} \sigma_{\varepsilon_i}^2 n_i \mathrm{Tr}(\boldsymbol{C}_1\boldsymbol{G}_0\boldsymbol{C}_i\boldsymbol{G}_0) \\
&= (1 + O(\epsilon))V_\sigma.
\end{aligned}$$

Thus $|\overline{\mathsf{V}}_1 - V_\sigma| \le C\epsilon V_\sigma$.

It remains to bound the correction $\lambda_n^2 \tau_{\boldsymbol{A}_*}^\top \boldsymbol{L}^{-1} \tau_{\boldsymbol{C}_1}$. The rescaled linear system for w_i has nonnegative coefficients and row sums smaller than one. Its Neumann series therefore has nonnegative entries. Applying the same argument to any nonnegative right-hand side and undoing the positive diagonal rescaling shows that $\boldsymbol{L}^{-1}$ is entrywise nonnegative.

For each i, we have

$$\begin{aligned}\lambda_n^2 \tau_{\boldsymbol{A}_*}[i] &= \lambda_n^2 \boldsymbol{\theta}_*^\top \overline{\boldsymbol{G}} \boldsymbol{C}_i \overline{\boldsymbol{G}} \boldsymbol{\theta}_* \\ &\le \lambda_n^2 \|\overline{\boldsymbol{G}}\|_{\mathrm{op}} \mathrm{Tr}(\boldsymbol{C}_i \overline{\boldsymbol{G}}) \|\boldsymbol{\theta}_*\|^2 \\ &\le C\lambda_n \epsilon,\end{aligned}$$

where we used $\|\overline{\boldsymbol{G}}\|_{\mathrm{op}} \le \lambda_n^{-1}$ and $\|\boldsymbol{\theta}_*\|^2 < \infty$. Since the noise variances are bounded below, this means that $\lambda_n^2 \tau_{\boldsymbol{A}_*} \le C\lambda_n \epsilon \boldsymbol{\sigma}$ entrywise. By symmetry and entrywise nonnegativity of $\boldsymbol{L}^{-1}$,

$$0 \le \lambda_n^2 \tau_{\boldsymbol{A}_*}^\top \boldsymbol{L}^{-1} \tau_{\boldsymbol{C}_1} \le C\lambda_n \epsilon \boldsymbol{\sigma}^\top \boldsymbol{L}^{-1} \tau_{\boldsymbol{C}_1} = C\lambda_n \epsilon \overline{\mathsf{V}}_1 \le C\lambda_n \epsilon V_\sigma.$$

Combining the three terms in the deterministic risk decomposition proves

$$|\overline{\mathsf{R}}_1(\lambda_n) - (B_0 + V_\sigma)| \le C\epsilon B_0 + C\epsilon V_\sigma + C\lambda_n \epsilon V_\sigma \le C(\epsilon + \lambda_n \epsilon)(B_0 + V_\sigma).$$

This establishes (33) and concludes the proof.

F Proofs of deterministic equivalences

F.1 Heuristic derivation of deterministic equivalences

First of all, it is easy to heuristicly derive the deterministic equivalence for the first order functional $\mathrm{Tr}(\boldsymbol{A}\boldsymbol{G}) \approx \mathrm{Tr}(\boldsymbol{A}\overline{\boldsymbol{G}})$. Next we heuristicly derive the deterministic equivalence for the second order functional of resolvent.

Deterministic equivalence of $\mathrm{Tr}(\boldsymbol{A}\boldsymbol{G}\boldsymbol{B}\boldsymbol{G})$. Let

$$\boldsymbol{H} = \sum_{i=1}^K \boldsymbol{X}_i^\top \boldsymbol{X}_i + \lambda \boldsymbol{I}.$$

For a deterministic PSD matrix $\boldsymbol{B}$,

$$\mathrm{Tr}(\boldsymbol{A}\boldsymbol{G}\boldsymbol{B}\boldsymbol{G}) = -\frac{\mathrm{d}}{\mathrm{d}t} \mathrm{Tr}\left(\boldsymbol{A}(\boldsymbol{H} + t\boldsymbol{B})^{-1}\right)\Big|_{t=0}.$$

The deterministic equivalence of $\mathrm{Tr}(\boldsymbol{A}\boldsymbol{G})$ gives

$$\mathrm{Tr}\left(\boldsymbol{A}(\boldsymbol{H} + t\boldsymbol{B})^{-1}\right) \approx \mathrm{Tr}(\boldsymbol{A}\overline{\boldsymbol{G}}(t)),$$

where

$$\overline{\boldsymbol{G}}(t) = \left(\sum_{i=1}^K \mu_i(t)\boldsymbol{C}_i + \lambda \boldsymbol{I} + t\boldsymbol{B}\right)^{-1}, \qquad \mu_i(t) = \frac{n_i}{1 + \mathrm{Tr}(\boldsymbol{C}_i \overline{\boldsymbol{G}}(t))}.$$

Differentiating the resolvent gives

$$\frac{\mathrm{d}}{\mathrm{d}t}\overline{\boldsymbol{G}}(t) = -\overline{\boldsymbol{G}}(t)\left(\sum_{j=1}^K \frac{\mathrm{d}\mu_j(t)}{\mathrm{d}t}\boldsymbol{C}_j + \boldsymbol{B}\right)\overline{\boldsymbol{G}}(t).$$

Differentiating the fixed-point equations and evaluating at $t = 0$ yields

$$\frac{n_i}{\mu_i^2}\dot{\mu}_i - \sum_{j=1}^K \mathrm{Tr}(\boldsymbol{C}_i \overline{\boldsymbol{G}} \boldsymbol{C}_j \overline{\boldsymbol{G}})\dot{\mu}_j = \mathrm{Tr}(\boldsymbol{C}_i \overline{\boldsymbol{G}} \boldsymbol{B} \overline{\boldsymbol{G}}).$$

Therefore,

$$\boldsymbol{L}\dot{\boldsymbol{\mu}} = \tau_{\boldsymbol{B}}.$$

A final differentiation gives

$$\begin{aligned} -\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{Tr}(\boldsymbol{A}\overline{\boldsymbol{G}}(t))\Big|_{t=0} &= \mathrm{Tr}(\boldsymbol{A}\overline{\boldsymbol{G}}\boldsymbol{B}\overline{\boldsymbol{G}}) + \tau_{\boldsymbol{A}}^\top \dot{\boldsymbol{\mu}} \\ &= \mathrm{Tr}(\boldsymbol{A}\overline{\boldsymbol{G}}\boldsymbol{B}\overline{\boldsymbol{G}}) + \tau_{\boldsymbol{A}}^\top \boldsymbol{L}^{-1}\tau_{\boldsymbol{B}}. \end{aligned}$$

This proves the first second-order formula in Theorem 2.

Deterministic equivalence of $\mathrm{Tr}(\boldsymbol{A}\boldsymbol{G}\boldsymbol{X}_k^\top\boldsymbol{X}_k\boldsymbol{G})$. Set $\boldsymbol{S}_k = \boldsymbol{X}_k^\top\boldsymbol{X}_k$. Then

$$\mathrm{Tr}(\boldsymbol{A}\boldsymbol{G}\boldsymbol{S}_k\boldsymbol{G}) = -\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{Tr}\left(\boldsymbol{A}(\boldsymbol{H}+t\boldsymbol{S}_k)^{-1}\right)\Big|_{t=0}.$$

Let $\beta_i(t) = 1 + t\mathbf{1}[i=k]$ and $r_i(t) = \beta_i(t)\mu_i(t)$. The first-order deterministic equivalent is

$$\overline{\boldsymbol{G}}_k(t) = \left(\sum_{i=1}^K r_i(t)\boldsymbol{C}_i + \lambda\boldsymbol{I}\right)^{-1}, \qquad \mu_i(t) = \frac{n_i}{1+\beta_i(t)\mathrm{Tr}(\boldsymbol{C}_i\overline{\boldsymbol{G}}_k(t))}.$$

Equivalently,

$$\frac{n_i}{r_i(t)} = \frac{1}{\beta_i(t)} + \mathrm{Tr}(\boldsymbol{C}_i\overline{\boldsymbol{G}}_k(t)).$$

Differentiating at $t=0$ gives

$$\frac{n_i}{\mu_i^2}\dot{r}_i - \sum_{j=1}^K \mathrm{Tr}(\boldsymbol{C}_i\overline{\boldsymbol{G}}\boldsymbol{C}_j\overline{\boldsymbol{G}})\dot{r}_j = \mathbf{1}[i=k],$$

and hence

$$\boldsymbol{L}\dot{\boldsymbol{r}} = \boldsymbol{e}_k.$$

Since

$$-\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{Tr}(\boldsymbol{A}\overline{\boldsymbol{G}}_k(t))\Big|_{t=0} = \tau_{\boldsymbol{A}}^\top\dot{\boldsymbol{r}},$$

we obtain

$$\mathrm{Tr}(\boldsymbol{A}\boldsymbol{G}\boldsymbol{S}_k\boldsymbol{G}) \approx \tau_{\boldsymbol{A}}^\top\boldsymbol{L}^{-1}\boldsymbol{e}_k.$$

Bias and variance. We apply the first second-order formula with the rank-one matrix $\boldsymbol{A} = \boldsymbol{\theta}_*\boldsymbol{\theta}_*^\top$ and $\boldsymbol{B} = \boldsymbol{C}_k$. Multiplication by λ^2 gives

$$\overline{\mathsf{B}}_k = \lambda^2\left[\langle\boldsymbol{\theta}_*, \overline{\boldsymbol{G}}\boldsymbol{C}_k\overline{\boldsymbol{G}}\boldsymbol{\theta}_*\rangle + \tau_{\boldsymbol{\theta}_*}^\top\boldsymbol{L}^{-1}\tau_{\boldsymbol{C}_k}\right].$$

Applying the second formula with $\boldsymbol{A} = \boldsymbol{C}_k$ and summing the independent noise contributions gives

$$\overline{\mathsf{V}}_k = \sum_{i=1}^K \sigma_{\varepsilon_i}^2\,\tau_{\boldsymbol{C}_k}^\top\boldsymbol{L}^{-1}\boldsymbol{e}_i.$$

Weighting the domain-specific risks by π_k^* gives the main results in Theorem 2.

F.2 Main theorem and proof sketch

We first present the main theorem and the proof sketch, and then specify the proof of each deterministic equivalence in the following subsections.

Before stating the main results, we first recall some notation below.

Let $\boldsymbol{X}_i = [\boldsymbol{x}_{i,1}^\top, \ldots, \boldsymbol{x}_{i,n_i}^\top]^\top \in \mathbb{R}^{n_i\times d}$ ($i\in[K]$) be K datasets, where $\mathbb{E}[\boldsymbol{x}_{i,r}\boldsymbol{x}_{i,r}^\top] = \boldsymbol{C}_i$. Define $\boldsymbol{G} = (\sum_{i=1}^K \boldsymbol{X}_i^\top\boldsymbol{X}_i + \lambda)^{-1}$, and $\overline{\boldsymbol{G}} = (\sum_{i=1}^K \mu_i\boldsymbol{C}_i + \lambda)^{-1}$. Here, μ_i is the fixed point of the self-consistence equation

$$\frac{n_i}{\mu_i} = 1 + \mathrm{Tr}(\boldsymbol{C}_i\overline{\boldsymbol{G}}), \quad i\in[K].$$

For later convenience, we define the following matrices that relate to the deterministic equivalence and the stability of the fixed point. Define a $K \times K$ diagonal matrix $\boldsymbol{D}$ with $\boldsymbol{D}_{ii} = \frac{n_i}{\mu_i^2}$, a $K \times K$ diagonal matrix $\boldsymbol{K}$ with $\boldsymbol{K}_{ij} = \mathrm{Tr}(\boldsymbol{C}_i\overline{\boldsymbol{G}}\boldsymbol{C}_j\overline{\boldsymbol{G}})$, and $\boldsymbol{L} = \boldsymbol{D} - \boldsymbol{K}$. Now, we state the main theorem that provides the deterministic equivalence of the three functionals appearing in Theorem 2.

Theorem 6. *Assume that, for $i \in [K]$, $\boldsymbol{C}_i$ satisfies Assumption 1 and these matrices commute. For each $\boldsymbol{C}_i$, following (Misiakiewicz & Saeed, 2024, Appendix A.1), we define*

$$\begin{aligned} r_{\boldsymbol{C}_i}(n_i) &= \max\left\{n_i, \max_{0\le m<\min(n_i,p_i)} \frac{\sum_{j>m}\xi_{i,j}}{\xi_{i,m+1}}\right\}, \\ \nu_\lambda^i(n_i) &:= 1 + \frac{\xi_{\lfloor \eta n_i\rfloor} r_{\boldsymbol{C}_i}(n_i)\sqrt{\log r_{\boldsymbol{C}_i}(n_i)}}{\lambda}, \qquad \nu := \max_{i\in[K]} \nu_\lambda^i(n_i). \end{aligned} \tag{34}$$

We further assume $\nu_\lambda^i(n_i)$ satisfies Assumption 2 for all $i \in [K]$. For any PSD matrix $\boldsymbol{A}$, define $\tau_A \in \mathbb{R}^K$ with $\tau_A[k] = \mathrm{Tr}(\boldsymbol{A}\overline{\boldsymbol{G}}\boldsymbol{C}_k\overline{\boldsymbol{G}})$. Then, we have with probability at least $1 - \sum_{i=1}^K n_i^{-D}$,

$$\begin{aligned} |\mathrm{Tr}(\boldsymbol{A}\boldsymbol{G}) - \mathrm{Tr}(\boldsymbol{A}\overline{\boldsymbol{G}})| &\le C_D\varepsilon_1 \mathrm{Tr}(\boldsymbol{A}\overline{\boldsymbol{G}}), \\ |\mathrm{Tr}(\boldsymbol{A}\boldsymbol{G}\boldsymbol{C}_k\boldsymbol{G}) - \boldsymbol{e}_k^\top \boldsymbol{D}\boldsymbol{L}^{-1}\tau_{\boldsymbol{A}}| &\le C_D\varepsilon_2 V_{\boldsymbol{A}}, \\ |\mathrm{Tr}(\boldsymbol{A}\boldsymbol{G}\boldsymbol{X}_k^\top\boldsymbol{X}_k\boldsymbol{G}) - \boldsymbol{e}_k^\top \boldsymbol{L}^{-1}\tau_{\boldsymbol{A}}| &\le C_D\varepsilon_3 V_{\boldsymbol{A}}, \end{aligned}$$

where

$$\begin{aligned} \varepsilon_1 &= e_K\left(\nu^6 + \nu^2\log^{\beta+1/2}(N)\right), \\ \varepsilon_2 &= e_K\left(\frac{\nu^{14}}{n_k} + \nu^8 e_K^2\log^{4\beta+3/2}(N)\right), \\ \varepsilon_3 &= e_K\left(\nu^{14} + \nu^5\log^{3\beta+3/2}(N)\right), \\ V_{\boldsymbol{A}} &= \sum_{i=1}^K n_i \mathrm{Tr}(\boldsymbol{A}\overline{\boldsymbol{G}}\boldsymbol{C}_i\overline{\boldsymbol{G}}), \qquad e_K = \left(\sum_{i=1}^K n_i^{-1}\right)^{1/2}. \end{aligned}$$

Connections to Misiakiewicz & Saeed (2024). (Misiakiewicz & Saeed, 2024, Appendix A) shows that all three functionals $\mathrm{Tr}(\boldsymbol{A}\boldsymbol{M})$, $\mathrm{Tr}(\boldsymbol{A}\boldsymbol{M}^2)$, and $n_k^{-1}\mathrm{Tr}(\boldsymbol{A}\boldsymbol{M}\boldsymbol{X}_k^\top\boldsymbol{X}_k\boldsymbol{M})$ enjoy multiplicative error bound with the coefficient scales as $O(n^{-1/2})$. In our case, under fixed K and comparable dataset sizes $n_i \asymp n$, all three ε_1, ε_2, and ε_3 also scale as $O(n^{-1/2})$.

To prove Theorem 6, we decompose the error into the deterministic bias part and the martingale part. For example, for $\mathrm{Tr}(\boldsymbol{A}\boldsymbol{G})$, we have:

$$|\mathrm{Tr}(\boldsymbol{A}\boldsymbol{G}) - \mathrm{Tr}(\boldsymbol{A}\overline{\boldsymbol{G}})| \le \underbrace{|\mathrm{Tr}(\boldsymbol{A}\boldsymbol{G}) - \mathbb{E}[\mathrm{Tr}(\boldsymbol{A}\boldsymbol{G})]|}_{\text{deterministic part}} + \underbrace{|\mathbb{E}[\mathrm{Tr}(\boldsymbol{A}\boldsymbol{G})] - \mathrm{Tr}(\boldsymbol{A}\overline{\boldsymbol{G}})|}_{\text{martingale part}}.$$

F.3 Deterministic equivalence of $\mathrm{Tr}(\boldsymbol{A}\boldsymbol{G})$

F.3.1 Deterministic part of $\mathrm{Tr}(\boldsymbol{A}\boldsymbol{G})$

Write

$$e_K = \left(\sum_{i=1}^K \frac{1}{n_i}\right)^{1/2}, \qquad \nu = \max_i \nu_\lambda^i(n).$$

Without loss of generality, we may assume $e_K \le 1$. By the deterministic bounds in Lemma 5, we have

$$t_i := \mathrm{Tr}(\boldsymbol{C}_i\overline{\boldsymbol{G}}) \le C\nu.$$

As in (Misiakiewicz & Saeed, 2024, Appendix A.3.1), we may assume $\boldsymbol{A} = \boldsymbol{\theta}\boldsymbol{\theta}^\top$. For dataset $\boldsymbol{X}_i$, define

$$\boldsymbol{H}_i = \boldsymbol{G}_{-(i,1)}, \qquad \kappa_i = \mathbb{E}[\mathrm{Tr}(\boldsymbol{C}_i\boldsymbol{H}_i)], \qquad \overline{\boldsymbol{G}}_- = \left(\sum_{\ell=1}^K \frac{n_\ell}{1+\kappa_\ell}\boldsymbol{C}_\ell + \lambda\right)^{-1},$$

where $\boldsymbol{G}_{-(i,1)}$ is obtained from $\boldsymbol{G}$ by leave-one-out on the i-th data point. We will bound $|\mathbb{E}[\mathrm{Tr}(\boldsymbol{A}\boldsymbol{G}) - \mathrm{Tr}(\boldsymbol{A}\overline{\boldsymbol{G}})|$ by splitting it into

$$|\mathbb{E}[\mathrm{Tr}(\boldsymbol{A}\boldsymbol{G}) - \mathrm{Tr}(\boldsymbol{A}\overline{\boldsymbol{G}})| \le |\mathrm{Tr}(\boldsymbol{A}(\mathbb{E}[\boldsymbol{G}] - \overline{\boldsymbol{G}}_-))| + |\mathrm{Tr}(\boldsymbol{A}(\overline{\boldsymbol{G}}_- - \overline{\boldsymbol{G}}))|. \tag{35}$$

For the first term in (35),

$$\mathbb{E}[\boldsymbol{G}] - \overline{\boldsymbol{G}}_- = \sum_{i=1}^{K} n_i \mathbb{E}\left[\boldsymbol{G}\left(\frac{\boldsymbol{C}_i}{1+\kappa_i} - \boldsymbol{x}\boldsymbol{x}^\top\right)\overline{\boldsymbol{G}}_-\right],$$

with $\boldsymbol{x} = \boldsymbol{x}_{i,1}$. Setting $z_i = \boldsymbol{x}^\top \boldsymbol{H}_i \boldsymbol{x}$ and following the approach in (Misiakiewicz & Saeed, 2024, Appendix A.3.1), we decompose

$$\boldsymbol{G}\left(\frac{\boldsymbol{C}_i}{1+\kappa_i} - \boldsymbol{x}\boldsymbol{x}^\top\right)\overline{\boldsymbol{G}}_- = \Delta_{i,1} + \Delta_{i,2} + \Delta_{i,3},$$

with

$$\Delta_{i,1} = \frac{\boldsymbol{H}_i(\boldsymbol{C}_i - \boldsymbol{x}\boldsymbol{x}^\top)\overline{\boldsymbol{G}}_-}{1+\kappa_i},$$
$$\Delta_{i,2} = \frac{\boldsymbol{H}_i\boldsymbol{x}\boldsymbol{x}^\top\overline{\boldsymbol{G}}_-(z_i - \kappa_i)}{(1+\kappa_i)(1+z_i)},$$
$$\Delta_{i,3} = -\frac{\boldsymbol{H}_i\boldsymbol{x}\boldsymbol{x}^\top\boldsymbol{H}_i\boldsymbol{C}_i\overline{\boldsymbol{G}}_-}{1+z_i}.$$

The independence between $\boldsymbol{x}$ and $\boldsymbol{H}_i$ implies $\mathbb{E}[\Delta_{i,1}] = 0$. Write $z_i - \kappa_i$ as

$$z_i - \kappa_i = (z_i - \mathrm{Tr}(\boldsymbol{C}_i\boldsymbol{H}_i)) + (\mathrm{Tr}(\boldsymbol{C}_i\boldsymbol{H}_i) - \mathbb{E}[\mathrm{Tr}(\boldsymbol{C}_i\boldsymbol{H}_i)]).$$

We control the first term by applying Lemma 4 conditionally on $\boldsymbol{H}_i$. For the second term, we use the first-order martingale estimation established in the next section. Altogether, we get

$$\|z_i - \kappa_i\|_{L^q} \le C_q\left(\frac{\nu}{\sqrt{n_i}} + \nu^2 e_K t_i\right) \le C_q \nu^3 e_K.$$

Using

$$\overline{\boldsymbol{G}}_- \preceq C\nu\overline{\boldsymbol{G}}, \qquad \overline{\boldsymbol{G}}_-\boldsymbol{C}_i\overline{\boldsymbol{G}}_- \preceq C\frac{\nu^2}{n_i}\overline{\boldsymbol{G}},$$

Lemma 4 to the squared linear forms, the basic deleted-resolvent moment bounds in Lemma 5, and Hölder's inequality, we get

$$\|\boldsymbol{\theta}^\top\boldsymbol{H}_i\boldsymbol{x}\|_{L^q} \le C_q\frac{\nu}{\sqrt{n_i}}\sqrt{\mathrm{Tr}(\boldsymbol{A}\overline{\boldsymbol{G}})},$$
$$\|\boldsymbol{x}^\top\overline{\boldsymbol{G}}_-\boldsymbol{\theta}\|_{L^q} \le C_q\frac{\nu}{\sqrt{n_i}}\sqrt{\mathrm{Tr}(\boldsymbol{A}\overline{\boldsymbol{G}})},$$
$$\|\boldsymbol{x}^\top\boldsymbol{H}_i\boldsymbol{C}_i\overline{\boldsymbol{G}}_-\boldsymbol{\theta}\|_{L^q} \le C_q\frac{\nu^2}{n_i^{3/2}}\sqrt{\mathrm{Tr}(\boldsymbol{A}\overline{\boldsymbol{G}})}.$$

By Hölder,

$$n_i|\mathbb{E}[\mathrm{Tr}(\boldsymbol{A}\Delta_{i,2})]| \le Cn_i\frac{\nu}{\sqrt{n_i}}\frac{\nu}{\sqrt{n_i}}(\nu^3 e_K)\mathrm{Tr}(\boldsymbol{A}\overline{\boldsymbol{G}}) = C\nu^5 e_K\mathrm{Tr}(\boldsymbol{A}\overline{\boldsymbol{G}}),$$
$$n_i|\mathbb{E}[\mathrm{Tr}(\boldsymbol{A}\Delta_{i,3})]| \le Cn_i\frac{\nu}{\sqrt{n_i}}\frac{\nu^2}{n_i^{3/2}}\mathrm{Tr}(\boldsymbol{A}\overline{\boldsymbol{G}}) = C\frac{\nu^3}{n_i}\mathrm{Tr}(\boldsymbol{A}\overline{\boldsymbol{G}}).$$

Combining the previous results, we get, for the first term in (35),

$$|\mathbb{E}[\mathrm{Tr}(\boldsymbol{A}(\mathbb{E}[\boldsymbol{G}] - \overline{\boldsymbol{G}}_-))] \le C(\nu^5 e_K + \nu^3 e_K^2)\mathrm{Tr}(\boldsymbol{A}\overline{\boldsymbol{G}}) \le C\nu^5 e_K\mathrm{Tr}(\boldsymbol{A}\overline{\boldsymbol{G}}).$$

Now we need to bound the second term in (35). Defining

$$\varepsilon = \max_i \frac{|\kappa_i - t_i|}{1+t_i},$$

we have

$$\begin{aligned}|\mathrm{Tr}(\boldsymbol{A}(\overline{\boldsymbol{G}}_- - \overline{\boldsymbol{G}}))| &\le \varepsilon \sum_{i=1}^{K} \frac{n_i}{1+\kappa_i} \mathrm{Tr}(\boldsymbol{A}\overline{\boldsymbol{G}}_- \boldsymbol{C}_i \overline{\boldsymbol{G}}) \\ &= \varepsilon \mathrm{Tr}(\boldsymbol{A}(\boldsymbol{I} - \lambda \overline{\boldsymbol{G}}_-)\overline{\boldsymbol{G}}) \\ &\le \varepsilon \mathrm{Tr}(\boldsymbol{A}\overline{\boldsymbol{G}}).\end{aligned}$$

Now we need to upper bound ε. Using the one-observation restoration estimate established in Lemma 5, together with the preceding bounds, we obtain

$$\begin{aligned}|\kappa_i - t_i| &\le |\mathbb{E}[\mathrm{Tr}(\boldsymbol{C}_i(\boldsymbol{H}_i - \boldsymbol{G}))]| + |\mathrm{Tr}(\boldsymbol{C}_i(\mathbb{E}[\boldsymbol{G}] - \overline{\boldsymbol{G}}_-))| + |\mathrm{Tr}(\boldsymbol{C}_i(\overline{\boldsymbol{G}}_- - \overline{\boldsymbol{G}}))| \\ &\le C\frac{\nu^2}{n_i} t_i + C\nu^5 e_K t_i + \varepsilon t_i.\end{aligned}$$

By definition, there is $i \in [K]$ such that $|\kappa_i - t_i| = \varepsilon(1+t_i)$. Substituting it to the previous inequality yields

$$\varepsilon \le C\left(\frac{\nu^2}{n_i} + \nu^5 e_K\right) t_i \le C\nu^6 e_K.$$

Therefore, for the second term on (35),

$$|\mathrm{Tr}(\boldsymbol{A}(\overline{\boldsymbol{G}}_- - \overline{\boldsymbol{G}}))| \le C\nu^6 e_K \mathrm{Tr}(\boldsymbol{A}\overline{\boldsymbol{G}}).$$

Combining these two terms,

$$|\mathbb{E}[\mathrm{Tr}(\boldsymbol{A}\boldsymbol{G})] - \mathrm{Tr}(\boldsymbol{A}\overline{\boldsymbol{G}})| \le C\nu^6 e_K \mathrm{Tr}(\boldsymbol{A}\overline{\boldsymbol{G}}).$$

F.3.2 Martingale part of $\mathrm{Tr}(\boldsymbol{A}\boldsymbol{G})$

Here we enumerate all $N = \sum_{i=1}^{K} n_i$ observations. Define $\mathcal{F}_\ell$ be the first ℓ observations, and write $\mathbb{E}_\ell[\cdot] = \mathbb{E}[\cdot|\mathcal{F}_\ell]$. Assume the ℓ-th feature is $\boldsymbol{X}_{i,r}$. Defining $\boldsymbol{H}_\ell = \boldsymbol{G}_{-(i,r)}$, we have

$$\mathrm{Tr}(\boldsymbol{A}\boldsymbol{G}) - \mathbb{E}[\mathrm{Tr}(\boldsymbol{A}\boldsymbol{G})] = \sum_{\ell=1}^{N} (\mathbb{E}_\ell - \mathbb{E}_{\ell-1})(\mathrm{Tr}(\boldsymbol{A}\boldsymbol{G}) - \mathrm{Tr}(\boldsymbol{A}\boldsymbol{H}_\ell)) =: \sum_{\ell=1}^{N} (\mathbb{E}_\ell - \mathbb{E}_{\ell-1})\xi_\ell.$$

Sherman-Morrison gives

$$\xi_\ell = \mathrm{Tr}(\boldsymbol{A}(\boldsymbol{G} - \boldsymbol{H}_\ell)) = -\frac{\boldsymbol{x}_\ell^\top \boldsymbol{H}_\ell \boldsymbol{A} \boldsymbol{H}_\ell \boldsymbol{x}_\ell}{1 + \boldsymbol{x}_\ell^\top \boldsymbol{H}_\ell \boldsymbol{x}_\ell}.$$

Applying Lemma 4 conditionally on $\boldsymbol{H}_i$, together with the basic deleted-resolvent moment bounds in Lemma 5, gives

$$\|\xi_\ell\|_{L^q} \le C_q \frac{\nu^2}{n_i} \mathrm{Tr}(\boldsymbol{A}\overline{\boldsymbol{G}}).$$

Therefore,

$$\|\mathrm{Tr}(\boldsymbol{A}\boldsymbol{G}) - \mathbb{E}[\mathrm{Tr}(\boldsymbol{A}\boldsymbol{G})]\|_{L^q} \le C_q \left(\sum_{j=1}^{K} n_j \frac{\nu^4}{n_j^2}\right)^{1/2} \mathrm{Tr}(\boldsymbol{A}\overline{\boldsymbol{G}}) = C_q \nu^2 e_K \mathrm{Tr}(\boldsymbol{A}\overline{\boldsymbol{G}}).$$

The same argument applies when one observation is already removed, since the resulting row increments involve resolvents with at most two observations removed. For the probability bound, we use the same approach as in (Misiakiewicz & Saeed, 2024, Appendix A.3.2). Set $n_{\min} = \min_i n_i$. The conditional concentration bound in Lemma 4 and the probabilistic bound of the basic deleted-resolvent give

$$\begin{aligned}\boldsymbol{x}_\ell^\top \boldsymbol{H}_\ell \boldsymbol{A} \boldsymbol{H}_\ell \boldsymbol{x}_\ell &\le C_M \log^\eta(N) \mathrm{Tr}(\boldsymbol{A}\boldsymbol{H}_\ell \boldsymbol{C}_i \boldsymbol{H}_\ell) \\ &\le C_M \frac{\nu^2}{n_i} \log^\eta(N) \mathrm{Tr}(\boldsymbol{A}\overline{\boldsymbol{G}}),\end{aligned}$$

except with probability at most $C_M n_{\min}^{-M}$. Here we used

$$\mathrm{Tr}(\boldsymbol{A}\boldsymbol{H}_\ell\boldsymbol{C}_i\boldsymbol{H}_\ell) \le \|\boldsymbol{C}_i^{1/2}\boldsymbol{H}_\ell\boldsymbol{C}_i^{1/2}\|_{\mathrm{op}}\|\overline{\boldsymbol{G}}^{-1/2}\boldsymbol{H}_\ell\overline{\boldsymbol{G}}^{-1/2}\|_{\mathrm{op}}\mathrm{Tr}(\boldsymbol{A}\overline{\boldsymbol{G}}).$$

Consequently,

$$\mathbb{P}(|\xi_\ell| > R_i) \le C_M n_{\min}^{-M}, \qquad R_i := C_M \frac{\nu^2}{n_i}\log^\beta(N)\mathrm{Tr}(\boldsymbol{A}\overline{\boldsymbol{G}}).$$

Define

$$\widetilde{\xi}_\ell = \xi_\ell \mathbf{1}_{\{|\xi_\ell|\le R_i\}}, \qquad \widetilde{d}_\ell = (\mathbb{E}_\ell - \mathbb{E}_{\ell-1})\widetilde{\xi}_\ell, \qquad d_\ell = (\mathbb{E}_\ell - \mathbb{E}_{\ell-1})\xi_\ell.$$

Then $|\widetilde{d}_\ell| \le 2R_i$ almost surely, and Cauchy–Schwarz gives

$$\mathbb{E}[|\xi_\ell|\mathbf{1}_{\{|\xi_\ell|>R_i\}}] \le C_M \frac{\nu^2}{n_i} n_{\min}^{-M/2}\mathrm{Tr}(\boldsymbol{A}\overline{\boldsymbol{G}}).$$

Summing and applying Markov's inequality, using $e_K \ge n_{\min}^{-1/2}$, for sufficiently large M and $n_{\min}$, with probability at least $1 - \frac{1}{2}\sum_i n_i^{-D}$,

$$\left|\sum_{\ell=1}^N (d_\ell - \widetilde{d}_\ell)\right| \le C\nu^2 e_K \mathrm{Tr}(\boldsymbol{A}\overline{\boldsymbol{G}}).$$

By Azuma–Hoeffding, with probability at least $1 - \frac{1}{2}\sum_i n_i^{-D}$,

$$\left|\sum_{\ell=1}^N \widetilde{d}_\ell\right| \le C_D\sqrt{\log(N)\sum_i n_i R_i^2} \le C_D\nu^2 e_K \log^{\eta+1/2}(N)\mathrm{Tr}(\boldsymbol{A}\overline{\boldsymbol{G}}).$$

Combining with the deterministic bound, with probability at least $1 - \sum_i n_i^{-D}$,

$$|\mathrm{Tr}(\boldsymbol{A}\boldsymbol{G}) - \mathrm{Tr}(\boldsymbol{A}\overline{\boldsymbol{G}})| \le C_D e_K\left(\nu^6 + \nu^2\log^{\eta+1/2}(N)\right)\mathrm{Tr}(\boldsymbol{A}\overline{\boldsymbol{G}}).$$

F.4 Deterministic equivalence of $\mathrm{Tr}(\boldsymbol{A}\boldsymbol{G}\boldsymbol{C}_k\boldsymbol{G})$

We follow the same idea as in (Misiakiewicz & Saeed, 2024, Appendix A.5). First, we split the term again in deterministic part and martingale part.

F.4.1 Deterministic part of $\mathrm{Tr}(\boldsymbol{A}\boldsymbol{G}\boldsymbol{C}_k\boldsymbol{G})$

Denote $\kappa_i = \mathbb{E}[\mathrm{Tr}(\boldsymbol{C}_i\boldsymbol{G}_{-(i,1)})]$. For the row $\boldsymbol{x} = \boldsymbol{x}_{i,r}$ under consideration, set

$$\boldsymbol{H} = \boldsymbol{G}_{-(i,r)}, \qquad q_i = \mathrm{Tr}(\boldsymbol{A}\overline{\boldsymbol{G}}_-\boldsymbol{C}_i\overline{\boldsymbol{G}}_-), \qquad s_i = 1 + \kappa_i,$$

and

$$c_k = \boldsymbol{x}^\top\boldsymbol{H}\boldsymbol{C}_k\boldsymbol{H}\boldsymbol{C}_i\overline{\boldsymbol{G}}_-\boldsymbol{\theta}.$$

By the first-order deterministic approximation and the one-observation restoration estimate established in the previous subsections, we have

$$\delta := \max_i \frac{|\kappa_i - t_i|}{1 + t_i} \le C\nu^6 e_K.$$

By the resolvent identity and commutativity,

$$-\delta\overline{\boldsymbol{G}} \preceq \overline{\boldsymbol{G}}_- - \overline{\boldsymbol{G}} \preceq \delta\overline{\boldsymbol{G}}.$$

Moreover, since $t_i \le C\nu$ and $\kappa_i \ge 0$,

$$\overline{\boldsymbol{G}}_- \preceq C\nu\overline{\boldsymbol{G}}.$$

Consequently,

$$q_i \le C\nu^2\tau_{\boldsymbol{A}}[i], \qquad |q_i - \tau_{\boldsymbol{A}}[i]| \le C\nu\delta\,\tau_{\boldsymbol{A}}[i].$$

Setting

$$T_{\boldsymbol{A},k} = \mathbb{E}[\mathrm{Tr}(\boldsymbol{A}\boldsymbol{G}\boldsymbol{C}_k\boldsymbol{G})],$$

we can decompose $T_{\boldsymbol{A},k}$ into

$$T_{\boldsymbol{A},k} = \mathrm{Tr}(\boldsymbol{A}\overline{\boldsymbol{G}}_{-}\boldsymbol{C}_k\overline{\boldsymbol{G}}_{-}) + \boldsymbol{\theta}^\top(\mathbb{E}[\boldsymbol{G}] - \overline{\boldsymbol{G}}_{-})\boldsymbol{C}_k\overline{\boldsymbol{G}}_{-}\boldsymbol{\theta} + \sum_{i=1}^{K} n_i \mathbb{E}\left[\boldsymbol{\theta}^\top\overline{\boldsymbol{G}}_{-}\left(\frac{\boldsymbol{C}_i}{1+\kappa_i} - \boldsymbol{x}\boldsymbol{x}^\top\right)\boldsymbol{G}\boldsymbol{C}_k\boldsymbol{G}\boldsymbol{\theta}\right],$$

where $\boldsymbol{x}$ has the same distribution as the features in $\boldsymbol{X}_i$. Defining

$$\boldsymbol{H} = \boldsymbol{G}_{-(i,r)}, \qquad z = \boldsymbol{x}^\top\boldsymbol{H}\boldsymbol{x}, \qquad f = 1 + z,$$

we write

$$\left(\frac{\boldsymbol{C}_i}{1+\kappa_i} - \boldsymbol{x}\boldsymbol{x}^\top\right)\boldsymbol{G} = \frac{(\boldsymbol{C}_i - \boldsymbol{x}\boldsymbol{x}^\top)\boldsymbol{H}}{1+\kappa_i} + \frac{\boldsymbol{x}\boldsymbol{x}^\top\boldsymbol{H}(z-\kappa_i)}{(1+\kappa_i)f} - \frac{\boldsymbol{C}_i\boldsymbol{H}\boldsymbol{x}\boldsymbol{x}^\top\boldsymbol{H}}{(1+\kappa_i)f},$$

where we imitate the idea of decomposing the deterministic part into $\Delta_1, \Delta_2, \Delta_3$ as in (Misiakiewicz & Saeed, 2024). Furthermore, defining

$$a = \boldsymbol{\theta}^\top\boldsymbol{H}\boldsymbol{x}, \quad b = \boldsymbol{x}^\top\overline{\boldsymbol{G}}_{-}\boldsymbol{\theta}, \quad w = \boldsymbol{x}^\top\boldsymbol{H}\boldsymbol{C}_k\boldsymbol{H}\boldsymbol{x}, \quad c = \boldsymbol{x}^\top\boldsymbol{H}\boldsymbol{C}_i\overline{\boldsymbol{G}}_{-}\boldsymbol{\theta},$$
$$a_k = \boldsymbol{x}^\top\boldsymbol{H}\boldsymbol{C}_k\boldsymbol{H}\boldsymbol{\theta}, \quad c_k = \boldsymbol{x}^\top\boldsymbol{H}\boldsymbol{C}_k\boldsymbol{H}\boldsymbol{C}_i\overline{\boldsymbol{G}}_{-}\boldsymbol{\theta},$$

we have

$$\mathbb{E}\left[\boldsymbol{\theta}^\top\overline{\boldsymbol{G}}_{-}\left(\frac{\boldsymbol{C}_i}{1+\kappa_i} - \boldsymbol{x}\boldsymbol{x}^\top\right)\boldsymbol{G}\boldsymbol{C}_k\boldsymbol{G}\boldsymbol{\theta}\right] = \underbrace{\mathbb{E}\left[\frac{abw}{f^2}\right]}_{\text{leading term}} + \underbrace{\mathbb{E}\left[\frac{(z-\kappa_i)ba_k}{(1+\kappa_i)f} - \frac{ca_k + ac_k}{(1+\kappa_i)f} + \frac{acw}{s_i f^2}\right]}_{\text{remainder term}}.$$

Therefore, next we want to give a deterministic approximation of the leading term, while proving the remainder terms are all small.

Bounding the leading term. Setting $h_i^- = \mathbb{E}[\mathrm{Tr}(\boldsymbol{C}_i\boldsymbol{H}\boldsymbol{C}_k\boldsymbol{H})]$, we claim

$$\mathbb{E}\left[\frac{abw}{f^2}\right] \approx \frac{q_i h_i^-}{(1+\kappa_i)^2}.$$

To establish this result, we first show that we can replace the denominator from f^2 to $(1+\kappa_i)^2$. Recall that

$$\|z - \kappa_i\|_{L^p} \le C_p\delta_i, \qquad \delta_i = C\left(\nu\sqrt{\frac{t_i}{n_i}} + \nu^2 e_K t_i\right) \le C\nu^3 e_K.$$

Since $f, 1+\kappa_i \ge 1$, we can bound $f^{-2} - (1+\kappa_i)^{-2}$ as

$$|f^{-2} - (1+\kappa_i)^{-2}| \le 2|z - \kappa_i|.$$

Applying Lemma 4 conditionally on $\boldsymbol{H}$, followed by Lemma 6, gives

$$\|a\|_{L^p} \le C_p\sqrt{\mathcal{Q}_{\boldsymbol{A}}[i]}, \qquad \|b\|_{L^p} \le C_p\sqrt{q_i} \le C_p\nu\sqrt{\tau_{\boldsymbol{A}}[i]},$$

where

$$\mathcal{Q}_{\boldsymbol{A}}[i] = \tau_{\boldsymbol{A}}[i] + \nu^3 m_i W_{\boldsymbol{A}}, \qquad W_{\boldsymbol{A}} = \sum_{j=1}^{K} n_j t_j \tau_{\boldsymbol{A}}[j].$$

The same conditional moment estimate and the basic deleted-resolvent bounds in Lemma 5 also give

$$\begin{aligned}\|w\|_{L^p} &\le C_p\|\mathrm{Tr}(\boldsymbol{C}_i\boldsymbol{H}\boldsymbol{C}_k\boldsymbol{H})\|_{L^p} \\ &\le C_p\left\|\|\boldsymbol{C}_k^{1/2}\boldsymbol{H}\boldsymbol{C}_k^{1/2}\|_{\mathrm{op}}\right\|_{L^{2p}}\|\mathrm{Tr}(\boldsymbol{C}_i\boldsymbol{H})\|_{L^{2p}} \le C_p\frac{\nu^2}{n_k}.\end{aligned}$$

In particular, $h_i^- \le C\nu^2/n_k$. Hölder then implies

$$\left|\mathbb{E}[abw(f^{-2} - (1+\kappa_i)^{-2})]\right| \le C\sqrt{\tau_{\boldsymbol{A}}[i]\mathcal{Q}_{\boldsymbol{A}}[i]}\,\delta_i\frac{\nu^3}{n_k},$$

Then, we show that we can replace w by h_i^-. Decompose

$$w - h_i^- = (w - \mathrm{Tr}(\boldsymbol{C}_i\boldsymbol{H}\boldsymbol{C}_k\boldsymbol{H})) + (\mathrm{Tr}(\boldsymbol{C}_i\boldsymbol{H}\boldsymbol{C}_k\boldsymbol{H}) - h_i^-).$$

For the first term, set $\boldsymbol{B} = \boldsymbol{C}_i^{1/2}\boldsymbol{H}\boldsymbol{C}_k\boldsymbol{H}\boldsymbol{C}_i^{1/2}$. Since

$$\|\boldsymbol{B}\|_{\mathrm{op}} \le a_i(\boldsymbol{H})a_k(\boldsymbol{H}), \qquad \mathrm{Tr}(\boldsymbol{B}) \le a_k(\boldsymbol{H})\mathrm{Tr}(\boldsymbol{C}_i\boldsymbol{H}),$$

Lemma 4 and $\|\boldsymbol{B}\|_F^2 \le \|\boldsymbol{B}\|_{\mathrm{op}}\mathrm{Tr}(\boldsymbol{B})$ give

$$\|w - \mathrm{Tr}(\boldsymbol{C}_i\boldsymbol{H}\boldsymbol{C}_k\boldsymbol{H})\|_{L^p} \le C_p\frac{\nu^2}{n_k\sqrt{n_i}}.$$

For the second term, apply the row-deletion martingale argument to $\mathrm{Tr}(\boldsymbol{C}_i\boldsymbol{H}\boldsymbol{C}_k\boldsymbol{H})$. For an additional row from dataset j, the conditional increment is bounded by

$$C_p a_j(\boldsymbol{R})a_k(\boldsymbol{R})\mathrm{Tr}(\boldsymbol{C}_i\boldsymbol{R})\big(1 + \mathrm{Tr}(\boldsymbol{C}_j\boldsymbol{R})\big),$$

where $\boldsymbol{R}$ has both rows removed. Lemma 5 therefore gives an unconditional increment bound $C_p\nu^4/(n_jn_k)$, and hence

$$\|\mathrm{Tr}(\boldsymbol{C}_i\boldsymbol{H}\boldsymbol{C}_k\boldsymbol{H}) - h_i^-\|_{L^p} \le C_p\frac{\nu^4 e_K}{n_k}.$$

Consequently,

$$\|w - h_i^-\|_{L^p} \le C_p\frac{\nu^4 e_K}{n_k}, \qquad |\mathbb{E}[ab(w - h_i^-)]| \le C\sqrt{\tau_{\boldsymbol{A}}[i]\mathcal{Q}_{\boldsymbol{A}}[i]}\frac{\nu^5 e_K}{n_k}.$$

Finally, we show that we can approximate $\mathbb{E}[ab]$ by q_i. A direct calculation gives

$$\mathbb{E}[ab] - q_i = \boldsymbol{\theta}^\top(\mathbb{E}[\boldsymbol{G}] - \overline{\boldsymbol{G}}_-)\boldsymbol{C}_i\overline{\boldsymbol{G}}_-\boldsymbol{\theta} + \boldsymbol{\theta}^\top\mathbb{E}[\boldsymbol{H} - \boldsymbol{G}]\boldsymbol{C}_i\overline{\boldsymbol{G}}_-\boldsymbol{\theta}.$$

For the second term, Sherman-Morrison gives

$$\boldsymbol{\theta}^\top(\boldsymbol{H} - \boldsymbol{G})\boldsymbol{C}_i\overline{\boldsymbol{G}}_-\boldsymbol{\theta} = \frac{ac}{f}.$$

Therefore, we have

$$\left|\boldsymbol{\theta}^\top(\mathbb{E}[\boldsymbol{H} - \boldsymbol{G}])\boldsymbol{C}_i\overline{\boldsymbol{G}}_-\boldsymbol{\theta}\right| \le C\frac{\nu^2}{n_i}\sqrt{\tau_{\boldsymbol{A}}[i]\mathcal{Q}_{\boldsymbol{A}}[i]}.$$

For the first term,

$$\left|\boldsymbol{\theta}^\top(\mathbb{E}[\boldsymbol{G}] - \overline{\boldsymbol{G}}_-)\boldsymbol{C}_i\overline{\boldsymbol{G}}_-\boldsymbol{\theta}\right| \le C\nu^2 m_i J_{\boldsymbol{A}},$$

where

$$m_i = \|\overline{\boldsymbol{G}}^{1/2}\boldsymbol{C}_i\overline{\boldsymbol{G}}^{1/2}\|_{\mathrm{op}}, \qquad J_{\boldsymbol{A}} = \sum_{j=1}^{K} n_j\sqrt{\tau_{\boldsymbol{A}}[j]\mathcal{Q}_{\boldsymbol{A}}[j]}\left(\delta_j + \frac{\nu}{n_j}\right).$$

Therefore,

$$|\mathbb{E}[ab] - q_i| \le C\left(\nu^2 m_i J_{\boldsymbol{A}} + \frac{\nu^2}{n_i}\sqrt{\tau_{\boldsymbol{A}}[i]\mathcal{Q}_{\boldsymbol{A}}[i]}\right).$$

Combining the three approximations,

$$\begin{aligned}\left|\mathbb{E}\left[\frac{abw}{f^2}\right] - \frac{q_i h_i^-}{(1+\kappa_i)^2}\right| &\le \left|\mathbb{E}[abw(f^{-2} - (1+\kappa_i)^{-2})]\right| + \left|\frac{1}{(1+\kappa_i)^2}\mathbb{E}[ab(w - h_i^-)]\right| + \left|\frac{h_i^-}{(1+\kappa_i)^2}(\mathbb{E}[ab] - q_i)\right| \\ &\le \frac{C\sqrt{\tau_{\boldsymbol{A}}[i]\mathcal{Q}_{\boldsymbol{A}}[i]}}{n_k}\left(\nu^3\delta_i + \nu^5 e_K + \frac{\nu^4}{n_i}\right) + \frac{C\nu^4 m_i}{n_k}J_{\boldsymbol{A}}.\end{aligned}$$

Bounding the remainder terms. For the remainder terms, applying the same conditional moment estimates and deleted-resolvent bounds as above, we obtain

$$\|a_k\|_{L^p} \le C_p\nu\sqrt{\frac{m_k\mathcal{Q}_{\boldsymbol{A}}[k]}{n_i}}, \qquad \|c_k\|_{L^p} \le C_p\frac{\nu^3 m_k}{n_i}\sqrt{\tau_{\boldsymbol{A}}[i]}.$$

Therefore, we can control the terms appearing in the remainder terms by

$$n_i\mathbb{E}[|(z-\kappa_i)ba_k|] \le C\nu^2\delta_i\sqrt{n_i m_k \tau_{\boldsymbol{A}}[i]\mathcal{Q}_{\boldsymbol{A}}[k]},$$
$$n_i\mathbb{E}[|ca_k|] \le C\nu^3\sqrt{m_k\tau_{\boldsymbol{A}}[i]\mathcal{Q}_{\boldsymbol{A}}[k]/n_i},$$
$$n_i\mathbb{E}[|ac_k|] \le C\nu^3 m_k\sqrt{\tau_{\boldsymbol{A}}[i]\mathcal{Q}_{\boldsymbol{A}}[i]},$$
$$n_i\mathbb{E}[|acw|] \le C\frac{\nu^4}{n_k}\sqrt{\tau_{\boldsymbol{A}}[i]\mathcal{Q}_{\boldsymbol{A}}[i]}.$$

Thus, we can write

$$T_{\boldsymbol{A},k} = q_k + \sum_{i=1}^{K}\frac{n_i}{(1+\kappa_i)^2}q_i h_i^- + \widetilde{r}_{\boldsymbol{A},k},$$

where $\widetilde{r}_{\boldsymbol{A},k}$ is the controlled error. By the same covariance-weighted row-increment estimate used above,

$$|h_i^- - T_{\boldsymbol{C}_i,k}| \le C\frac{\nu^4}{n_i n_k}, \qquad h_i^- + T_{\boldsymbol{C}_i,k} \le C\frac{\nu^2}{n_k}.$$

Also, since $(1+t_i)/(1+\kappa_i) \le C\nu$,

$$\left|\frac{n_i}{(1+\kappa_i)^2} - \frac{\mu_i^2}{n_i}\right| \le C\nu^2\delta\,\frac{\mu_i^2}{n_i}.$$

Together with the bounds on q_i, this gives

$$\left|\frac{n_i q_i}{(1+\kappa_i)^2} - \frac{\mu_i^2}{n_i}\tau_{\boldsymbol{A}}[i]\right| \le C n_i\nu^4\delta\,\tau_{\boldsymbol{A}}[i].$$

Therefore, we can write

$$T_{\boldsymbol{A},k} = \tau_{\boldsymbol{A}}[k] + \sum_{i=1}^{K}\frac{\mu_i^2}{n_i}\tau_{\boldsymbol{A}}[i]T_{\boldsymbol{C}_i,k} + r_{\boldsymbol{A},k}. \tag{36}$$

Now, we want to bound $\widetilde{r}_{\boldsymbol{A},k}$ and $r_{\boldsymbol{A},k}$. Recall that

$$\mathcal{Q}_{\boldsymbol{A}}[i] \le C\frac{\nu^5}{n_i}V_{\boldsymbol{A}}, \qquad \sqrt{\tau_{\boldsymbol{A}}[i]\mathcal{Q}_{\boldsymbol{A}}[i]} \le C\frac{\nu^3}{n_i}V_{\boldsymbol{A}}, \qquad J_{\boldsymbol{A}} \le C\nu^6 e_K V_{\boldsymbol{A}}.$$

Substituting these bounds into the preceding leading-term and remainder estimates gives

$$|\widetilde{r}_{\boldsymbol{A},k}| \le C\frac{\nu^{11}e_K}{n_k}V_{\boldsymbol{A}}.$$

The replacement of q_k, the coefficients, and h_i^- contributes at most

$$C\nu\delta\,\tau_{\boldsymbol{A}}[k] + C\frac{\nu^6\delta}{n_k}V_{\boldsymbol{A}} + C\frac{\nu^6}{n_k}\sum_i \tau_{\boldsymbol{A}}[i].$$

Using $\delta \le C\nu^6 e_K$, $\nu \ge 1$, and $e_K^2 \le \sqrt{K}\,e_K$, we get

$$|r_{\boldsymbol{A},k}| \le C\frac{V_{\boldsymbol{A}}}{n_k}\left(\nu^{11}e_K + \nu\delta + \nu^6\delta + \nu^6 e_K^2\right)$$
$$\le C\frac{\nu^{12}e_K}{n_k}V_{\boldsymbol{A}}.$$

Having established a bound on $T_{\boldsymbol{A},k}$, now we consider the coupled system. Define $h_j = T_{\boldsymbol{C}_j,k} = \mathbb{E}[\mathrm{Tr}(\boldsymbol{C}_j\boldsymbol{G}\boldsymbol{C}_k\boldsymbol{G})]$, and write

$$\boldsymbol{h} = \begin{pmatrix} h_1 \\ h_2 \\ \vdots \\ h_K \end{pmatrix}.$$

Setting $\boldsymbol{A} = \boldsymbol{C}_j$, we have for all $j \in [K]$,

$$h_j = \tau_{\boldsymbol{C}_j}[k] + \sum_{i=1}^{K} \frac{\mu_i^2}{n_i} \tau_{\boldsymbol{C}_j}[i] h_i + r_{\boldsymbol{C}_j,k}. \tag{37}$$

Recall that

$$\boldsymbol{K}_{j,i} = \operatorname{Tr}(\boldsymbol{C}_j \overline{\boldsymbol{G}} \boldsymbol{C}_i \overline{\boldsymbol{G}}), \qquad \boldsymbol{D} = \operatorname{diag}\left(\frac{n_1}{\mu_1^2}, \cdots, \frac{n_K}{\mu_K^2}\right), \qquad \boldsymbol{L} = \boldsymbol{D} - \boldsymbol{K}.$$

Then, we can write (37) as

$$\boldsymbol{h} = \boldsymbol{K}\boldsymbol{e}_k + \boldsymbol{K}\boldsymbol{D}^{-1}\boldsymbol{h} + \boldsymbol{r}.$$

Substituting it back to (36), we get

$$T_{\boldsymbol{A},k} = \underbrace{\tau_{\boldsymbol{A}}[k] + \tau_{\boldsymbol{A}}^\top \boldsymbol{L}^{-1} \tau_{\boldsymbol{C}_k}}_{\text{deterministic equivalent}} + \underbrace{r_{\boldsymbol{A},k} + \sum_{1 \le i,j \le K} \tau_{\boldsymbol{A}}[i] (\boldsymbol{L}^{-1})_{i,j} \boldsymbol{r}_{\boldsymbol{C}_j,k}}_{\text{remaining error}}.$$

We finally have

$$\left|\mathbb{E}[\operatorname{Tr}(\boldsymbol{A}\boldsymbol{G}\boldsymbol{C}_k\boldsymbol{G})] - \boldsymbol{e}_k^\top \boldsymbol{D}\boldsymbol{L}^{-1}\tau_{\boldsymbol{A}}\right| \le C\frac{\nu^{12} e_K}{n_k} \left(V_{\boldsymbol{A}} + \sum_{i,j} \tau_{\boldsymbol{A}}[i] (\boldsymbol{L}^{-1})_{ij} V_{\boldsymbol{C}_j}\right).$$

Now we use established upper bounds to control the terms appearing in the above expression. Since

$$\sum_i n_i \boldsymbol{C}_i = \sum_i \mu_i (1 + t_i) \boldsymbol{C}_i \preceq C\nu \sum_i \mu_i \boldsymbol{C}_i \preceq C\nu \overline{\boldsymbol{G}}^{-1},$$

we have

$$V_{\boldsymbol{C}_j} = \operatorname{Tr}\left(\boldsymbol{C}_j \overline{\boldsymbol{G}} \left(\sum_i n_i \boldsymbol{C}_i\right) \overline{\boldsymbol{G}}\right) \le C\nu t_j \le C\nu^2.$$

Furthermore, because

$$\boldsymbol{L}^{-1} \ge 0 \text{ entrywise}, \qquad \boldsymbol{L}\boldsymbol{\mu} = \mathbf{1} + \lambda \tau_{\boldsymbol{I}}, \qquad \boldsymbol{L}^{-1}\mathbf{1} \le \boldsymbol{\mu},$$

we have

$$\begin{aligned} \sum_{i,j} \tau_{\boldsymbol{A}}[i] (\boldsymbol{L}^{-1})_{ij} V_{\boldsymbol{C}_j} &\le C\nu^2 \tau_{\boldsymbol{A}}^\top \boldsymbol{L}^{-1}\mathbf{1} \\ &\le C\nu^2 \sum_i \mu_i \tau_{\boldsymbol{A}}[i] \le C\nu^2 V_{\boldsymbol{A}}. \end{aligned}$$

Combining these bounds, we obtain

$$\left|\mathbb{E}[\operatorname{Tr}(\boldsymbol{A}\boldsymbol{G}\boldsymbol{C}_k\boldsymbol{G})] - \boldsymbol{e}_k^\top \boldsymbol{D}\boldsymbol{L}^{-1}\tau_{\boldsymbol{A}}\right| \le C\frac{\nu^{14} e_K}{n_k} V_{\boldsymbol{A}}.$$

F.4.2 Martingale part of $\operatorname{Tr}(\boldsymbol{A}\boldsymbol{G}\boldsymbol{C}_k\boldsymbol{G})$

For $i \in [K]$ and $r \in [n_i]$, define

$$\boldsymbol{H} = \boldsymbol{G}_{-i,r}, \qquad s = \boldsymbol{x}_{i,r}^\top \boldsymbol{H} \boldsymbol{x}_{i,r}.$$

We use Sherman-Morrison to get

$$\operatorname{Tr}(\boldsymbol{A}\boldsymbol{G}\boldsymbol{C}_k\boldsymbol{G}) - \operatorname{Tr}(\boldsymbol{A}\boldsymbol{H}\boldsymbol{C}_k\boldsymbol{H}) = \underbrace{-\frac{2\boldsymbol{x}^\top \boldsymbol{H}\boldsymbol{A}\boldsymbol{H}\boldsymbol{C}_k\boldsymbol{H}\boldsymbol{x}}{1+s}}_{\text{cross term}} + \underbrace{\frac{(\boldsymbol{x}^\top \boldsymbol{H}\boldsymbol{A}\boldsymbol{H}\boldsymbol{x})(\boldsymbol{x}^\top \boldsymbol{H}\boldsymbol{C}_k\boldsymbol{H}\boldsymbol{x})}{(1+s)^2}}_{\text{product term}}. \tag{38}$$

Define

$$t_j = \operatorname{Tr}(\boldsymbol{C}_j \overline{\boldsymbol{G}}), \qquad W_{\boldsymbol{A}} = \sum_{j=1}^{K} n_j t_j \tau_{\boldsymbol{A}}[j], \qquad Q_j = \tau_{\boldsymbol{A}}[j] + \nu^3 \|\overline{\boldsymbol{G}}^{1/2} \boldsymbol{C}_j \overline{\boldsymbol{G}}^{1/2}\|_{\mathrm{op}} W_{\boldsymbol{A}}.$$

By Lemma 6, we have

$$\|\mathrm{Tr}(\boldsymbol{AHC}_j\boldsymbol{H})\|_{L^p} \le C_p Q_j.$$

Now we need to bound the cross term and the product term respectively in the decomposition of (38).

For the cross term, as in (Misiakiewicz & Saeed, 2024, Appendix A.3.1), it suffices to assume $\boldsymbol{A} = \boldsymbol{\theta}\boldsymbol{\theta}^\top$. Setting $a_j(\boldsymbol{H}) = \|\boldsymbol{C}_j^{1/2}\boldsymbol{H}\boldsymbol{C}_j^{1/2}\|_{\mathrm{op}}$, we can bound $\|a_j(\boldsymbol{H})\|_{L^p} \le C_p\nu/n_j$ with Lemma 5. Using Lemma 4 and Hölder's inequality, we have

$$\|\boldsymbol{x}^\top\boldsymbol{HAHC}_k\boldsymbol{Hx}\|_{L^p(\boldsymbol{x}|\boldsymbol{H})} \le C_p\sqrt{a_i(\boldsymbol{H})a_k(\boldsymbol{H})\mathrm{Tr}(\boldsymbol{AHC}_i\boldsymbol{H})\mathrm{Tr}(\boldsymbol{AHC}_k\boldsymbol{H})}.$$

Taking the remaining expectation and applying Hölder, we get

$$\|\boldsymbol{x}^\top\boldsymbol{HAHC}_k\boldsymbol{Hx}\|_{L^p} \le C_p\nu\sqrt{\frac{Q_iQ_k}{n_in_k}}.$$

For the product term, similarly we have

$$\|\boldsymbol{x}^\top\boldsymbol{HAHxx}^\top\boldsymbol{HC}_k\boldsymbol{Hx}\|_{L^p(\boldsymbol{x}|\boldsymbol{H})} \le C_p\mathrm{Tr}(\boldsymbol{AHC}_i\boldsymbol{H})\mathrm{Tr}(\boldsymbol{C}_i\boldsymbol{HC}_k\boldsymbol{H}).$$

We have already obtained a bound on $\mathrm{Tr}(\boldsymbol{AHC}_i\boldsymbol{H})$. For $\mathrm{Tr}(\boldsymbol{C}_i\boldsymbol{HC}_k\boldsymbol{H})$,

$$\mathrm{Tr}(\boldsymbol{C}_i\boldsymbol{HC}_k\boldsymbol{H}) \le a_i(\boldsymbol{H})\mathrm{Tr}(C_k\boldsymbol{H}) \wedge a_k(\boldsymbol{H})\mathrm{Tr}(\boldsymbol{C}_i\boldsymbol{H}).$$

Taking moments and applying Hölder's inequality, together with the same basic deleted-resolvent bounds, gives

$$\|\mathrm{Tr}(\boldsymbol{C}_i\boldsymbol{HC}_k\boldsymbol{H})\|_{L^p} \le C_p\nu^2\left(\frac{t_k}{n_i} \wedge \frac{t_i}{n_k}\right).$$

Combining the previous results and setting

$$b_i = \nu\sqrt{\frac{Q_iQ_k}{n_in_k}} + \nu^2Q_i\left(\frac{t_i}{n_k} \wedge \frac{t_k}{n_i}\right),$$

we have

$$\|\mathrm{Tr}(\boldsymbol{AGC}_k\boldsymbol{G}) - \mathrm{Tr}(\boldsymbol{AHC}_k\boldsymbol{H})\|_{L^p} \le C_pb_i.$$

As before, we construct a martingale difference sequence and write

$$\begin{aligned}\mathrm{Tr}(\boldsymbol{AGC}_k\boldsymbol{G}) - \mathbb{E}[\mathrm{Tr}(\boldsymbol{AGC}_k\boldsymbol{G})] &= \sum_{\ell=1}^{N}(\mathbb{E}_\ell - \mathbb{E}_{\ell-1})\mathrm{Tr}(\boldsymbol{AGC}_k\boldsymbol{G}) \\ &= \sum_{\ell=1}^{N}(\mathbb{E}_\ell - \mathbb{E}_{\ell-1})(\mathrm{Tr}(\boldsymbol{AGC}_k\boldsymbol{G}) - \mathrm{Tr}(\boldsymbol{AG}_{-\ell}\boldsymbol{C}_k\boldsymbol{G}_{-\ell})),\end{aligned}$$

where $N = \sum_{k=1}^K n_k$ and the ℓ-th feature we removed belongs to $\boldsymbol{X}_{i,r}$. For convenience, we denote

$$\Delta_\ell = \mathrm{Tr}(\boldsymbol{AGC}_k\boldsymbol{G}) - \mathrm{Tr}(\boldsymbol{AG}_{-\ell}\boldsymbol{C}_k\boldsymbol{G}_{-\ell}), \qquad d_\ell = (\mathbb{E}_\ell - \mathbb{E}_{\ell-1})\Delta_\ell.$$

By the previous bound on $\mathrm{Tr}(\boldsymbol{AGC}_k\boldsymbol{G}) - \mathrm{Tr}(\boldsymbol{AG}_-\boldsymbol{C}_k\boldsymbol{G}_-)$, we have

$$\|(\mathbb{E}_\ell - \mathbb{E}_{\ell-1})(\mathrm{Tr}(\boldsymbol{AGC}_k\boldsymbol{G}) - \mathrm{Tr}(\boldsymbol{AG}_{-\ell}\boldsymbol{C}_k\boldsymbol{G}_{-\ell}))\|_{L^p} \le 2\|\Delta_\ell\|_{L^p} \le C_pb_i.$$

By the martingale moment inequality, we have

$$\|\mathrm{Tr}(\boldsymbol{AGC}_k\boldsymbol{G}) - \mathbb{E}[\mathrm{Tr}(\boldsymbol{AGC}_k\boldsymbol{G})]\|_{L^p} \le C_p\left(\sum_{i=1}^K n_ib_i^2\right)^{1/2} \le C_p\nu^8e_K^3V_{\boldsymbol{A}}.$$

For the probability bound, set $n_{\min} = \min_i n_i$. Applying Lemma 4 conditionally on $\boldsymbol{H}$ and the probabilistic results in Lemma 5 and 6 gives

$$\mathbb{P}(|\Delta_\ell| > R_i) \le C_Mn_{\min}^{-M}, \qquad R_i := C_M\log^{4\beta+1}(N)b_i,$$

where observation ℓ belongs to dataset i. Define

$$\widetilde{\Delta}_\ell = \Delta_\ell\mathbf{1}_{\{|\Delta_\ell|\le R_i\}}, \qquad \widetilde{d}_\ell = (\mathbb{E}_\ell - \mathbb{E}_{\ell-1})\widetilde{\Delta}_\ell.$$

Then $|\widetilde{d}_\ell| \le 2R_i$ almost surely. By Cauchy–Schwarz,

$$\mathbb{E}[|\Delta_\ell|\mathbf{1}_{\{|\Delta_\ell|>R_i\}}] \le C_M b_i n_{\min}^{-M/2}.$$

The previously established bounds imply

$$b_i \le CV_{\boldsymbol{A}}\left(\frac{\nu^6}{n_i n_k} + \frac{\nu^8}{n_i^2}\right), \qquad \sum_i n_i b_i \le C_K \nu^8 e_K^2 V_{\boldsymbol{A}}.$$

Thus, summing and applying Markov's inequality, using $e_K \ge n_{\min}^{-1/2}$, gives

$$\mathbb{P}\left(\left|\sum_{\ell=1}^N (d_\ell - \widetilde{d}_\ell)\right| > \nu^8 e_K^3 V_{\boldsymbol{A}}\right) \le C_{M,K} n_{\min}^{-(M-1)/2} \le \frac{1}{2}\sum_i n_i^{-D}$$

for sufficiently large M and $n_{\min}$.

By Azuma–Hoeffding, with probability at least $1 - \frac{1}{2}\sum_i n_i^{-D}$,

$$\begin{aligned}\left|\sum_{\ell=1}^N \widetilde{d}_\ell\right| &\le C_D\sqrt{\log(N)\sum_i n_i R_i^2} \\ &\le C_D \log^{4\eta+3/2}(N)\left(\sum_i n_i b_i^2\right)^{1/2} \\ &\le C_D \nu^8 e_K^3 \log^{4\eta+3/2}(N) V_{\boldsymbol{A}}.\end{aligned}$$

Therefore, with probability at least $1 - \sum_i n_i^{-D}$,

$$|\mathrm{Tr}(\boldsymbol{AGC}_k\boldsymbol{G}) - \mathbb{E}[\mathrm{Tr}(\boldsymbol{AGC}_k\boldsymbol{G})]| \le C_D \nu^8 e_K^3 \log^{4\beta+3/2}(N) V_{\boldsymbol{A}}.$$

Combined with the deterministic part of $\mathrm{Tr}(\boldsymbol{AGC}_k\boldsymbol{G})$, we get

$$\left|\mathrm{Tr}(\boldsymbol{AGC}_k\boldsymbol{G}) - \boldsymbol{e}_k^\top \boldsymbol{D}\boldsymbol{L}^{-1}\tau_{\boldsymbol{A}}\right| \le C_D e_K\left(\frac{\nu^{14}}{n_k} + \nu^8 e_K^2 \log^{4\beta+3/2}(N)\right) V_{\boldsymbol{A}}.$$

F.5 Deterministic equivalence of $\mathrm{Tr}(\boldsymbol{AGX}_k^\top \boldsymbol{X}_k\boldsymbol{G})$

F.5.1 Deterministic part of $\mathrm{Tr}(\boldsymbol{AGX}_k^\top \boldsymbol{X}_k\boldsymbol{G})$

Let $\boldsymbol{x} = \boldsymbol{x}_{k,1}$ and $\boldsymbol{H}$ be the resolvent with this feature removed. Set

$$z = \boldsymbol{x}^\top \boldsymbol{H}\boldsymbol{x}, \qquad \kappa_k = \mathbb{E}[\mathrm{Tr}(\boldsymbol{C}_k\boldsymbol{H})], \qquad y = \boldsymbol{x}^\top \boldsymbol{H}\boldsymbol{A}\boldsymbol{H}\boldsymbol{x}.$$

Then, we have

$$\mathbb{E}[\mathrm{Tr}(\boldsymbol{AGX}_k^\top \boldsymbol{X}_k\boldsymbol{G})] = n_k \mathbb{E}\left[\frac{y}{(1+z)^2}\right].$$

First, we show we can replace the denominator by $(1+\kappa_k)^2$ up to a small error. We have

$$\left|\frac{1}{(1+z)^2} - \frac{1}{(1+\kappa_k)^2}\right| \le 2|z - \kappa_k|.$$

By the same conditional quadratic-form and first-order martingale estimates used in the deterministic part of $\mathrm{Tr}(\boldsymbol{AG})$, we have

$$\|z - \kappa_k\|_{L^q} \le C_q\left(\frac{\nu}{\sqrt{n_k}} + \nu^2 e_K t_k\right) \le C_q \nu^3 e_K.$$

Moreover, applying Lemma 4 conditionally on $\boldsymbol{H}$, followed by Lemma 6, gives

$$\|y\|_{L^q} \le C_q \|\mathrm{Tr}(\boldsymbol{AHC}_k\boldsymbol{H})\|_{L^q} \le C_q \frac{\nu^5}{n_k}\sum_{i=1}^K n_i \tau_{\boldsymbol{A}}[i].$$

Therefore,

$$\left|\mathbb{E}[\mathrm{Tr}(\boldsymbol{A}\boldsymbol{G}\boldsymbol{X}_k^\top\boldsymbol{X}_k\boldsymbol{G})]-\frac{n_k}{(1+\kappa_k)^2}\mathbb{E}[\mathrm{Tr}(\boldsymbol{A}\boldsymbol{H}\boldsymbol{C}_k\boldsymbol{H})]\right|\le C\nu^8 e_K\sum_{i=1}^K n_i\tau_{\boldsymbol{A}}[i].$$

Now, we show we can approximate $\mathrm{Tr}(\boldsymbol{A}\boldsymbol{H}\boldsymbol{C}_k\boldsymbol{H})$ by $\mathrm{Tr}(\boldsymbol{A}\boldsymbol{G}\boldsymbol{C}_k\boldsymbol{G})$, by noticing that

$$\mathrm{Tr}(\boldsymbol{A}\boldsymbol{H}\boldsymbol{C}_k\boldsymbol{H})-\mathrm{Tr}(\boldsymbol{A}\boldsymbol{G}\boldsymbol{C}_k\boldsymbol{G})=2\frac{\boldsymbol{x}^\top\boldsymbol{H}\boldsymbol{A}\boldsymbol{H}\boldsymbol{C}_k\boldsymbol{H}\boldsymbol{x}}{1+z}-\frac{(\boldsymbol{x}^\top\boldsymbol{H}\boldsymbol{A}\boldsymbol{H}\boldsymbol{x})(\boldsymbol{x}^\top\boldsymbol{H}\boldsymbol{C}_k\boldsymbol{H}\boldsymbol{x})}{(1+z)^2}.$$

Applying the conditional moment estimates of Lemma 4, we have

$$\begin{aligned}\mathbb{E}_{\boldsymbol{x}}[|\boldsymbol{x}^\top\boldsymbol{H}\boldsymbol{A}\boldsymbol{H}\boldsymbol{C}_k\boldsymbol{H}\boldsymbol{x}|]&\le C\|\boldsymbol{C}_k^{1/2}\boldsymbol{H}\boldsymbol{C}_k^{1/2}\|_{\mathrm{op}}\mathrm{Tr}(\boldsymbol{A}\boldsymbol{H}\boldsymbol{C}_k\boldsymbol{H}),\\ \mathbb{E}_{\boldsymbol{x}}[(\boldsymbol{x}^\top\boldsymbol{H}\boldsymbol{A}\boldsymbol{H}\boldsymbol{x})(\boldsymbol{x}^\top\boldsymbol{H}\boldsymbol{C}_k\boldsymbol{H}\boldsymbol{x})]&\le C\mathrm{Tr}(\boldsymbol{A}\boldsymbol{H}\boldsymbol{C}_k\boldsymbol{H})\mathrm{Tr}(\boldsymbol{C}_k\boldsymbol{H}\boldsymbol{C}_k\boldsymbol{H}).\end{aligned}$$

Again, by the basic deleted-resolvent moment bounds in Lemma 5, together with

$$\mathrm{Tr}(\boldsymbol{C}_k\boldsymbol{H}\boldsymbol{C}_k\boldsymbol{H})\le\|\boldsymbol{C}_k^{1/2}\boldsymbol{H}\boldsymbol{C}_k^{1/2}\|_{\mathrm{op}}\mathrm{Tr}(\boldsymbol{C}_k\boldsymbol{H}),\qquad\|\|\boldsymbol{C}_k^{1/2}\boldsymbol{H}\boldsymbol{C}_k^{1/2}\|_{\mathrm{op}}\|_{L^q}\le C_q\frac{\nu}{n_k},$$

$$\|\mathrm{Tr}(\boldsymbol{C}_k\boldsymbol{H})\|_{L^q}\le C_q\nu,$$

we have

$$\frac{n_k}{(1+\kappa_k)^2}\mathbb{E}[|\mathrm{Tr}(\boldsymbol{A}\boldsymbol{H}\boldsymbol{C}_k\boldsymbol{H})-\mathrm{Tr}(\boldsymbol{A}\boldsymbol{G}\boldsymbol{C}_k\boldsymbol{G})|]\le C\frac{\nu^7}{n_k^2}\sum_{i=1}^K n_i\tau_{\boldsymbol{A}}[i]\le C\nu^8 e_K\sum_{i=1}^K n_i\tau_{\boldsymbol{A}}[i].$$

Therefore,

$$\left|\mathbb{E}[\mathrm{Tr}(\boldsymbol{A}\boldsymbol{G}\boldsymbol{X}_k^\top\boldsymbol{X}_k\boldsymbol{G})]-\frac{n_k}{(1+\kappa_k)^2}\mathbb{E}[\mathrm{Tr}(\boldsymbol{A}\boldsymbol{G}\boldsymbol{C}_k\boldsymbol{G})]\right|\le C\nu^8 e_K\sum_{i=1}^K n_i\tau_{\boldsymbol{A}}[i].$$

Finally, using the deterministic equivalent of $\mathrm{Tr}(\boldsymbol{A}\boldsymbol{G}\boldsymbol{C}_k\boldsymbol{G})$,

$$|\mathbb{E}[\mathrm{Tr}(\boldsymbol{A}\boldsymbol{G}\boldsymbol{C}_k\boldsymbol{G})]-\boldsymbol{e}_k^\top\boldsymbol{D}\boldsymbol{L}^{-1}\tau_{\boldsymbol{A}}|\le C\frac{\nu^{14}e_K}{n_k}V_{\boldsymbol{A}},$$

we get

$$\begin{aligned}&|\mathbb{E}[\mathrm{Tr}(\boldsymbol{A}\boldsymbol{G}\boldsymbol{X}_k^\top\boldsymbol{X}_k\boldsymbol{G})]-\boldsymbol{e}_k^\top\boldsymbol{L}^{-1}\tau_{\boldsymbol{A}}|\\ &\le\underbrace{\left|\mathbb{E}[\mathrm{Tr}(\boldsymbol{A}\boldsymbol{G}\boldsymbol{X}_k^\top\boldsymbol{X}_k\boldsymbol{G})]-\frac{n_k}{(1+\kappa_k)^2}\mathbb{E}[\mathrm{Tr}(\boldsymbol{A}\boldsymbol{G}\boldsymbol{C}_k\boldsymbol{G})]\right|}_{(I)}+\underbrace{\left|\frac{n_k}{(1+\kappa_k)^2}\mathbb{E}[\mathrm{Tr}(\boldsymbol{A}\boldsymbol{G}\boldsymbol{C}_k\boldsymbol{G})]-\boldsymbol{e}_k^\top\boldsymbol{L}^{-1}\tau_{\boldsymbol{A}}\right|}_{(II)}.\end{aligned}$$

For (I), we have

$$(I)\le C\nu^8 e_K V_{\boldsymbol{A}}.$$

For (II), similar to the deterministic part of $\mathrm{Tr}(\boldsymbol{A}\boldsymbol{G}\boldsymbol{C}_k\boldsymbol{G})$, and the coefficient comparison established in the preceding subsection, we have

$$\begin{aligned}(II)&\le\frac{n_k}{(1+\kappa_k)^2}\left|\mathbb{E}[\mathrm{Tr}(\boldsymbol{A}\boldsymbol{G}\boldsymbol{C}_k\boldsymbol{G})]-\boldsymbol{e}_k^\top\boldsymbol{D}\boldsymbol{L}^{-1}\tau_{\boldsymbol{A}}\right|+\left|\frac{n_k}{(1+\kappa_k)^2}-\frac{\mu_k^2}{n_k}\right|\boldsymbol{e}_k^\top\boldsymbol{D}\boldsymbol{L}^{-1}\tau_{\boldsymbol{A}}\\ &\le Cn_k\frac{\nu^{14}e_K}{n_k}V_{\boldsymbol{A}}+C\nu^8 e_K\boldsymbol{e}_k^\top\boldsymbol{L}^{-1}\tau_{\boldsymbol{A}},\\ &\le C\nu^{14}e_K V_{\boldsymbol{A}},\end{aligned}$$

where we use $\boldsymbol{D}_{k,k}=n_k/\mu_k^2$, $\boldsymbol{e}_k^\top\boldsymbol{L}^{-1}\tau_{\boldsymbol{A}}\le V_{\boldsymbol{A}}$, and $\nu\ge 1$. Combining the previous bounds, we obtain

$$\left|\mathbb{E}[\mathrm{Tr}(\boldsymbol{A}\boldsymbol{G}\boldsymbol{X}_k^\top\boldsymbol{X}_k\boldsymbol{G})]-\boldsymbol{e}_k^\top\boldsymbol{L}^{-1}\tau_{\boldsymbol{A}}\right|\le C\nu^{14}e_K V_{\boldsymbol{A}}.$$

F.5.2 MARTINGALE PART OF $\mathrm{Tr}(\boldsymbol{A}\boldsymbol{G}\boldsymbol{X}_k^\top \boldsymbol{X}_k \boldsymbol{G})$

For convenience, we denote $\boldsymbol{S}_k = \boldsymbol{X}_k^\top \boldsymbol{X}_k$. For $\boldsymbol{x} = \boldsymbol{x}_{i,r}$, we set

$$\boldsymbol{S}_k = \boldsymbol{S}_{k,-} + \mathbf{1}_{\{i=k\}} \boldsymbol{x}\boldsymbol{x}^\top.$$

Note that $\mathrm{Tr}(\boldsymbol{A}\boldsymbol{H}\boldsymbol{S}_{k,-}\boldsymbol{H})$ is independent of $\boldsymbol{x}$. Setting $z = \boldsymbol{x}^\top \boldsymbol{H}\boldsymbol{x}$, we get

$$\begin{aligned}\xi &:= \mathrm{Tr}(\boldsymbol{A}\boldsymbol{G}\boldsymbol{S}_k\boldsymbol{G}) - \mathrm{Tr}(\boldsymbol{A}\boldsymbol{H}\boldsymbol{S}_{k,-}\boldsymbol{H}) \\ &= -2\frac{\boldsymbol{x}^\top \boldsymbol{H}\boldsymbol{A}\boldsymbol{H}\boldsymbol{S}_{k,-}\boldsymbol{H}\boldsymbol{x}}{1+z} + \frac{(\boldsymbol{x}^\top \boldsymbol{H}\boldsymbol{A}\boldsymbol{H}\boldsymbol{x})(\boldsymbol{x}^\top \boldsymbol{H}\boldsymbol{S}_{k,-}\boldsymbol{H}\boldsymbol{x})}{(1+z)^2} + \mathbf{1}_{\{i=k\}}\frac{\boldsymbol{x}^\top \boldsymbol{H}\boldsymbol{A}\boldsymbol{H}\boldsymbol{x}}{(1+z)^2}.\end{aligned}$$

Since $0 \preceq \boldsymbol{S}_{k,-} \preceq \boldsymbol{H}^{-1}$, we have $\boldsymbol{x}^\top \boldsymbol{H}\boldsymbol{S}_{k,-}\boldsymbol{H}\boldsymbol{x} \le z$. This shows that the sum of the last two terms is bounded by $2\boldsymbol{x}^\top \boldsymbol{H}\boldsymbol{A}\boldsymbol{H}\boldsymbol{x}$, whose moments, by Lemma 4 conditionally on $\boldsymbol{H}$, followed by Lemma 6, satisfy

$$\|\boldsymbol{x}^\top \boldsymbol{H}\boldsymbol{A}\boldsymbol{H}\boldsymbol{x}\|_{L^q} \le C_q \frac{\nu^5}{n_i}\sum_{j=1}^K n_j \tau_{\boldsymbol{A}}[j].$$

For the cross term, conditioning on the remaining observations and applying Lemma 4 and Cauchy–Schwarz give

$$\begin{aligned}\|\boldsymbol{x}^\top \boldsymbol{H}\boldsymbol{A}\boldsymbol{H}\boldsymbol{S}_{k,-}\boldsymbol{H}\boldsymbol{x}\|_{L^q(\boldsymbol{x}|\cdot)} &\le \|\boldsymbol{x}^\top \boldsymbol{H}\boldsymbol{A}\boldsymbol{H}\boldsymbol{x}\|^{1/2}_{L^q(\boldsymbol{x}|\cdot)} \|\boldsymbol{x}^\top \boldsymbol{H}\boldsymbol{S}_{k,-}\boldsymbol{H}\boldsymbol{A}\boldsymbol{H}\boldsymbol{S}_{k,-}\boldsymbol{H}\boldsymbol{x}\|^{1/2}_{L^q(\boldsymbol{x}|\cdot)} \\ &\le C_q \sqrt{\mathrm{Tr}(\boldsymbol{A}\boldsymbol{H}\boldsymbol{C}_i\boldsymbol{H})\mathrm{Tr}(\boldsymbol{A}\boldsymbol{H}\boldsymbol{S}_{k,-}\boldsymbol{H}\boldsymbol{C}_i\boldsymbol{H}\boldsymbol{S}_{k,-}\boldsymbol{H})} \\ &\le C_q \sqrt{\|\boldsymbol{C}_i^{1/2}\boldsymbol{H}\boldsymbol{C}_i^{1/2}\|_{\mathrm{op}}\mathrm{Tr}(\boldsymbol{A}\boldsymbol{H}\boldsymbol{C}_i\boldsymbol{H})\mathrm{Tr}(\boldsymbol{A}\boldsymbol{H}\boldsymbol{S}_{k,-}\boldsymbol{H})}.\end{aligned}$$

With the basic deleted-resolvent moment bounds, Lemma 6, and the empirical-covariance estimate

$$\|\boldsymbol{C}_i^{1/2}\boldsymbol{H}\boldsymbol{C}_i^{1/2}\|_{\mathrm{op}} \le C_i\frac{\nu}{n_i}, \qquad \|\mathrm{Tr}(\boldsymbol{A}\boldsymbol{H}\boldsymbol{C}_i\boldsymbol{H})\|_{L^q} \le C_q\frac{\nu^5}{n_i}\sum_{j=1}^K n_j\tau_{\boldsymbol{A}}[j],$$

$$\|\mathrm{Tr}(\boldsymbol{A}\boldsymbol{H}\boldsymbol{S}_{k,-}\boldsymbol{H})\|_{L^q} \le C_q\nu^3\sum_{j=1}^K n_j\tau_{\boldsymbol{A}}[j],$$

we have

$$\|\boldsymbol{x}^\top \boldsymbol{H}\boldsymbol{A}\boldsymbol{H}\boldsymbol{S}_{k,-}\boldsymbol{H}\boldsymbol{x}\|_{L^q} \le C_q\frac{\nu^5}{n_i}\sum_{j=1}^K n_j\tau_{\boldsymbol{A}}[j].$$

Here, the moment bound on $\mathrm{Tr}(\boldsymbol{A}\boldsymbol{H}\boldsymbol{S}_{k,-}\boldsymbol{H})$ holds because

$$\begin{aligned}\mathrm{Tr}(\boldsymbol{A}\boldsymbol{H}\boldsymbol{S}_{k,-}\boldsymbol{H}) &\le 2\mathrm{Tr}(\boldsymbol{A}\overline{\boldsymbol{G}}\boldsymbol{S}_{k,-}\overline{\boldsymbol{G}}) + 2\|\overline{\boldsymbol{G}}^{-1/2}\boldsymbol{H}\overline{\boldsymbol{G}}^{-1/2}\|_{\mathrm{op}}\|\boldsymbol{Y}\|_F^2 \\ &\le C_q(n_k\tau_{\boldsymbol{A}}[k] + \nu^2 W_{\boldsymbol{A}}) \\ &\le C_q\nu^3 V_{\boldsymbol{A}}.\end{aligned}$$

Now we obtain the bound

$$\|\xi\|_{L^q} \le C_q\frac{\nu^5}{n_i}\sum_{j=1}^K n_j\tau_{\boldsymbol{A}}[j].$$

The rest of the proof follows similarly. We again define the martingale difference sequence, and get

$$\begin{aligned}\|\mathrm{Tr}(\boldsymbol{AGS}_k\boldsymbol{G}) - \mathbb{E}[\mathrm{Tr}(\boldsymbol{AGS}_k\boldsymbol{G})]\|_{L^q} &= \left\|\sum_{\ell=1}^{K}(\mathbb{E}_\ell - \mathbb{E}_{\ell-1})\mathrm{Tr}(\boldsymbol{AGS}_k\boldsymbol{G})\right\|_{L^q} \\ &\le C_q\left(\sum_{\ell=1}^{K}\|(\mathbb{E}_\ell - \mathbb{E}_{\ell-1})\mathrm{Tr}(\boldsymbol{AGS}_k\boldsymbol{G})\|_{L^q}^2\right)^{1/2} \\ &\le C_q\nu^5\sum_{j=1}^{K} n_k\tau_{\boldsymbol{A}}[j]\left(\sum_{j=1}^{K}\frac{1}{n_j}\right)^{1/2} \\ &= C_q\nu^5 e_K\sum_{j=1}^{K} n_j\tau_{\boldsymbol{A}}[j].\end{aligned}$$

For a probabilistic bound, set $n_{\min} = \min_i n_i$. Similar to the martingale part of $\mathrm{Tr}(\boldsymbol{AGC}_k\boldsymbol{G})$, we have

$$\mathbb{P}(|\xi_\ell| > R_i) \le C_M n_{\min}^{-M}, \qquad R_i := C_M\frac{\nu^5}{n_i}\log^{3\eta+1}(N)V_{\boldsymbol{A}}.$$

Define

$$\widetilde{\xi}_\ell = \xi_\ell\mathbf{1}_{\{|\xi_\ell|\le R_i\}}, \qquad \widetilde{d}_\ell = (\mathbb{E}_\ell - \mathbb{E}_{\ell-1})\widetilde{\xi}_\ell.$$

Then $|\widetilde{d}_\ell| \le 2R_i$ almost surely, and Cauchy–Schwarz gives

$$\mathbb{E}[|\xi_\ell|\mathbf{1}_{\{|\xi_\ell|>R_i\}}] \le C_M\frac{\nu^5}{n_i}n_{\min}^{-M/2}V_{\boldsymbol{A}}.$$

Summing and applying Markov's inequality, for sufficiently large M and $n_{\min}$, with probability at least $1 - \frac{1}{2}\sum_i n_i^{-D}$,

$$\left|\sum_{\ell=1}^{N}(d_\ell - \widetilde{d}_\ell)\right| \le C\nu^5 e_K V_{\boldsymbol{A}}.$$

By Azuma–Hoeffding, with probability at least $1 - \frac{1}{2}\sum_i n_i^{-D}$,

$$\left|\sum_{\ell=1}^{N}\widetilde{d}_\ell\right| \le C_D\sqrt{\log(N)\sum_{i=1}^{K} n_i R_i^2} \le C_D\nu^5 e_K\log^{3\eta+3/2}(N)V_{\boldsymbol{A}}.$$

Therefore, with probability at least $1 - \sum_i n_i^{-D}$,

$$|\mathrm{Tr}(\boldsymbol{AGX}_k^\top\boldsymbol{X}_k\boldsymbol{G}) - \mathbb{E}[\mathrm{Tr}(\boldsymbol{AGX}_k^\top\boldsymbol{X}_k\boldsymbol{G})]| \le C_D\nu^5 e_K\log^{3\eta+3/2}(N)V_{\boldsymbol{A}}.$$

Combining the previously established moment and deterministic bounds, we obtain

$$\|\mathrm{Tr}(\boldsymbol{AGX}_k^\top\boldsymbol{X}_k\boldsymbol{G}) - \boldsymbol{e}_k^\top\boldsymbol{L}^{-1}\tau_{\boldsymbol{A}}\|_{L^q} \le C_q\nu^{14}e_K V_{\boldsymbol{A}},$$

and, with probability at least $1 - \sum_i n_i^{-D}$,

$$|\mathrm{Tr}(\boldsymbol{AGX}_k^\top\boldsymbol{X}_k\boldsymbol{G}) - \boldsymbol{e}_k^\top\boldsymbol{L}^{-1}\tau_{\boldsymbol{A}}| \le C_D e_K\left(\nu^{14} + \nu^5\log^{3\eta+3/2}(N)\right)V_{\boldsymbol{A}}.$$

F.6 Technical lemma

Lemma 3. *Under Assumption 1, define $\boldsymbol{M}_i = \boldsymbol{C}_i^{1/2}\boldsymbol{G}\boldsymbol{C}_i^{1/2}$. For any $i \in [K]$, we have, with probability at least $1 - n_i^{-D}$,*

$$\|\boldsymbol{M}_i\|_{op} \le C_D\frac{\nu_{\lambda,i}(n_i)}{n_i}, \quad \|\boldsymbol{M}_i\|_F \le C_D\frac{\nu_{\lambda,i}(n_i)}{\sqrt{n_i}}, \quad \mathrm{Tr}(\boldsymbol{M}_i) \le C_D\nu_{\lambda,i}(n_i).$$

Proof of Lemma 3. The result directly follows from (Misiakiewicz & Saeed, 2024, Proof of Lemma 1), by noting that

$$\boldsymbol{M}_i = \boldsymbol{C}_i^{1/2}\left(\sum_{i=1}^{K}\boldsymbol{X}_i^\top\boldsymbol{X}_i+\lambda\right)^{-1}\boldsymbol{C}_i^{1/2} \preceq \boldsymbol{C}_i^{1/2}\left(\boldsymbol{X}_i^\top\boldsymbol{X}_i+\lambda\right)^{-1}\boldsymbol{C}_i^{1/2}.$$

□

Next, we state (Misiakiewicz & Saeed, 2024, Lemma 2).

Lemma 4. *Assume the features $\boldsymbol{x}_1, \boldsymbol{x}_2$ are independent and satisfy Assumption 1. Then for all PSD matrices $\boldsymbol{B}$ independent of $\boldsymbol{x}_1, \boldsymbol{x}_2$, with probability at least $1-n^{-D}$ over the randomness of $\boldsymbol{x}_1, \boldsymbol{x}_2$, we have*

$$|\boldsymbol{x}_1^\top\boldsymbol{B}\boldsymbol{x}_1 - \mathrm{Tr}(\boldsymbol{C}\boldsymbol{B})| \le C_{x,D}\cdot\log^\beta(n)\|\boldsymbol{C}^{1/2}\boldsymbol{B}\boldsymbol{C}^{1/2}\|_F,$$
$$|\boldsymbol{x}_1^\top\boldsymbol{B}\boldsymbol{x}_2| \le C_{x,D}\cdot\log^\beta(n)\left(\|\boldsymbol{C}^{1/2}\boldsymbol{B}\boldsymbol{C}^{1/2}\|_F + \|\boldsymbol{C}^{1/2}\boldsymbol{B}\boldsymbol{C}\boldsymbol{B}\boldsymbol{C}^{1/2}\|_F^{1/2}\right).$$

Moreover, for all $q\in\mathbb{N}$, we have

$$\mathbb{E}_{\boldsymbol{x}_1}\left[|\boldsymbol{x}_1^\top\boldsymbol{B}\boldsymbol{x}_1 - \mathrm{Tr}(\boldsymbol{C}\boldsymbol{B})|^q\right]^{1/q} \le C_{x,q}\cdot\|\boldsymbol{C}^{1/2}\boldsymbol{B}\boldsymbol{C}^{1/2}\|_F,$$
$$\mathbb{E}_{\boldsymbol{x}_1,\boldsymbol{x}_2}\left[|\boldsymbol{x}_1^\top\boldsymbol{B}\boldsymbol{x}_2|^q\right]^{1/q} \le C_{x,q}\left(\|\boldsymbol{C}^{1/2}\boldsymbol{B}\boldsymbol{C}^{1/2}\|_F + \|\boldsymbol{C}^{1/2}\boldsymbol{B}\boldsymbol{C}\boldsymbol{B}\boldsymbol{C}^{1/2}\|_F^{1/2}\right).$$

Proof. The proof is the same as in (Misiakiewicz & Saeed, 2024, Lemma 2). □

Lemma 5. *Under Assumption 1 and 2, let $\boldsymbol{G}_-$ be the resolvent after at most 2 features respectively from $\boldsymbol{X}_\ell$ and $\boldsymbol{X}_j$ have been removed (here ℓ and j could be the same). Then, we have*

$$\mathrm{Tr}(\boldsymbol{C}_i\overline{\boldsymbol{G}}) \le C\nu, \qquad \|\overline{\boldsymbol{G}}^{1/2}\boldsymbol{C}_i\overline{\boldsymbol{G}}^{1/2}\|_{\mathrm{op}} \le \frac{C\nu}{n_i},$$

$$\|\mathrm{Tr}(\boldsymbol{C}_i\boldsymbol{G}_-) - \mathrm{Tr}(\boldsymbol{C}_i\overline{\boldsymbol{G}})\|_{L^q} \le C_q\left(\nu^2\left(\frac{1}{n_\ell}+\frac{1}{n_j}\right) + \nu^6\log^{C+1/2}(eN)\sum_{k=1}^{K}n_k^{-1/2}\right)\mathrm{Tr}(\boldsymbol{C}_i\overline{\boldsymbol{G}}).$$

Moreover, for every fixed $M>0$, with probability at least $1-\sum_{j=1}^{K}n_j^{-M}$,

$$\|\overline{\boldsymbol{G}}^{-1/2}\boldsymbol{G}_-\overline{\boldsymbol{G}}^{-1/2}\|_{\mathrm{op}} \le C_M\nu, \qquad \|\boldsymbol{C}_i^{1/2}\boldsymbol{G}_-\boldsymbol{C}_i^{1/2}\|_{\mathrm{op}} \le C_M\frac{\nu}{n_i}.$$

Therefore, for every deterministic PSD matrix $\boldsymbol{A}$,

$$\mathrm{Tr}(\boldsymbol{A}\boldsymbol{G}_-) \le C_M\nu\mathrm{Tr}(\boldsymbol{A}\overline{\boldsymbol{G}}) \qquad \textit{with probability at least } 1-\sum_{j=1}^{K}n_j^{-M}.$$

Proof. Write

$$\nu = \nu_{\lambda,*}, \qquad e_K = \left(\sum_{k=1}^{K}n_k^{-1}\right)^{1/2}, \qquad t_i = \mathrm{Tr}(\boldsymbol{C}_i\overline{\boldsymbol{G}}).$$

First, setting $h_i = d\wedge\lfloor\eta n_i\rfloor$, we have

$$\begin{aligned} n_i &= \mu_i + \mathrm{Tr}(\mu_i\boldsymbol{C}_i\overline{\boldsymbol{G}}) \\ &\le \mu_i + \sum_{\ell=1}^{d}\frac{\mu_i\xi_\ell(\boldsymbol{C}_i)}{\mu_i\xi_\ell(\boldsymbol{C}_i)+\lambda} \\ &\le \mu_i + h_i + \frac{\mu_i}{\lambda}\sum_{\ell>h_i}\xi_\ell(\boldsymbol{C}_i) \le \nu\mu_i + \eta n_i. \end{aligned}$$

Consequently,

$$\mu_i \ge \frac{1-\eta}{\nu}n_i, \qquad t_i \le \frac{\nu}{1-\eta}-1 \le C\nu, \qquad \|\overline{\boldsymbol{G}}^{1/2}\boldsymbol{C}_i\overline{\boldsymbol{G}}^{1/2}\|_{\mathrm{op}} \le \mu_i^{-1} \le \frac{C\nu}{n_i}.$$

We next prove the random resolvent bounds. Fix $|\mathcal{S}| \le 2$ and write $\boldsymbol{G}_{-\mathcal{S}} := \boldsymbol{G}_-$ to be explicit on the features removed from the resolvent. Since

$$\boldsymbol{G}_- \preceq \left(\lambda \boldsymbol{I} + \sum_{(i,r)\notin\mathcal{S}} \boldsymbol{x}_{i,r}\boldsymbol{x}_{i,r}^\top \right)^{-1},$$

the single-dataset moment estimates in (Misiakiewicz & Saeed, 2024, Lemma 4(b)), applied to the remaining observations, give

$$\left\| \|\boldsymbol{C}_i^{1/2}\boldsymbol{G}_-\boldsymbol{C}_i^{1/2}\|_{\mathrm{op}} \right\|_{L^q} \le C_q \frac{\nu}{n_i}, \qquad \|\mathrm{Tr}(\boldsymbol{C}_i\boldsymbol{G}_-)\|_{L^q} \le C_q\nu.$$

Here the same estimates apply after at most two deletions under our standing assumptions.

Moreover, $\lambda\boldsymbol{G}_- \preceq \boldsymbol{I}$ and $\mu_i \le n_i$. Thus,

$$\begin{aligned}
\|\overline{\boldsymbol{G}}^{-1/2}\boldsymbol{G}_-\overline{\boldsymbol{G}}^{-1/2}\|_{\mathrm{op}} &= \|\boldsymbol{G}_-^{1/2}\overline{\boldsymbol{G}}^{-1}\boldsymbol{G}_-^{1/2}\|_{\mathrm{op}} \\
&\le 1 + \sum_{i=1}^K \mu_i \|\boldsymbol{G}_-^{1/2}\boldsymbol{C}_i\boldsymbol{G}_-^{1/2}\|_{\mathrm{op}} \\
&= 1 + \sum_{i=1}^K \mu_i \|\boldsymbol{C}_i^{1/2}\boldsymbol{G}_-\boldsymbol{C}_i^{1/2}\|_{\mathrm{op}}.
\end{aligned}$$

Taking L^q norms proves

$$\left\| \|\overline{\boldsymbol{G}}^{-1/2}\boldsymbol{G}_-\overline{\boldsymbol{G}}^{-1/2}\|_{\mathrm{op}} \right\|_{L^q} \le C_q\nu,$$

where the constant may depend on K.

Now add back one removed observation $\boldsymbol{u} = \boldsymbol{x}_{j,r}$, and set $\boldsymbol{G}_+^{-1} = \boldsymbol{G}_-^{-1} + \boldsymbol{u}\boldsymbol{u}^\top$. By Sherman–Morrison,

$$0 \le \mathrm{Tr}(\boldsymbol{C}_i(\boldsymbol{G}_- - \boldsymbol{G}_+)) = \frac{\boldsymbol{u}^\top\boldsymbol{G}_-\boldsymbol{C}_i\boldsymbol{G}_-\boldsymbol{u}}{1+\boldsymbol{u}^\top\boldsymbol{G}_-\boldsymbol{u}} \le \boldsymbol{u}^\top\boldsymbol{G}_-\boldsymbol{C}_i\boldsymbol{G}_-\boldsymbol{u}.$$

Conditional on $\boldsymbol{G}_-$, the quadratic-form moment bound gives

$$\|\mathrm{Tr}(\boldsymbol{C}_i(\boldsymbol{G}_- - \boldsymbol{G}_+))\|_{L^q(\boldsymbol{u}|\boldsymbol{G}_-)} \le C_q\mathrm{Tr}(\boldsymbol{C}_j\boldsymbol{G}_-\boldsymbol{C}_i\boldsymbol{G}_-).$$

Also,

$$\mathrm{Tr}(\boldsymbol{C}_j\boldsymbol{G}_-\boldsymbol{C}_i\boldsymbol{G}_-) \le \|\boldsymbol{C}_j^{1/2}\boldsymbol{G}_-\boldsymbol{C}_j^{1/2}\|_{\mathrm{op}}\|\overline{\boldsymbol{G}}^{-1/2}\boldsymbol{G}_-\overline{\boldsymbol{G}}^{-1/2}\|_{\mathrm{op}}t_i.$$

Taking L^q norms, applying Hölder and the preceding bounds at exponent $2q$, we obtain

$$\|\mathrm{Tr}(\boldsymbol{C}_i(\boldsymbol{G}_- - \boldsymbol{G}_+))\|_{L^q} \le C_q\frac{\nu^2}{n_j}t_i.$$

Repeating this for each removed observation yields

$$\|\mathrm{Tr}(\boldsymbol{C}_i(\boldsymbol{G}_{-\mathcal{S}} - \boldsymbol{G}))\|_{L^q} \le C_q\nu^2 \left(\sum_{(j,r)\in\mathcal{S}} \frac{1}{n_j} \right) t_i.$$

Finally, the previously established first-order moment bound gives

$$\|\mathrm{Tr}(\boldsymbol{C}_i(\boldsymbol{G} - \overline{\boldsymbol{G}}))\|_{L^q} \le C_q\nu^6 e_K t_i.$$

The triangle inequality therefore implies

$$\|\mathrm{Tr}(\boldsymbol{C}_i\boldsymbol{G}_-) - \mathrm{Tr}(\boldsymbol{A}\overline{\boldsymbol{G}})\|_{L^q} \le C_q \left(\nu^2 \sum_{(j,r)\in\mathcal{S}} \frac{1}{n_j} + \nu^6 e_K \right) \mathrm{Tr}(\boldsymbol{A}\overline{\boldsymbol{G}}).$$

For the probabilistic bound, applying the single-dataset resolvent estimates in Lemma 3 to the retained observations and taking the union bounds over the datasets, we get

$$\|\boldsymbol{C}_i^{1/2}\boldsymbol{G}_-\boldsymbol{C}_i^{1/2}\|_{\mathrm{op}} \le C_M\frac{\nu}{n_i}.$$

Since $\lambda \boldsymbol{G}_- \preceq \boldsymbol{I}$ and $\mu_i \le n_i$, we get

$$\begin{aligned}\|\overline{\boldsymbol{G}}^{-1/2}\boldsymbol{G}_-\overline{\boldsymbol{G}}^{-1/2}\|_{\mathrm{op}} &= \|\boldsymbol{G}_-^{1/2}\overline{\boldsymbol{G}}^{-1}\boldsymbol{G}_-^{1/2}\|_{\mathrm{op}} \\ &\le 1 + \sum_i \mu_i \|\boldsymbol{C}_i^{1/2}\boldsymbol{G}_-\boldsymbol{C}_i^{1/2}\|_{\mathrm{op}} \le C_D\nu.\end{aligned}$$

The trace bound then follows immediately. □

Lemma 6. *Under the assumptions of Lemmas 4 and 5, let $\boldsymbol{G}_-$ denote the resolvent after a fixed set of at most two observations has been removed. For deterministic PSD matrices $\boldsymbol{A}, \boldsymbol{B}$ and every fixed $q \ge 1$,*

$$\|\mathrm{Tr}(\boldsymbol{A}\boldsymbol{G}_-\boldsymbol{B}\boldsymbol{G}_-)\|_{L^q} \le C_q\left(\mathrm{Tr}(\boldsymbol{A}\overline{\boldsymbol{G}}\boldsymbol{B}\overline{\boldsymbol{G}}) + \nu^3\|\overline{\boldsymbol{G}}^{1/2}\boldsymbol{B}\overline{\boldsymbol{G}}^{1/2}\|_{\mathrm{op}}W_{\boldsymbol{A}}\right),$$

where

$$t_i = \mathrm{Tr}(\boldsymbol{C}_i\overline{\boldsymbol{G}}), \qquad \tau_{\boldsymbol{A}}[i] = \mathrm{Tr}(\boldsymbol{A}\overline{\boldsymbol{G}}\boldsymbol{C}_i\overline{\boldsymbol{G}}), \qquad W_{\boldsymbol{A}} = \sum_{i=1}^K n_i t_i \tau_{\boldsymbol{A}}[i].$$

Moreover, for every fixed $D > 0$, with probability at least $1 - \sum_{i=1}^K n_i^{-D}$,

$$\mathrm{Tr}(\boldsymbol{A}\boldsymbol{G}_-\boldsymbol{B}\boldsymbol{G}_-) \le C_D \log^{2\beta+1}(N)\left(\mathrm{Tr}(\boldsymbol{A}\overline{\boldsymbol{G}}\boldsymbol{B}\overline{\boldsymbol{G}}) + \nu^3\|\overline{\boldsymbol{G}}^{1/2}\boldsymbol{B}\overline{\boldsymbol{G}}^{1/2}\|_{\mathrm{op}}W_{\boldsymbol{A}}\right).$$

Proof. We adopt the same notation as in the proof of Lemma 5. Defining

$$\boldsymbol{S}_- := \boldsymbol{G}_-^{-1} - \lambda\boldsymbol{I} = \sum_{(i,r)\notin\mathcal{S}} \boldsymbol{x}_{i,r}\boldsymbol{x}_{i,r}^\top,$$

we can write

$$\mathrm{Tr}(\boldsymbol{A}\boldsymbol{G}_-\boldsymbol{B}\boldsymbol{G}_-) = \|\boldsymbol{B}^{1/2}(\overline{\boldsymbol{G}}\boldsymbol{A}^{1/2} + \boldsymbol{G}_-\overline{\boldsymbol{G}}^{-1/2}\boldsymbol{Y})\|_F^2,$$

where

$$\boldsymbol{Y} = \overline{\boldsymbol{G}}^{1/2}\left(\sum_{k=1}^K \mu_k\boldsymbol{C}_k - \boldsymbol{S}_-\right)\overline{\boldsymbol{G}}\boldsymbol{A}^{1/2}.$$

Next, we want to bound $\boldsymbol{Y}$. From (Misiakiewicz & Saeed, 2024, Lemma 2), we have for all (i,r),

$$\begin{aligned}\left\|\|\overline{\boldsymbol{G}}^{1/2}\boldsymbol{x}_{i,r}\boldsymbol{x}_{i,r}^\top\overline{\boldsymbol{G}}\boldsymbol{A}^{1/2}\|_F\right\|_{L^p} &\le \left\|\boldsymbol{x}_{i,r}^\top\overline{\boldsymbol{G}}\boldsymbol{x}_{i,r}\right\|_{L^{2p}}^{1/2} \cdot \left\|\boldsymbol{x}_{i,r}\overline{\boldsymbol{G}}\boldsymbol{A}\overline{\boldsymbol{G}}\boldsymbol{u}\right\|_{L^{2p}}^{1/2} \\ &\le C_p\sqrt{t_i\tau_{\boldsymbol{A}}[i]}.\end{aligned}$$

Therefore, we also have

$$\left\|\|\overline{\boldsymbol{G}}^{1/2}(\boldsymbol{x}_{i,r}\boldsymbol{x}_{i,r}^\top - \boldsymbol{C}_i)\overline{\boldsymbol{G}}\boldsymbol{A}^{1/2}\|_F\right\|_{L^p} \le C_p\sqrt{t_i\tau_{\boldsymbol{A}}[i]}.$$

Since we want to obtain a bound on $\boldsymbol{Y}$, we first consider controlling

$$\boldsymbol{Y} - \mathbb{E}[\boldsymbol{Y}] = -\sum_{(i,r)} \overline{\boldsymbol{G}}^{1/2}(\boldsymbol{x}_{i,r}\boldsymbol{x}_{i,r}^\top - \boldsymbol{C}_i)\overline{\boldsymbol{G}}\boldsymbol{A}^{1/2} =: -\sum_{(i,r)} \boldsymbol{Z}_{i,r.}$$

By symmetrization and the Hilbert-space Khintchine–Kahane inequality (Ledoux & Talagrand, 1991, Chapter 4 and Section 6.1), applied to the independent centered matrices $\boldsymbol{Z}_{i,r}$ with the Frobenius norm, , for $p \ge 2$,

$$\begin{aligned}\left\|\left\|\sum_{(i,r)}\boldsymbol{Z}_{i,r}\right\|_F\right\|_{L^p} &\le C\sqrt{p}\left(\sum_{(i,r)}\|\|\boldsymbol{Z}_{i,r}\|_F\|_{L^p}^2\right)^{1/2} \\ &\le C_p\left(\sum_{i=1}^K n_i t_i \tau_{\boldsymbol{A}}[i]\right)^{1/2} \\ &= C_p\sqrt{W_{\boldsymbol{A}}}.\end{aligned}$$

Next, we need to bound $\mathbb{E}[\boldsymbol{Y}]$. Define d_i as the number of observations deleted from $\boldsymbol{X}_i$. By assumption, we have $d_i \in \{0, 1, 2\}$ for all $i \in [K]$. We can write $\mathbb{E}[\boldsymbol{Y}]$ as

$$\mathbb{E}[\boldsymbol{Y}] = -\overline{\boldsymbol{G}}^{1/2} \sum_{i=1}^{K} (n_i - \mu_i) \boldsymbol{C}_i \overline{\boldsymbol{G}} \boldsymbol{A}^{1/2} + \sum_{i=1}^{K} d_i \overline{\boldsymbol{G}}^{1/2} \boldsymbol{C}_i \overline{\boldsymbol{G}} \boldsymbol{A}^{1/2}.$$

Define

$$\boldsymbol{H}_i = \overline{\boldsymbol{G}}^{1/2} \boldsymbol{C}_i \overline{\boldsymbol{G}}^{1/2}, \qquad \boldsymbol{R} = \overline{\boldsymbol{G}}^{1/2} \boldsymbol{A}^{1/2}, \qquad \boldsymbol{P} = \sum_{i=1}^{K} (n_i - \mu_i) \boldsymbol{H}_i.$$

Then, we can write

$$-\overline{\boldsymbol{G}}^{1/2} \sum_{i=1}^{K} (n_i - \mu_i) \boldsymbol{C}_i \overline{\boldsymbol{G}} \boldsymbol{A}^{1/2} = -\boldsymbol{P}\boldsymbol{R}.$$

Since $n_i - \mu_i = \mu_i t_i$ and

$$\sum_{i=1}^{K} \mu_i \boldsymbol{H}_i = \boldsymbol{I} - \lambda \overline{\boldsymbol{G}} \preceq \boldsymbol{I},$$

we have

$$0 \preceq \boldsymbol{P} = \sum_{i=1}^{K} \mu_i t_i \boldsymbol{H}_i \preceq \max_i t_i \, \boldsymbol{I} \preceq C\nu \boldsymbol{I}.$$

Consequently,

$$\begin{aligned} \|\boldsymbol{P}\boldsymbol{R}\|_F^2 &\leq C\nu \mathrm{Tr}(\boldsymbol{R}^\top \boldsymbol{P} \boldsymbol{R}) \\ &= C\nu \sum_{i=1}^{K} \mu_i t_i \tau_{\boldsymbol{A}}[i] \\ &\leq C\nu W_{\boldsymbol{A}}. \end{aligned}$$

Let d_i be the number of observations removed from dataset i. Then

$$0 \leq d_i \leq n_i, \qquad \sum_{i=1}^{K} d_i \leq 2,$$

and we can write

$$\sum_{i=1}^{K} d_i \overline{\boldsymbol{G}}^{1/2} \boldsymbol{C}_i \overline{\boldsymbol{G}} \boldsymbol{A}^{1/2} = \sum_{i=1}^{K} d_i \boldsymbol{H}_i \boldsymbol{R}.$$

For each i,

$$\|\boldsymbol{H}_i \boldsymbol{R}\|_F^2 = \mathrm{Tr}(\boldsymbol{R}^\top \boldsymbol{H}_i^2 \boldsymbol{R}) \leq t_i \mathrm{Tr}(\boldsymbol{R}^\top \boldsymbol{H}_i \boldsymbol{R}) = t_i \tau_{\boldsymbol{A}}[i].$$

Since there are at most two observations being deleted, using the inequality

$$\|\boldsymbol{H}_i \boldsymbol{R} + \boldsymbol{H}_j \boldsymbol{R}\|_F^2 \leq 2\|\boldsymbol{H}_i \boldsymbol{R}\|_F^2 + 2\|\boldsymbol{H}_j \boldsymbol{R}\|_F^2,$$

we have

$$\left\| \sum_{i=1}^{K} d_i \boldsymbol{H}_i \boldsymbol{R} \right\|_F^2 \leq 2 \sum_{i=1}^{K} d_i t_i \tau_{\boldsymbol{A}}[i] \leq 2W_{\boldsymbol{A}}.$$

Combined with previous results, we get

$$\|\mathbb{E}[\boldsymbol{Y}]\|_F^2 \leq C\nu W_{\boldsymbol{A}}.$$

Therefore, we bound the L^p norm of $\|\boldsymbol{Y}\|_F$ as

$$\big\| \|\boldsymbol{Y}\|_F \big\|_{L^p}^2 \leq \left(\big\| \|\boldsymbol{Y} - \mathbb{E}[\boldsymbol{Y}]\|_F \big\|_{L^p} + \big\| \|\mathbb{E}[\boldsymbol{Y}]\|_F \big\|_{L^p} \right)^2 \leq C_p \nu W_{\boldsymbol{A}}.$$

With this bound on $\boldsymbol{Y}$, we can control $\mathrm{Tr}(\boldsymbol{A}\boldsymbol{G}_-\boldsymbol{B}\boldsymbol{G}_-)$. A direct computation gives

$$\begin{aligned} \mathrm{Tr}(\boldsymbol{A}\boldsymbol{G}_-\boldsymbol{B}\boldsymbol{G}_-) &= \|\boldsymbol{B}^{1/2}\boldsymbol{G}_-\boldsymbol{A}^{1/2}\|_F^2 \\ &\leq 2\|\boldsymbol{B}^{1/2}\overline{\boldsymbol{G}}\boldsymbol{A}^{1/2}\|_F^2 + 2\|\boldsymbol{B}^{1/2}\boldsymbol{G}_-\overline{\boldsymbol{G}}^{-1/2}\|_{\mathrm{op}}^2 \|\boldsymbol{Y}\|_F^2 \\ &\leq 2\mathrm{Tr}(\boldsymbol{A}\overline{\boldsymbol{G}}\boldsymbol{B}\overline{\boldsymbol{G}}) + 2\|\overline{\boldsymbol{G}}^{1/2}\boldsymbol{B}\overline{\boldsymbol{G}}^{1/2}\|_{\mathrm{op}} \|\overline{\boldsymbol{G}}^{-1/2}\boldsymbol{G}_-\overline{\boldsymbol{G}}^{-1/2}\|_{\mathrm{op}}^2 \|\boldsymbol{Y}\|_F^2. \end{aligned}$$

Finally, we apply Hölder's inequality on the last term and get

$$\|\|\overline{\boldsymbol{G}}^{-1/2}\boldsymbol{G}_{-}\overline{\boldsymbol{G}}^{-1/2}\|_{\mathrm{op}}^2\|\boldsymbol{Y}\|_F^2\|_{L^q} \le \|\|\overline{\boldsymbol{G}}^{-1/2}\boldsymbol{G}_{-}\overline{\boldsymbol{G}}^{-1/2}\|_{\mathrm{op}}\|_{L^{4q}}^2\|\|\boldsymbol{Y}\|_F\|_{L^{4q}}^2 \le C_q\nu^3 W_{\boldsymbol{A}}.$$

For the probabilistic result the idea is similar to the martingale parts of the functionals. First, we have

$$\mathbb{P}(\|\boldsymbol{Z}_{i,r}\|_F \ge R_i) \le C_M N^{-M}, \qquad R_i = C_M \log^{\beta}(N)\sqrt{t_i\tau_{\boldsymbol{A}}[i]}.$$

Set $\widetilde{\boldsymbol{Z}}_{i,r} = \boldsymbol{Z}_{i,r}\mathbf{1}_{\{\|Z_{i,r}\|_F \le L_i\}}$. Since $\mathbb{E}[\boldsymbol{Z}_{i,r}] = 0$,

$$\left\|\sum_{i,r}\mathbb{E}[\widetilde{\boldsymbol{Z}}_{i,r}]\right\|_F \le C_M N^{-M/2}\sum_{i=1}^{K} n_i\sqrt{t_i\tau_{\boldsymbol{A}}[i]} \le C_M N^{(1-M)/2}\sqrt{W_{\boldsymbol{A}}} \le C_M\sqrt{W_A}$$

for M sufficiently large. Bounded-differences concentration then gives with probability at least $1 - N^{-M}$,

$$\left\|\sum_{i,r}(\widetilde{\boldsymbol{Z}}_{i,r} - \mathbb{E}[\widetilde{\boldsymbol{Z}}_{i,r}])\right\|_F \le C_M\sqrt{\log(N)\sum_{i=1}^{K} n_i R_i^2} \le C_M \log^{\beta+1/2}(N)\sqrt{W_{\boldsymbol{A}}}.$$

Applying union bounds and choosing M sufficiently large, we have with probability at least $1 - \frac{1}{2}\sum_{i=1}^{K} n_i^D$,

$$\|\boldsymbol{Y} - \mathbb{E}[\boldsymbol{Y}]\|_F \le C_D \log^{\beta+1/2}(N)\sqrt{W_{\boldsymbol{A}}}.$$

Along with $\|\mathbb{E}[\boldsymbol{Y}]\|_F^2 \le C\nu W_{\boldsymbol{A}}$, we get

$$\|\boldsymbol{Y}\|_F^2 \le C_D\nu \log^{2\beta+1}(N) W_{\boldsymbol{A}}.$$

Therefore, we have with probability at least $1 - \sum_{i=1}^{K} n_i^{-D}$,

$$\mathrm{Tr}(\boldsymbol{A}\boldsymbol{G}_{-}\boldsymbol{B}\boldsymbol{G}_{-}) \le C_D \log^{2\beta+1}(N)\left(\mathrm{Tr}(\boldsymbol{A}\overline{\boldsymbol{G}}\boldsymbol{B}\overline{\boldsymbol{G}}) + \nu^3\|\overline{\boldsymbol{G}}^{1/2}\boldsymbol{B}\overline{\boldsymbol{G}}^{1/2}\|_{\mathrm{op}} W_{\boldsymbol{A}}\right).$$

□

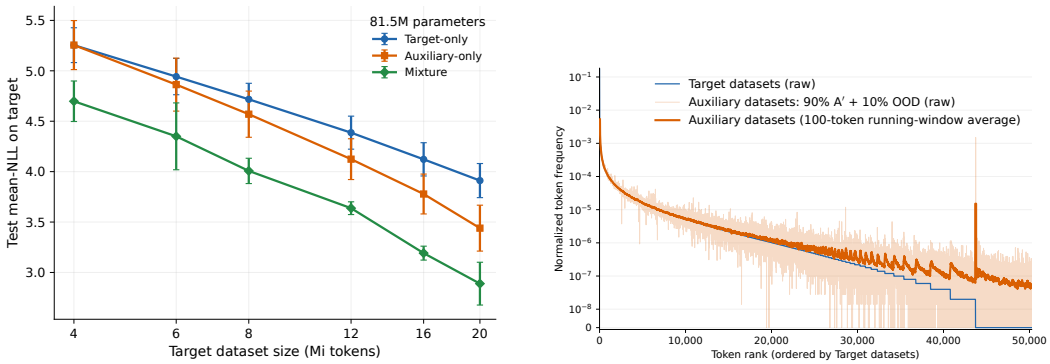
G Omitted parts in Section 6

G.1 Details of numerical experiments

Datasets and data preprocessing. We use SlimPajama (Cerebras Systems, 2023) as the pretraining corpus. The target dataset consists of documents from the `StackExchange` domain. We consider the following constructions of the auxiliary dataset:

- In the case where the mixture improves the scaling law compared to using each datasets individually, we exclude `StackExchange` and combine `GitHub`, `C4`, `arXiv`, `Common Crawl`, `Books`, and `Wikipedia` with weights 0.637, 0.099, 0.095, 0.076, 0.048, and 0.045, respectively, following Fan et al. (2024).
- In the case where the auxiliary distribution is close to the target distribution, we combine 90% independent `StackExchange` data with 10% of the six-domain OOD mixture defined above.

Before tokenization, we normalize each document using Unicode NFC, canonicalize newline characters, and remove leading and trailing whitespace. We then assign documents to the training, validation, and test sets using normalized-content hashes, with a $98\%/1\%/1\%$ split. This document-level splitting ensures that documents with the same normalized content cannot occur in different splits. We tokenize the documents using the GPT-2 byte-level BPE tokenizer with a vocabulary size of $50{,}257$ and append an end-of-text token to every document. The resulting token streams are packed into non-overlapping blocks of 512 tokens. For data preprocessing, we discard incomplete trailing blocks and low-diversity blocks for which fewer than 10% of the tokens are unique. All dataset sizes reported in our experiments refer to prediction tokens remaining after this preprocessing.



(a) $\gamma_2 = 1.25$

(b) Token-frequency distributions

Figure 4: Target-domain test negative log-likelihood (NLL) scaling of an 81.5M-parameter GPT-2-style language model trained from scratch on SlimPajama. Panel (a) reports mean NLL on a fixed target-domain test set as a function of the reference target budget n. We use $n \in \{4, 6, 8, 12, 16, 20\}$ Mi, except for the $\gamma_2 = 1.75$ experiment that uses $n \leq 16$ Mi. Panel (b) compares the normalized target and auxiliary token frequencies. Token types are ordered by decreasing target-domain frequency. The light auxiliary curve shows the raw frequencies, and the darker, thicker curve shows their 100-token running-window average.

Model and optimization. All reported results use an 81.5M-parameter decoder-only transformer with 6 layers, hidden width 768, 12 attention heads, feed-forward width 3072, context length 512, and a 50,257-token GPT-2 vocabulary. Input and output token embeddings are tied. We use three seeds . We use AdamW with peak learning rate 6×10^{-4}, $\beta = (0.9, 0.95)$, $\epsilon = 10^{-8}$, weight decay 0.1, gradient clipping at 1, and a global batch of 32,768 prediction tokens. We use the following learning rate scheduler:

$$\eta(t) = 6 \times 10^{-4} \min\left(1, \frac{t}{262144}\right) \left(1 + \frac{\max(t - 262144, 0)}{16777216}\right)^{-1/2},$$

where t is the number of processed prediction tokens before the update. Training uses PyTorch automatic mixed precision with BF16 for eligible operations. Model parameters and AdamW optimizer states are maintained in FP32, and TF32 is enabled for the remaining FP32 matrix-multiplication operations.

G.2 Experiments on datasets with close distributions

Figure 4 shows that when the token frequency of the two datasets is close, for $\gamma_2 = 1.25$, the data mixture does not improve the scaling law compared to using one of the individual datasets. Table 1 shows that this phenomenon occurs across different scales γ_2. This qualitatively supports our theoretical finding that data mixtures improve scaling only when spectral decay rates are sufficiently different.

Table 1: Linear OLS scaling slopes on the target test set for the 81.5M-parameter model over 4–20 M target datasets tokens. More-negative values indicate a faster decrease in target NLL, i.e. better scaling law. Each slope is obtained by fitting NLL $= \alpha + s \ln(n)$ over the six Target dataset sizes $n \in \{4, 6, 8, 12, 16, 20\}$ M within each seed and then averaging over three seeds. Improvement is $100\,(|s_{\mathrm{Mixture}}| - |s_X|)/|s_X|$, reported against target-only and auxiliary-only, respectively.

γ_2	Target-only	Auxiliary-only	Mixture	Mixture improvement (%) Target / Auxiliary
0.75	-0.8319	-0.6085	-0.8742	+5.1% / +43.7%
1	-0.8319	-0.8389	-1.0272	+23.5% / +22.4%
1.25	-0.8319	-1.1155	-1.1225	+34.9% / +0.6%
1.5	-0.8319	-1.4015	-1.2513	+50.4% / -10.7%
1.75	-0.8319	-1.5971	-1.3208	+58.8% / -17.3%
2	-0.8319	-1.7332	-1.4361	+72.6% / -17.1%