HIGH-DIMENSIONAL ASYMPTOTICS AND DATASET SELECTION FOR PRIVATE TRANSFER LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

To commit to buying external data or participate in collaborative learning, one must decide whether the additional data will improve prediction enough to justify the cost. This comes with several challenges: *(i)* the decision often relies only on aggregated statistics available publicly, rather than individual-level data; *(ii)* covariate and model shifts can induce negative transfer, so the additional data deteriorates rather than improves performance; *(iii)* if the data is sensitive, its privatization requires the injection of noise, which can also offset the benefit of a larger sample size. In this paper, we model the problem of dataset selection through high-dimensional regression with multiple heterogeneous sources and a weighted ridge estimator. Our approach uses only summary statistics and it gives privacy guarantees either on labels only or jointly on features and labels, in terms of ρ-zero-concentrated differential privacy. The main technical contribution is a deterministic equivalent of the test error, which captures the interactions between sample size, covariance structure, model shift, regularization and privacy noise. Our theory allows to optimize hyperparameters (weights and ridge regularizers) and, more broadly, to decide when private external datasets are useful without accessing the data itself but only relying on population-level quantities. This provides a theoretically tractable foundation for private transfer learning, which we support via experiments on both synthetic and real-world datasets.

1 INTRODUCTION

Should one buy external data or participate in collaborative learning? The answer depends on whether the extra data improves prediction enough to justify its cost. When handling sensitive data, this decision may rely on aggregate statistics available before access to individual records. For example, the UK Biobank Data Showcase provides public summaries to assess relevance before applying for access (Sudlow et al., 2015) and the US Census Bureau releases statistics aggregated at county or state levels (Bureau, 2026; Abowd et al., 2022). Collaboration between hospitals raises the same question for heterogeneous, sensitive datasets (Ogier du Terrail et al., 2022): are the efforts and risks worth the accuracy gain?

Indeed, incorporating source data does not necessarily improve target prediction. Differences in feature distributions (covariate shift, (Sugiyama et al., 2007; Ma et al., 2023)) and in the relationship between features and labels (model shift, (Yang et al., 2025; Li et al., 2022)) can limit the benefits of transfer and, depending on how source data are incorporated, lead to negative transfer (Pan & Yang, 2009; Wang et al., 2019). To understand when transfer remains beneficial despite these shifts, recent work has developed precise characterizations of the prediction risk for high-dimensional linear regression (Song et al., 2024; Yang et al., 2025).

Differential privacy adds another drawback: the noise injected to protect individuals can offset the benefit of additional samples (Dwork & Roth, 2014). Thus, a rich line of theoretical work has quantified privacy-accuracy trade-offs in empirical risk minimization (Chaudhuri et al., 2011; Bassily et al., 2014) and stochastic convex optimization (Bassily et al., 2019), alongside minimax guarantees for statistical estimation (Duchi et al., 2018; Cai et al., 2021). In collaborative settings, unequal privacy noise levels across sources affect how their contributions should be combined (Liu et al., 2021), and recent work has characterized the accuracy achievable under privacy constraints in federated transfer learning with heterogeneous sources (Li et al., 2024). We also note that, although estimating

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.

the gains from collaboration has been widely studied in federated learning (McMahan et al., 2017; Sattler et al., 2020), the cost of participation motivates assessing the value of collaboration *before* committing. A central question therefore remains:

> *Can we predict whether and by how much heterogeneous private data will improve target prediction, before processing the data itself?*

We address this question for a high-dimensional linear model by analyzing weighted ridge regression with multiple heterogeneous and differentially private sources. This approach uses only summary statistics and it allows both model and covariance shift. Our main result is a deterministic equivalent for the risk which provides a principled basis to select sources, having access only to population-level quantities (as for, e.g., the US Census (Bureau, 2026; Abowd et al., 2022) or the UK Biobank (Sudlow et al., 2015)). More precisely, we summarize our contributions below:

- **Deterministic equivalent for the risk.** In the regime where sample sizes and feature dimension scale proportionally, we provide a precise characterization for the bias and variance of the weighted ridge estimator under label privatization (Theorem 4.1), and extend the analysis to joint feature and label privatization for Gaussian data (Corollary 4.4). The characterizations depend only on population quantities and they quantitatively capture covariate shift, model shift, regularization, and privacy noise.
- **Optimal tuning.** For a Gaussian prior, we identify source weights and ridge penalties that minimize the risk and show that the resulting ridge estimator is Bayes-optimal (Proposition 4.3).
- **Risk prediction and dataset selection in real-world scenarios.** We demonstrate the validity of our analysis beyond linear models. On three real-world datasets, we compare deterministic equivalent predictions with empirical changes in target prediction risk as auxiliary data is added (Figure 2), as well as the effectiveness of the resulting selection of datasets (Figure 3).

2 Related work

Transfer learning under covariate and model shifts. The usefulness of auxiliary data depends on differences between source and target distributions and on how information is shared across tasks. Statistical approaches provide estimation and prediction guarantees, together with methods for identifying and exploiting informative sources in high-dimensional regression (Li et al., 2022; Tian & Feng, 2023). The benefits of sharing information also depend on task similarity, data geometry, and the representations learned across tasks, as studied through joint training and representation transfer (Wu et al., 2020; Dhifallah & Lu, 2021; Sarao Mannelli et al., 2025; Gerace et al., 2022; Ingrosso et al., 2025). Related work examines how regularization and minimum-norm interpolation affect prediction under distribution shift (Patil et al., 2024; Mallinar et al., 2024). Closer to our setting, precise high-dimensional risk analyses show that adding source data can worsen target prediction, with positive and negative transfer depending on distribution shifts, noise levels, and sample sizes (Yang et al., 2025; Song et al., 2024). Source weighting and regularization can improve prediction by accounting for task relatedness and noise (Zhang & Li, 2023; Ilbert et al., 2024). Similar questions arise when combining real and surrogate data or selecting data under weak supervision, where risk analyses inform data weighting and selection (Jain et al., 2024; Kolossov et al., 2024; Rezaei et al., 2026). We highlight that none of these analyses address source-specific privacy constraints. We instead characterise how privatising labels, or both features and labels, changes target prediction risk under covariate and model shifts.

Data acquisition and differential privacy. Privacy-preserving methods have long addressed how to provide sufficient protection while retaining the benefits of collaboration, through the pricing of noisy query answers (Li et al., 2013) and data acquisition more generally. Fallah et al. (2022); Cummings et al. (2023) for instance examine the interplay between privacy costs and statistical accuracy. Privacy can undermine the benefits of collaboration, motivating the need for decision rules on whether or not joining forces, which has been studied from a federated learning perspective (Tsoy et al., 2024; McMahan et al., 2017; Sattler et al., 2020; Gu et al., 2025). Another related line of work combines public and private data, showing how a large private dataset can complement a small number of observations that do not require differential privacy protection (Alon et al., 2019; Bie et al., 2022; Ullah et al., 2024; Block et al., 2024). These approaches focus on how to combine

the available data, rather than predicting whether to collaborate, and often assume the same data distribution or have no theoretical guarantees for optimization (Sarpatwar et al., 2019). We instead predict the value of heterogeneous private datasets, accounting for covariate and model shifts as well as privacy noise. In collaborative settings, aggregate feature statistics are sometimes publicly available. When the features themselves are public and only the labels are sensitive, label differential privacy provides an appropriate model (Ghazi et al., 2021). This setting has also been studied for regression: Ghazi et al. (2023) derive optimal label-randomization mechanisms for a given prior distribution over labels and regression loss. We use this setting in Section 4.1 before considering the privatization of both features and labels in Section 4.2.

3 PRELIMINARIES

Linear model for transfer learning. We consider a linear regression model with C clients, where $C \geq 2$ does not grow with the dimension. Client $c \in \mathcal{C} \coloneqq \{1, \ldots, C\}$ observes a local dataset (X_c, y_c), where the design matrix $X_c \in \mathbb{R}^{n_c \times d}$ and the observation $y_c \in \mathbb{R}^{n_c}$ are given by

$$y_c = X_c \beta_c + \varepsilon_c,$$

where the entries of ε_c are independent (between themselves and w.r.t. $\{(X_c, \beta_c)_{c=1}^C\}$), have zero mean, variance σ_c^2 and uniformly bounded 8-th moments. The regression vectors $\beta_c \in \mathbb{R}^d$ and the feature distributions may vary across clients. In particular,

$$X_c = Z_c \Sigma_c^{1/2},$$

where $\Sigma_c \in \mathbb{R}^{d \times d}$ is the deterministic population covariance of client c, and the entries of $Z_c \in \mathbb{R}^{n_c \times d}$ are random. Without loss of generality, client 1 is the target client, and the prediction performance is evaluated on new observations drawn from the same distribution as client 1. The remaining clients provide auxiliary data that can be shared only in a privatized way, and they form the fixed set $\mathcal{A} \coloneqq \{2, \ldots, C\}$. We simplify the model shift notation by introducing

$$\beta \coloneqq \beta_1, \qquad \delta_c \coloneqq \beta_c - \beta,\ c \in \mathcal{A}, \qquad \theta \coloneqq [\beta \quad \delta_2 \quad \ldots \quad \delta_C]^\top \in \mathbb{R}^{Cd}.$$

Privatizing data. Differential privacy (DP) relies on the concept of adjacent datasets: two datasets D and D' are adjacent, written $D \sim D'$, if one is obtained from the other by replacing a single input-label pair. We quantify DP via ρ-zero-concentrated differential privacy (ρ-zCDP) which uses the α-Rényi divergence to measure the distinguishability between two probability distributions $P \ll Q$:

$$D_\alpha(P \| Q) \coloneqq \frac{1}{\alpha - 1} \log \int \left(\frac{\mathrm{d}P}{\mathrm{d}Q} \right)^\alpha \mathrm{d}Q, \qquad \alpha > 1.$$

Definition 3.1 (Zero-concentrated differential privacy (Bun & Steinke, 2016))**.** *A randomized algorithm $\mathcal{M}$ satisfies ρ-zero-concentrated differential privacy (ρ-zCDP) if, for every pair of adjacent datasets D, D' and every $\alpha > 1$,*

$$D_\alpha \left(\mathcal{M}(D) \| \mathcal{M}(D') \right) \leq \rho \alpha.$$

Guarantees for ρ-zCDP can be translated to the standard notion of (ϵ, δ)-differential privacy by taking $\epsilon = \rho + 2\sqrt{\rho \log(1/\delta)}$ (Bun & Steinke, 2016). We achieve ρ-zCDP via the Gaussian mechanism (Dwork & Roth, 2014) and provide the following two different privatization guarantees.

Label privatization: We privatize only the labels y_c by releasing

$$\tilde{y}_c = \mathrm{Clip}_{K_{y,c}}(y_c) + \xi_c, \qquad \xi_c \sim \mathcal{N}(0, \sigma_{y,\mathrm{DP},c}^2 I_{n_c}), \qquad \sigma_{y,\mathrm{DP},c}^2 = \frac{\min(\Delta_{y,c}^2, K_{y,c}^2)}{2\rho_{y,c}}. \tag{3.1}$$

Here, $\mathrm{Clip}_{K_{y,c}}(x) = x \min\{1, K_{y,c}/(2\,|x|)\}$ is the entry-wise clipping function with a constant $K_{y,c}$ that could depend on n_c and d, and $\Delta_{y,c} = \max_{y_c' \sim y_c''} \|y_c' - y_c''\|_2$ denotes the l_2-sensitivity of the labels of client c. The release in (3.1) satisfies $\rho_{y,c}$-zCDP (Bun & Steinke, 2016).

Feature and label privatization: We privatize both features X_c and labels y_c by releasing

$$\tilde{X}_c = \mathrm{Clip}_{K_{X,c}}(X_c) + \Xi_c, \qquad \tilde{y}_c = \mathrm{Clip}_{K_{y,c}}(y_c) + \xi_c, \tag{3.2}$$

where $\mathrm{Clip}_{K_{X,c}}(v) = v \min\{1, \sqrt{d} K_{X,c}/(2\|v\|_2)\}$ is the row-wise clipping function with a constant $K_{X,c}$ that could depend on n_c and d, entries of Ξ_c are i.i.d. $\sim \mathcal{N}(0, \sigma_{X,\mathrm{DP},c}^2)$, $\sigma_{X,\mathrm{DP},c}^2 = \min(\Delta_{X,c}^2, dK_{X,c}^2)/2\rho_{X,c}$, ξ_c is in (3.1), and $\Delta_{X,c} = \max_{X_c' \sim X_c''} \|X_c' - X_c''\|_2$ is the sensitivity of the features of client c. The release in (3.2) satisfies $(\rho_{X,c} + \rho_{y,c})$-zCDP (Bun & Steinke, 2016). All noises $(\{\varepsilon_c\}_{c \in \mathcal{C}}, \{\xi_c\}_{c \in \mathcal{A}}, \{\Xi_c\}_{c \in \mathcal{A}})$ are independent among themselves and of $(\theta, X_1, \ldots, X_C)$.

Assumptions. We consider the asymptotic regime where $d, n_1, \dots, n_C \to \infty$ and set $n := \sum_{c=1}^{C} n_c$, $\pi_c := \frac{n_c}{n}$, and $\gamma_c := \frac{d}{n_c}$. We assume an underparameterized regime **(A1)**, so that $\pi_c \to \bar{\pi}_c \in (0,1)$ and $\gamma_c \to \bar{\gamma}_c \in (0,1)$ for $c \in \{1, \dots, C\}$. We further assume almost surely $\sup_{c,n,d} \|\beta_c\|_2 < \infty$ and all entries of Z_c to be independent (between themselves and w.r.t. $\{\beta_1, \dots, \beta_C\}$), identically distributed within each client, zero-mean, unit-variance and having uniformly bounded 8-th moment **(A2)**. The population covariances are uniformly well-conditioned **(A3)**, i.e., $\omega_\Sigma I_d \preceq \Sigma_c \preceq \alpha_\Sigma I_d$ uniformly in c and d, for constants $0 < \omega_\Sigma < \alpha_\Sigma < \infty$.

We assume that noise levels, σ_c^2, $\sigma_{y,\mathrm{DP},c}^2$ and $\sigma_{X,\mathrm{DP},c}^2$, are constants independent of d, n **(A4)** and that clipping does not occur almost surely for large enough n **(A5)**. This last assumption is technical and we defer the analysis of clipping to future work, see Section 6. We note that this covers several practical settings: if the data are normalized so that the entries of X_c, y_c are bounded, then we can take $K_{y,c}, K_{X,c}$ of constant order and (3.1)-(3.2) hold with $\rho_{y,c}$ of constant order and $\rho_{X,c}$ of order d; if the data has sub-Gaussian tails, then $K_{y,c}, K_{X,c}$ are of order $\sqrt{\log d}$, which adds a logarithmic factor to $\rho_{y,c}, \rho_{X,c}$ as well (see Appendix A.1 for details). Whereas taking $\rho_{y,c}$ of constant order is standard in privacy, letting $\rho_{X,c}$ grow with d is due to the fact that features are d-dimensional: $\rho_{X,c}$ of order d corresponds to a noise of constant order per each coordinate. Moreover, we adopt a strong privacy model, for which large budgets are known to be needed (Bhowmick et al., 2018; Cyffers, 2025), and recent evidence indicates that the budget can grow with the dimension while protecting from reconstruction attacks (Liu et al., 2025; Hayes et al., 2023; Guo et al., 2022).

Weighted ridge estimator. For label privatization, let $X_c^{\mathrm{eff}} = X_c$ for every $c \in \mathcal{A}$. For feature and label privatization, X_c^{eff} is the conditionally denoised design defined in Section 4.2. In both cases, set $X_1^{\mathrm{eff}} = X_1$ and $\tilde{y}_1 = y_1$. We consider the weighted joint ridge estimator

$$(\hat{\beta}, \hat{\delta}_2, \dots, \hat{\delta}_C) = \underset{\hat{b}, \hat{h}_2, \dots, \hat{h}_C}{\arg\min} \left\{ \frac{\|y_1 - X_1 \hat{b}\|_2^2}{n} + \sum_{c \in \mathcal{A}} w_c \frac{\|\tilde{y}_c - X_c^{\mathrm{eff}}(\hat{b} + \hat{h}_c)\|_2^2}{n} + \lambda_\beta \|\hat{b}\|_2^2 + \sum_{c \in \mathcal{A}} \lambda_{\delta_c} \|\hat{h}_c\|_2^2 \right\}, \tag{3.3}$$

where $\{w_c\}_{c=2}^{C} =: \mathbf{w}$ are the weights, $\lambda_\beta \geq 0$ the ridge corresponding to the target β, and $\{\lambda_{\delta_c}\}_{c=2}^{C} =: \boldsymbol{\lambda}_\delta \succeq 0$ the ridges corresponding to model shift.

For $c \in \{1, \dots, C\}$, we denote the empirical covariance and cross-correlation terms as

$$\hat{\Sigma}_c := (X_c^{\mathrm{eff}})^\top X_c^{\mathrm{eff}} / n, \qquad \hat{v}_c := (X_c^{\mathrm{eff}})^\top \tilde{y}_c / n. \tag{3.4}$$

Next, we define the deterministic block-selection matrices $L_c \in \mathbb{R}^{d \times Cd}$ as $L_1 \theta = \beta$ and $L_c \theta = \beta + \delta_c$, for $c \in \mathcal{A}$. Equivalently, $L_1 = [I_d \quad 0 \quad \dots \quad 0]$, while L_c has an I_d in the β block and an I_d in the δ_c block. Hence, the model for client c is

$$y_c = X_c L_c \theta + \varepsilon_c.$$

Since it will be useful to view the problem as a regression in Cd dimensions, we introduce the joint estimator $\hat{\theta} := \begin{bmatrix} \hat{\beta}^\top & \hat{\delta}_2^\top & \dots & \hat{\delta}_C^\top \end{bmatrix}^\top$, as well as empirical matrices and vectors

$$\hat{S}_c := L_c^\top \hat{\Sigma}_c L_c \in \mathbb{R}^{Cd \times Cd}, \qquad \hat{U}_c := L_c^\top \hat{v}_c \in \mathbb{R}^{Cd}.$$

Let P_β be the projection onto the first block and P_{δ_c} the projection onto the δ_c block. Then, we can express the joint ridge term via $\Lambda := \lambda_\beta P_\beta + \sum_{c \in \mathcal{A}} \lambda_{\delta_c} P_{\delta_c} = \mathrm{diag}\,(\lambda_\beta I_d, \lambda_{\delta_2} I_d, \dots, \lambda_{\delta_C} I_d)$. With $\hat{S} := \hat{S}_1 + \sum_{c \in \mathcal{A}} w_c \hat{S}_c$, $\hat{U} := \hat{U}_1 + \sum_{c \in \mathcal{A}} w_c \hat{U}_c$, $\hat{Q} := (\hat{S} + \Lambda)^{-1}$, we obtain

$$\hat{\theta} = \hat{Q} \hat{U}. \tag{3.5}$$

The formulation in (3.3) contains the main estimators considered throughout. For example, setting $w_c = 0$ corresponds to a regression setting without the auxiliary client c. Similarly, setting $\lambda_{\delta_c} = 0$ makes the contribution of client c independent of b after minimizing over h_c, so that client c does not affect $\hat{\beta}$. On the other hand, sending $\lambda_{\delta_c} \to \infty$ forces $\hat{\delta}_c \to 0$ and pools client c with the target model. In particular, if every $\lambda_{\delta_c} \to \infty$, then the estimator in (3.3) reduces to the weighted pooled ridge estimator considered in Song et al. (2024); Yang et al. (2025); Rezaei et al. (2026), which ignores model shift and admits the closed-form solution

$$\hat{\beta} \longrightarrow \left(\hat{\Sigma}_1 + \sum_{c \in \mathcal{A}} w_c \hat{\Sigma}_c + \lambda_\beta I_d \right)^{-1} \left(v_1 + \sum_{c \in \mathcal{A}} w_c \hat{v}_c \right).$$

Finally, ridgeless estimators are obtained by taking the corresponding ridge parameters to zero.

Prediction risk. We evaluate all estimators on the distribution of the target client. If x_1 is an independent target covariate, then the conditional out-of-sample excess risk is

$$R(\hat{\beta};\beta) \coloneqq \mathbb{E}\left[(x_1^\top\hat{\beta} - x_1^\top\beta)^2 \,\middle|\, X^{\text{eff}}, \theta\right] = \mathbb{E}\left[\|\hat{\theta}-\theta\|_{S_1}^2 \,\middle|\, X^{\text{eff}}, \theta\right] = B(\hat{\theta};\theta) + V(\hat{\theta};\theta), \quad (3.6)$$

with $X^{\text{eff}} = (X_1^{\text{eff}} \dots X_C^{\text{eff}})$ and $B(\hat{\theta};\theta) \coloneqq \|\mathbb{E}[\hat{\theta} \mid X^{\text{eff}}, \theta] - \theta\|_{S_1}^2$, $V(\hat{\theta};\theta) \coloneqq \text{Tr}[\text{Cov}(\hat{\theta} \mid X^{\text{eff}}, \theta) S_1]$ denoting bias and variance of the risk. The expectation is taken over the noise in the data and the private noise. Define the augmented population covariance $S_c \coloneqq L_c^\top \Sigma_c L_c \in \mathbb{R}^{Cd\times Cd}$, and the total noise level of the privatized client $\sigma_{t,c}^2 \coloneqq \sigma_c^2 + \sigma_{y,\text{DP},c}^2$. Throughout the rest of the analysis, B, V, and R refer to the estimator with clipping omitted. Then, under label privatization, the following exact expressions (proved in Appendix A.2) hold

$$B(\hat{\theta};\theta) = \theta^\top \Lambda \hat{Q} S_1 \hat{Q} \Lambda \theta, \qquad V(\hat{\theta};\theta) = \frac{1}{n}\,\text{Tr}\left[S_1\hat{Q}\left(\sigma_1^2\hat{S}_1 + \sum_{c\in\mathcal{A}} w_c^2\sigma_{t,c}^2\hat{S}_c\right)\hat{Q}\right]. \quad (3.7)$$

4 Theoretical results

We derive asymptotic characterizations of the bias and variance in (3.7), under both label privatization (Section 4.1) and joint feature and label privatization (Section 4.2), and we characterize the optimal tuning parameters of the estimator in (3.3). We start with an additional assumption **(A6)**:

$$\beta = \frac{r_\beta}{\sqrt{d}}\,u_\beta, \qquad \delta_c = \frac{r_{\delta_c}}{\sqrt{d}}\,u_{\delta_c}, \quad c \in \mathcal{A}, \quad (4.1)$$

where $u_\beta,\ u_{\delta_2}, \dots, u_{\delta_C}$ have independent, centered, unit-variance coordinates with a uniformly bounded 8-th moment and $r_\beta, r_{\delta_c} > 0$ are constants independent of n, d. This represents an average-case setting in which no feature direction is singled out a priori, and it yields a characterization depending on the magnitudes of target signal and model shifts, rather than their alignment. Let us further define $\tau_\beta \coloneqq nr_\beta^2/d,\ \tau_{\delta_c} \coloneqq nr_{\delta_c}^2/d$. We will also assume that there exist constants w_-, w_+ such that $0 < w_- < w_c < w_+ < +\infty$ for all $c \in \mathcal{A}$.

4.1 Label privatization

General deterministic equivalent. We start by privatizing only labels via (3.1). We introduce the total noise terms $s_1^2 \coloneqq \sigma_1^2$ and $s_c^2 \coloneqq \sigma_c^2 + \sigma_{y,\text{DP},c}^2$, $c \in \mathcal{A}$. Without loss of generality, we consider $w_1 = 1$ and define $\boldsymbol{\pi} = (\pi_c)_{c=1}^C$ and $\boldsymbol{\Sigma} \coloneqq \{\Sigma_c\}_{c=1}^C$. Let $\bar{Q} = \bar{Q}(\mathbf{w}, \lambda_\beta, \boldsymbol{\lambda}_\delta, \boldsymbol{\pi}, \boldsymbol{\Sigma})$ be the unique positive-definite solution of

$$\bar{Q} = \left[\Lambda + \sum_{c=1}^C \frac{\pi_c w_c S_c}{1+e_c}\right]^{-1}, \quad (4.2)$$

where $e_c = e_c(\mathbf{w}, \lambda_\beta, \boldsymbol{\lambda}_\delta, \boldsymbol{\pi}, \boldsymbol{\Sigma})$ are the C fixed point coefficients satisfying

$$e_c = \frac{w_c}{n}\,\text{Tr}(S_c\bar{Q}), \qquad c \in \{1,\dots,C\}. \quad (4.3)$$

Moreover, we define the matrix $J \in \mathbb{R}^{C\times C}$, the vectors $b, g \in \mathbb{R}^C$ and the matrix $\mathcal{K} \in \mathbb{R}^{Cd\times Cd}$ as

$$J_{ck} = \frac{\pi_k w_c w_k}{n(1+e_k)^2}\,\text{Tr}(S_c\bar{Q}S_k\bar{Q}), \quad b_c = \frac{w_c}{n}\,\text{Tr}(S_c\bar{Q}S_1\bar{Q}),\ c,k \in \mathcal{C}, \quad g = (I_C - J)^{-1}b. \quad (4.4)$$

$$\mathcal{K} = \bar{Q}\left[S_1 + \sum_{c=1}^C \frac{\pi_c w_c g_c}{(1+e_c)^2} S_c\right]\bar{Q}. \quad (4.5)$$

At this point, we can state our result which establishes a deterministic equivalent of the bias and variance terms, depending on signal parameters $\tau_\beta, \tau_{\delta_c}$, alignment of β_c's weighted by the covariance, sample complexities $\boldsymbol{\pi}$, covariance structures $\boldsymbol{\Sigma}$, intrinsic noise levels $\{\sigma_c^2\}_{c=1}^C$, weights $\{w_c\}_{c=2}^C$, ridge parameters $\lambda_\beta,\ \{\lambda_{\delta_c}\}_{c=2}^C$, and privacy noise intensities $\{\sigma_{y,\text{DP},c}^2\}_{c=2}^C$.

Theorem 4.1 (Deterministic equivalent for label privatization)**.** *Under the Assumptions (A1) - (A5), it holds, almost surely, that*

$$\lim_{n\to\infty}\left|B(\hat{\theta},\theta) - \mathcal{B}\right| = 0, \qquad \lim_{n\to\infty}\left|V(\hat{\theta},\theta) - \mathcal{V}\right| = 0, \quad (4.6)$$

where

$$\mathcal{B} = \theta^\top \Lambda \mathcal{K} \Lambda \theta, \qquad \mathcal{V} = \sum_{c=1}^{C} \frac{\pi_c w_c g_c s_c^2}{(1+e_c)^2}. \tag{4.7}$$

If we additionally assume that (A6) *holds, then* (4.6) *holds with* $\mathcal{V}$ *as in* (4.7) *and*

$$\mathcal{B} = \frac{\tau_\beta \lambda_\beta^2}{n} \operatorname{Tr}(P_\beta \mathcal{K}) + \sum_{c \in \mathcal{A}} \frac{\tau_{\delta_c} \lambda_{\delta_c}^2}{n} \operatorname{Tr}(P_{\delta_c} \mathcal{K}). \tag{4.8}$$

As an immediate corollary, $R(\hat{\beta}; \beta)$ converges almost surely to $\mathcal{B} + \mathcal{V}$, and it does so almost surely even without averaging over observation and privacy noises (Appendix B.3). Thus, although the estimator $\hat{\beta}$ depends on random, high-dimensional samples, its asymptotic risk is characterized by the deterministic quantities $\mathcal{K}$ and g. In Figure 1, we show that the deterministic equivalent closely approximates the empirical risk at moderate dimension. The plots further highlight how the risk can be substantially influenced by the sources of heterogeneity, such as **(a)** covariance shift, **(b)** model shift, and **(c)** privacy budget. We next sketch the argument deferring details to Appendix B.1.

Idea of the proof. The argument relies on reducing both bias and variance to the scalar resolvent functional $F = n^{-1} \operatorname{Tr}(S_1 \widehat{Q})$ and its derivatives. The deterministic equivalent for resolvents of random matrices with heterogeneous covariance structure, such as $\hat{Q}$, is well studied and an application of (Wagner et al., 2012, Theorem 1) gives that $|F - e_1| \to 0$, for e_1 defined by (4.2)-(4.3). However, to be able to apply such result, one first has to introduce an additional small positive regularization, which can be later removed using the uniform lower spectral bound on sample covariance matrices in the underparameterized regime. Then, uniform bounds on the second derivatives of F and e_1 allow the convergence to be transferred to their first derivatives. We finally lift assumption (A6) and using (Louart, 2026, Theorem 1.6), we get the directional concentration needed to show (4.7). □

Remark 4.2. *Theorem 4.1 is related to the high-dimensional multi-task linear regression analysis of Ilbert et al. (2024), who also consider a shared signal with task-specific deviations under heterogeneous feature covariances. The settings, though, differ in several respects. In particular, the assumptions on the features are different: whereas Ilbert et al. (2024) assume Lipschitz concentration of the features, we work under the bounded moment condition (A2). Most importantly, rather than studying multi-task prediction, we have a single target client and account for auxiliary weights and privacy mechanisms, which are central to the dataset selection problem we consider.*

Bayes-optimal tuning. If we impose further restrictions on the distribution of β and δ, then we can show optimality of our estimator and characterize the parameters $(\mathbf{w}, \lambda_\beta, \boldsymbol{\lambda}_\delta)$ which achieve it.

Proposition 4.3 (Bayes-optimal privacy weights and shift regularization)**.** *Consider the setting of Theorem 4.1 with the unclipped release* $y_c + \xi_c$ *for* $c \in \mathcal{A}$ *and Gaussian noise* $\{\varepsilon_j\}_{j=1}^C$*. Assume* $\sigma_1^2 > 0$*,* $\sigma_{t,c}^2 > 0$ *for* $c \in \mathcal{A}$ *and*

$$\beta \sim \mathcal{N}\left(0, \frac{r_\beta^2}{d} I_d\right), \qquad \delta_c \sim \mathcal{N}\left(0, \frac{r_{\delta_c}^2}{d} I_d\right),\ c \in \mathcal{A}, \tag{4.9}$$

independently. Then, conditional on the designs, the posterior mean of the stacked parameter, i.e. $\mathbb{E}\left[\theta \mid X_1, \ldots, X_C, y_1, \tilde{y}_2, \ldots, \tilde{y}_C\right]$*, is exactly the estimator in* (3.3) *with*

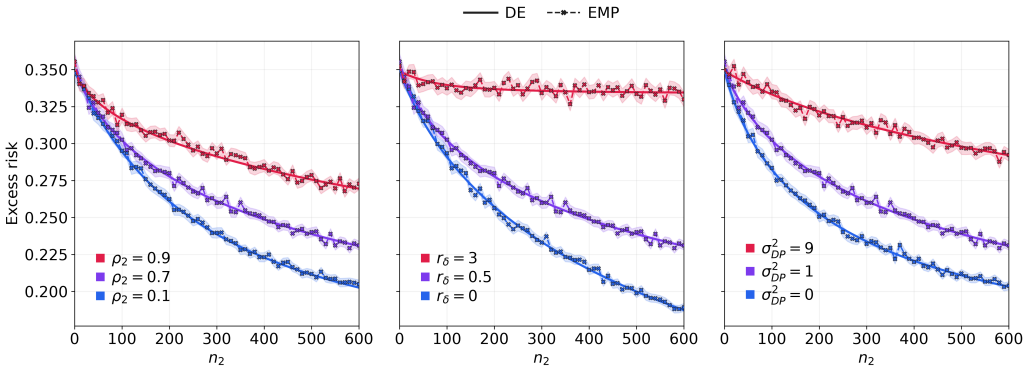
$$w_c^\circ = \frac{\sigma_1^2}{\sigma_{t,c}^2}, \qquad \lambda_\beta^\circ = \frac{\sigma_1^2}{\tau_\beta} = \frac{\sigma_1^2 d}{n r_\beta^2}, \qquad \lambda_{\delta_c}^\circ = \frac{\sigma_1^2}{\tau_{\delta_c}} = \frac{\sigma_1^2 d}{n r_{\delta_c}^2}. \tag{4.10}$$

This estimator minimizes $\mathbb{E}_\theta[R(\hat{\beta}; \beta)]$ *among all measurable estimators.*

In words, under Gaussian priors on target signal and model shifts, the weighted joint ridge estimator with parameters in (4.10) is Bayes-optimal. The argument is deferred to Appendix B.4.

4.2 Feature and label privatization

We next consider the case in which both features and labels are made private by the Gaussian mechanism in (3.2). To analyze this case, we assume Gaussian noise $\{\varepsilon_j\}_{j=1}^C$ and strengthen



(a) Varying covariance shift **(b)** Varying model shift **(c)** Varying privacy noise

Figure 1: Dependence of target excess risk on covariance shift, model shift and privacy noise in a synthetic linear regression experiment with two clients. We take $x_1 \sim \mathcal{N}(0, \Sigma_1)$ and $x_2 \sim \mathcal{N}(0, \Sigma_2)$, with Σ_1 and Σ_2 of the form $[K(\rho)]_{ij} = \rho^{|i-j|}$. We set $\Sigma_1 = K(\rho_1)$ with $\rho_1 = 0.1$ and $\Sigma_2 = a_2 K(\rho_2)$, so that $\mathrm{Tr}(\Sigma_2 \Sigma_1^{-1}) = d$. We take $d = 80$, $n_1 = 200$ and plot $\|\hat{\beta} - \beta\|^2_{\Sigma_1}$ as a function of n_2. Solid lines show the deterministic equivalent prediction and dashed lines with crosses show the empirical mean over 400 i.i.d. repetitions, with shaded 90% confidence intervals. **(a)** Covariance shift: $\rho_2 \in \{0.1, 0.7, 0.9\}$, with model shift $r_\delta = 0.5$ and privacy noise variance $\sigma^2_{\mathrm{DP}} = 1$. **(b)** Model shift: $r_\delta \in \{0, 0.5, 3\}$, $\rho_2 = 0.7$ and $\sigma^2_{\mathrm{DP}} = 1$. **(c)** Privacy noise: $\sigma^2_{\mathrm{DP}} \in \{0, 1, 9\}$, with $\rho_2 = 0.7$ and $r_\delta = 0.5$. The middle curve is the same in all three panels.

Assumption (A2) by assuming **(A7)** that the rows of X_c are i.i.d. Gaussian with covariance Σ_c, i.e., $x_{ci} \sim \mathcal{N}(0, \Sigma_c)$ for $c \in \mathcal{A}$ and $i \in \{1, \dots, n_c\}$. The reduction to an equivalent regression problem can then be performed separately for each auxiliary client. In particular, for $c \in \mathcal{A}$ and $i \in \{1, \dots, n_c\}$, we write

$$\tilde{x}_{ci} = x_{ci} + \xi_{X,ci}, \qquad \xi_{X,ci} \sim \mathcal{N}(0, \sigma^2_{X,\mathrm{DP},c} I_d).$$

Gaussian conditioning gives $x_{ci} \mid \tilde{x}_{ci} \sim \mathcal{N}(K_c \tilde{x}_{ci}, T_c)$, where $K_c := \Sigma_c(\Sigma_c + \sigma^2_{X,\mathrm{DP},c} I_d)^{-1}$ and $T_c := \Sigma_c - \Sigma_c(\Sigma_c + \sigma^2_{X,\mathrm{DP},c} I_d)^{-1}\Sigma_c$. We then define the conditionally denoised design as

$$X_c^{\mathrm{eff}} = \tilde{X}_c K_c.$$

Writing $X_c = X_c^{\mathrm{eff}} + E_c$, the rows of E_c are independent $\mathcal{N}(0, T_c)$ vectors, which are also independent of X^{eff}. Since $y_c = X_c(\beta + \delta_c) + \varepsilon_c$, the privatized response satisfies

$$\tilde{y}_c = X_c^{\mathrm{eff}}(\beta + \delta_c) + \xi_c^{\mathrm{eff}}, \qquad \xi_c^{\mathrm{eff}} := E_c(\beta + \delta_c) + \varepsilon_c + \xi_c.$$

Moreover, conditional on (β, δ_c), it follows that $\xi_c^{\mathrm{eff}} \sim \mathcal{N}(0, \nu_c^2 I_{n_c})$ is independent of X_c^{eff}, where $\nu_c^2 = \sigma_c^2 + \sigma^2_{y,\mathrm{DP},c} + (\beta + \delta_c)^\top T_c(\beta + \delta_c)$. The denoised population covariance is

$$\Sigma_c^{\mathrm{eff}} := \mathrm{Cov}(x_c^{\mathrm{eff}}) = \Sigma_c(\Sigma_c + \sigma^2_{X,\mathrm{DP},c} I_d)^{-1}\Sigma_c.$$

Thus, feature privatization has two effects: $\Sigma_c \longmapsto \Sigma_c^{\mathrm{eff}}$ and $\sigma_c^2 + \sigma^2_{y,\mathrm{DP},c} \longmapsto \nu_c^2$, which, respectively, add covariance shift and increase effective noise.

We now proceed to use the estimator in (3.3) with X_c^{eff} as defined above. Let g^{eff}, $\mathcal{K}^{\mathrm{eff}}$, and e_c^{eff} denote the quantities g, $\mathcal{K}$, and e_c defined in (4.2)-(4.5) with $\mathbf{\Sigma}$ replaced by $\mathbf{\Sigma}^{\mathrm{eff}} := \{\Sigma_c^{\mathrm{eff}}\}_{c=1}^C$. Moreover, note that under the Assumption (A6), it holds

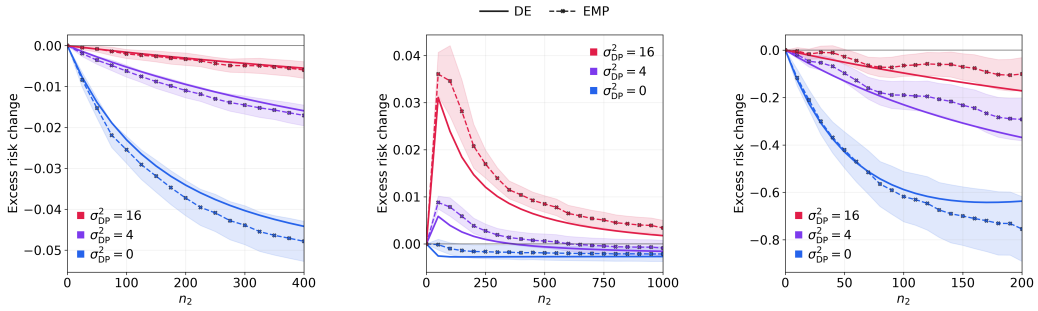
$$\lim_{n\to\infty} \max_{c \in \mathcal{A}} \left|\nu_c^2 - \bar{\nu}_c^2\right| = 0, \qquad \bar{\nu}_c^2 := \sigma_c^2 + \sigma^2_{y,\mathrm{DP},c} + \frac{r_\beta^2 + r_{\delta_c}^2}{d}\,\mathrm{Tr}(T_c), \tag{4.11}$$

almost surely. Thus, an application of Theorem 4.1 gives the corollary below.

Corollary 4.4 (Deterministic equivalent for label and feature privatization)**.** *Under the Assumptions (A1) - (A7), it holds, almost surely, that*

$$\lim_{n\to\infty} \left|R(\hat{\beta}; \beta) - \mathcal{R}\right| = 0,$$

where $\mathcal{R} = \theta^\top \Lambda \mathcal{K}^{\mathrm{eff}} \Lambda \theta + \sum_{c=1}^C (\pi_c w_c g_c^{\mathrm{eff}} \bar{\nu}_c^2)/(1 + e_c^{\mathrm{eff}})^2$.



(a) LSAC **(b)** Lianjia **(c)** Student Performance

Figure 2: Change in target risk relative to the target-only baseline as auxiliary data are added. **(a) LSAC:** Tier 3 is the target client and Tier 4 the auxiliary client. **(b) Lianjia:** District 8 is the target client and District 7 the auxiliary client. **(c) Student Performance:** Mousinho da Silveira is the target school and Gabriel Pereira the auxiliary school.

5 Experimental results

In this section, we evaluate our theoretical predictions on three real-world regression datasets. We first assess how accurately the deterministic equivalents (4.7)-(4.8) predict changes in target risk as auxiliary sample size and privacy noise vary (Figure 2). We then use these predictions to select beneficial auxiliary clients and compare against several heuristic baselines (Figure 3).

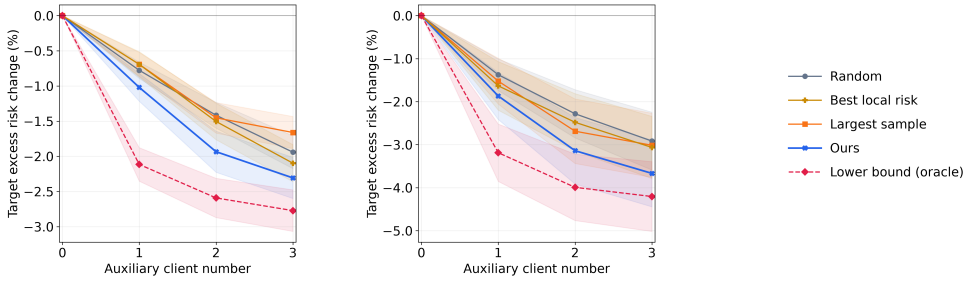
Datasets. We consider three publicly available regression datasets: *(i)* in *Student Performance*[1] (Cortez & Silva, 2008), we predict the final Portuguese-course grade G3 and treat the two schools as clients; *(ii)* in *LSAC*[2](Wightman, 1998), we predict standardized first-year GPA and treat law-school tiers as clients; *(iii)* in *Lianjia*[3], we predict log housing transaction prices in Beijing and treat city districts as clients. Dataset preprocessing and additional statistics are given in Appendix C.1.

Experimental setup. In each experiment, one client is designated as the target and the remaining clients as auxiliary with privacy requirements. We run 100 evaluation repetitions. In each repetition, the target training size is fixed at $n_1 = 100$, approximately 20% of the target observations are used for validation, and the remainder are used for testing. Auxiliary observations are sampled without replacement, with n_2 ranging up to 200, 400, and 1000 for Student Performance, LSAC, and Lianjia, respectively. Following (3.1), independent Gaussian noise is added to auxiliary labels while covariates remain unchanged. Estimation of the distributional quantities entering the deterministic equivalent and the hyperparameter tuning procedures are described in Appendix C.1.

Prediction of target risk. Figure 2 shows that the deterministic equivalent prediction accurately tracks empirical performance, as the auxiliary sample size n_2 and the privacy noise level σ^2_{DP} vary. We consider $\sigma^2_{\mathrm{DP}} \in \{0, 4, 16\}$ and plot the change in target risk relative to the target-only baseline (i.e., the difference between using n_2 client samples and not using client samples at all), as a function of n_2. Negative values indicate an improvement over the baseline using only target data. Dashed lines with crosses are empirical averages, with shaded regions denoting 90% confidence intervals for the empirical mean. Solid lines in plots show the simplified deterministic equivalent predictions (4.8), illustrating that a more tractable expression gives accurate predictions in this setting. Similar results are obtained in a semi-synthetic setting in which a linear model is planted on the real-world design matrix, and we use the general deterministic equivalent predictions (4.7), see Appendix C.2.

Dataset selection. Figure 3 demonstrates that the deterministic equivalent effectively guides auxiliary client selection in the multi-client LSAC and Lianjia datasets. Student Performance contains only two clients and is therefore excluded from this evaluation. Starting from the target-only estimator, each method sequentially selects auxiliary clients. All auxiliary clients have a fixed number of samples and receive an independent Gaussian privacy noise with variance $\sigma^2_{\mathrm{DP}} = 4$. After selecting k clients, we report the improvement over the baseline using only target data, expressed in percentage. As before, negative values indicate improvement over the target-only baseline.

[1] https://archive.ics.uci.edu/dataset/320/student+performance
[2] https://www.kaggle.com/datasets/danofer/law-school-admissions-bar-passage
[3] https://www.kaggle.com/datasets/ruiqurm/lianjia



(a) LSAC **(b)** Lianjia

Figure 3: Percentage change in target test risk relative to the target-only baseline under sequential auxiliary client selection. **(a) LSAC:** Tier 3 is the target client and Tiers 1, 2, 4, 5, and 6 are candidate auxiliary clients. **(b) Lianjia:** District 8 is the target client and Districts 1, 2, 4, 6, 7, 10, 11 and 13 are candidate auxiliary clients.

The method *Ours* greedily adds the auxiliary client that minimizes the deterministic equivalent prediction for the target risk from Theorem 4.1 using hyperparameters from Proposition 4.3. Our approach employs only population-level quantities that do not require access to client data and could be estimated from a publicly available dataset. This is quite common, e.g., when dealing with data from US Census (Bureau, 2026; Abowd et al., 2022) or UK Biobank (Sudlow et al., 2015). In our context, we assume the availability of a non-private dataset used to estimate the population quantities $\{\Sigma_c, \sigma_c\}_{c=1}^C$, together with either the relevant inner products between $\{\beta_c\}_{c=1}^C$ or τ_β and $\{\tau_{\delta_c}\}_{c=2}^C$, which are plugged into (4.7) or (4.8), respectively. More precisely, the non-private dataset consists of the remaining data not used for the experiment of Figure 3. Plots **(a)** and **(b)** use the deterministic equivalent (4.7), although the simplified version (4.8) also improves LSAC client selection.

We compare against three heuristic baselines: *Random*, which selects clients uniformly at random; *Largest sample*, which orders clients by decreasing size of their training set; and *Best local risk*, which ranks clients by performance on their own local distributions. Finally, we report *Lower bound (oracle)*, which at each step greedily adds the client that yields the lowest empirical target risk on the test set. Because it uses test data, this baseline cannot be achieved in practice and serves only as a reference. We report empirical averages (squares) and 90% confidence intervals (shaded regions). Figure 3 shows that our approach outperforms the three baselines, and similar results for different privacy noise levels are reported in Appendix C.3.

6 Concluding Remarks

In this work, we propose a principled way to decide which datasets will provide value for ridge regression, under differential privacy (DP) constraints. Our setting accommodates covariate and model shifts and provides useful predictions in experiments on real datasets, even though the underlying models are likely to be more complex than a simple linear model. Our results show that deterministic equivalents of the risk are accurate enough to guide dataset selection and identify the Bayes-optimal way to combine sources under our model assumptions.

An interesting future direction involves the analysis of nonlinear models. A first step would be to consider clipping of features and labels, which makes the models in (3.1)-(3.2) nonlinear. In fact, the importance of clipping is well reported in practice (De et al., 2022) and recent theoretical work has started to analyze the impact of gradient clipping in high-dimensional regression (Bombari et al., 2026; Dwork et al., 2026). A natural next step would be then to consider multi-index models, for instance when early feature learning is followed by ridge regression (Ba et al., 2022; Cui et al., 2024). In fact, combining linear probing with feature adaptation is an established approach to transfer learning (Collins et al., 2024; Tahir et al., 2025; Kumar et al., 2022). Finally, more sophisticated privacy primitives, e.g., the private covariance estimators developed by Amin et al. (2019); Kamath et al. (2019), could be used for both data selection and privatization in order to improve the privacy-utility tradeoff.

AI USE STATEMENT

In this work, we used generative AI tools to polish the writing, implement straightforward parts of the code (e.g., Matplotlib code for generating figures), fill in few details of proofs for which we had already identified the key methods and arguments, and improve the completeness of the related work section. We did not use generative AI tools to develop the paper's core ideas, define the problem setting, or prove the main theoretical results. We take full responsibility for the final content of this work, including any text, claims, or artifacts produced with the assistance of generative AI.

REPRODUCIBILITY STATEMENT

All experiments were conducted on widely used, publicly available datasets, which are referenced in Section 5. We describe the experimental setup in sufficient detail to enable reproduction of the paper's figures. The code will be made publicly available on GitHub upon acceptance of the paper. All proofs of theoretical results are provided in Appendices A-B.

REFERENCES

John Abowd, Robert Ashmead, Ryan Cumings-Menon, Simson Garfinkel, Micah Heineck, Christine Heiss, Robert Johns, Daniel Kifer, Philip Leclerc, Ashwin Machanavajjhala, Brett Moran, William Sexton, Matthew Spence, and Pavel Zhuravlev. The 2020 census disclosure avoidance system topdown algorithm. *Harvard Data Science Review*, (Special Issue 2), 2022. doi: 10.1162/99608f92.529e3cb9.

Noga Alon, Raef Bassily, and Shay Moran. Limits of private learning with access to public data. In *Advances in neural information processing systems*, volume 32, 2019.

Kareem Amin, Travis Dick, Alex Kulesza, Andres Munoz, and Sergei Vassilvitskii. Differentially private covariance estimation. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

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.

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

Raef Bassily, Adam Smith, and Abhradeep Thakurta. Private empirical risk minimization: Efficient algorithms and tight error bounds. In *55th Annual Symposium on Foundations of Computer Science*. IEEE, 2014.

Raef Bassily, Vitaly Feldman, Kunal Talwar, and Abhradeep Guha Thakurta. Private stochastic convex optimization with optimal rates. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Abhishek Bhowmick, John C. Duchi, Julien Freudiger, Gaurav Kapoor, and Ryan M. Rogers. Protection against reconstruction and its applications in private federated learning. *ArXiv*, abs/1812.00984, 2018.

Alex Bie, Gautam Kamath, and Vikrant Singhal. Private estimation with public data. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

Adam Block, Mark Bun, Rathin Desai, Abhishek Shetty, and Zhiwei S Wu. Oracle-efficient differentially private learning with public data. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Simone Bombari, Jialei Luo, Inbar Seroussi, and Marco Mondelli. High-dimensional private linear regression with optimal rates, 2026.

Mark Bun and Thomas Steinke. Concentrated Differential Privacy: Simplifications, Extensions, and Lower Bounds. In *Theory of Cryptography*, Lecture Notes in Computer Science, pp. 635–658. Springer, 2016.

United States Census Bureau. Census datasets, 2026.

T Tony Cai, Yichen Wang, and Linjun Zhang. The cost of privacy: Optimal rates of convergence for parameter estimation with differential privacy. *The Annals of Statistics*, 49(5), 2021.

Kamalika Chaudhuri, Claire Monteleoni, and Anand D. Sarwate. Differentially private empirical risk minimization. *Journal of Machine Learning Research*, 12(29):1069–1109, 2011.

Liam Collins, Hamed Hassani, Mahdi Soltanolkotabi, Aryan Mokhtari, and Sanjay Shakkottai. Provable multi-task representation learning by two-layer ReLU neural networks. In *International Conference on Machine Learning*, volume 235. PMLR, 2024.

Paulo Cortez and Alice Maria Gonçalves Silva. Using data mining to predict secondary school student performance. In *Proceedings of the 5th Future Business Technology Conference (FUBUTEC 2008)*, pp. 5–12, Porto, Portugal, 2008. EUROSIS.

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.

Rachel Cummings, Hadi Elzayn, Emmanouil Pountourakis, Vasilis Gkatzelis, and Juba Ziani. Optimal data acquisition with privacy-aware agents. In *IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*, 2023.

Edwige Cyffers. Setting epsilon is not the issue in differential privacy. In *Advances in Neural Information Processing Systems*, 2025.

Soham De, Leonard Berrada, Jamie Hayes, Samuel L. Smith, and Borja Balle. Unlocking high-accuracy differentially private image classification through scale. *ArXiv*, abs/2204.13650, 2022.

Oussama Dhifallah and Yue M Lu. Phase transitions in transfer learning for high-dimensional perceptrons. *Entropy*, 23(4), 2021.

John C Duchi, Michael I Jordan, and Martin J Wainwright. Minimax optimal procedures for locally private estimation. *Journal of the American Statistical Association*, 113(521), 2018.

Cynthia Dwork and Aaron Roth. The algorithmic foundations of differential privacy. *Foundations and trends® in theoretical computer science*, 9(3–4), 2014.

Cynthia Dwork, Pranay Tankala, and Linjun Zhang. Differentially private learning beyond the classical dimensionality regime. In Benny Applebaum and Huijia (Rachel) Lin (eds.), *Theory of Cryptography*, Cham, 2026. Springer Nature Switzerland. ISBN 978-3-032-12290-2.

Alireza Fallah, Ali Makhdoumi, Azarakhsh Malekian, and Asuman E. Ozdaglar. Optimal and differentially private data acquisition: Central and local mechanisms. In *23rd ACM Conference on Economics and Computation*, 2022.

Federica Gerace, Luca Saglietti, Stefano Sarao Mannelli, Andrew Saxe, and Lenka Zdeborová. Probing transfer learning with a model of synthetic correlated datasets. *Machine Learning: Science and Technology*, 3(1), 2022.

Badih Ghazi, Noah Golowich, Ravi Kumar, Pasin Manurangsi, and Chiyuan Zhang. Deep learning with label differential privacy. In *Advances in Neural Information Processing Systems*, volume 34, 2021.

Badih Ghazi, Pritish Kamath, Ravi Kumar, Ethan Leeman, Pasin Manurangsi, Avinash V Varadarajan, and Chiyuan Zhang. Regression with label differential privacy. In *International Conference on Learning Representations*, 2023.

Xin Gu, Gautam Kamath, and Zhiwei Steven Wu. Choosing public datasets for private machine learning via gradient subspace distance. In *IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*, 2025.

Chuan Guo, Brian Karrer, Kamalika Chaudhuri, and Laurens van der Maaten. Bounding training data reconstruction in private (deep) learning. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*. PMLR, 17–23 Jul 2022.

Jamie Hayes, Saeed Mahloujifar, and Borja Balle. Bounding training data reconstruction in dp-sgd. In *Advances in Neural Information Processing Systems*, 2023.

Romain Ilbert, Malik Tiomoko, Cosme Louart, Ambroise Odonnat, Vasilii Feofanov, Themis Palpanas, and Ievgen Redko. Analysing multi-task regression via random matrix theory with application to time series forecasting. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Alessandro Ingrosso, Rosalba Pacelli, Pietro Rotondo, and Federica Gerace. Statistical mechanics of transfer learning in fully connected networks in the proportional limit. *Physical Review Letters*, 134(17), 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, 2024.

Gautam Kamath, Jerry Li, Vikrant Singhal, and Jonathan Ullman. Privately learning high-dimensional distributions. In *Conference on Learning Theory*. PMLR, 2019.

Germain Kolossov, Andrea Montanari, and Pulkit Tandon. Towards a statistical theory of data selection under weak supervision. In *International Conference on Learning Representations*, volume 2024, 2024.

Ananya Kumar, Aditi Raghunathan, Robbie Matthew Jones, Tengyu Ma, and Percy Liang. Fine-tuning can distort pretrained features and underperform out-of-distribution. In *International Conference on Learning Representations*, 2022.

Chao Li, Daniel Yang Li, Gerome Miklau, and Dan Suciu. A theory of pricing private data. In *Proceedings of the 16th International Conference on Database Theory*, 2013.

Mengchu Li, Ye Tian, Yang Feng, and Yi Yu. Federated transfer learning with differential privacy. *arXiv preprint arXiv:2403.11343*, 2024.

Sai Li, T Tony Cai, and Hongzhe Li. Transfer learning for high-dimensional linear regression: Prediction, estimation and minimax optimality. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 84(1), 2022.

Junxu Liu, Jian Lou, Li Xiong, Jinfei Liu, and Xiaofeng Meng. Projected federated averaging with heterogeneous differential privacy. *Proceedings of the VLDB Endowment*, 15(4), 2021.

Sheng Liu, Zihan Wang, Yuxiao Chen, and Qi Lei. Data reconstruction attacks and defenses: A systematic evaluation. In *International Conference on Artificial Intelligence and Statistics*, volume 258. PMLR, 2025.

Cosme Louart. Resolvent convergence for sample covariance matrices with general covariance profiles and quadratic-form control. *arXiv preprint arXiv:2109.02644v7*, 2026.

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. doi: 10.1214/23-AOS2268.

Neil Rohit Mallinar, Austin Zane, Spencer Frei, and Bin Yu. Minimum-norm interpolation under covariate shift. In *International Conference on Machine Learning*, 2024.

Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *International Conference on Artificial Intelligence and Statistics*. PMLR, 2017.

Jean Ogier du Terrail, Samy-Safwan Ayed, Edwige Cyffers, Felix Grimberg, Chaoyang He, Regis Loeb, Paul Mangold, Tanguy Marchand, Othmane Marfoq, Erum Mushtaq, Boris Muzellec, Constantin Philippenko, Santiago Silva, Maria Teleńczuk, Shadi Albarqouni, Salman Avestimehr, Aurélien Bellet, Aymeric Dieuleveut, Martin Jaggi, Sai Praneeth Karimireddy, Marco Lorenzi, Giovanni Neglia, Marc Tommasi, and Mathieu Andreux. Flamby: datasets and benchmarks for cross-silo federated learning in realistic healthcare settings. In *Advances in Neural Information Processing Systems*, 2022.

Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10), 2009.

Pratik Patil, Jin-Hong Du, and Ryan J Tibshirani. Optimal ridge regularization for out-of-distribution prediction. In *International Conference on Machine Learning*, 2024.

Parham Rezaei, Filip Kovačević, Francesco Locatello, and Marco Mondelli. High-dimensional analysis of synthetic data selection. In *International Conference on Learning Representations*, 2026.

Stefano Sarao Mannelli, Federica Gerace, Negar Rostamzadeh, and Luca Saglietti. Bias-inducing geometries: An exactly solvable data model with fairness implications. *Physical Review E*, 112 (2), 2025.

Kanthi Sarpatwar, Karthikeyan Shanmugam, Venkata Sitaramagiridharganesh Ganapavarapu, Ashish Jagmohan, and Roman Vaculin. Differentially private distributed data summarization under covariate shift. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Felix Sattler, Klaus-Robert Müller, and Wojciech Samek. Clustered federated learning: Model-agnostic distributed multitask optimization under privacy constraints. *IEEE Transactions on Neural Networks and Learning Systems*, 32(8), 2020.

Yanke Song, Kenneth Gu, Sohom Bhattacharya, and Pragya Sur. Generalization error of min-norm interpolators in transfer learning. *arXiv preprint arXiv:2406.13944*, 2024.

Cathie Sudlow, John Gallacher, Naomi Allen, Valerie Beral, Paul Burton, John Danesh, Paul Downey, Paul Elliott, Jane Green, Martin Landray, Bette Liu, Paul Matthews, Giok Ong, Jill Pell, Alan Silman, Alan Young, Tim Sprosen, Tim Peakman, and Rory Collins. UK biobank: An open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLOS Medicine*, 12(3):e1001779, 2015.

Masashi Sugiyama, Matthias Krauledat, and Klaus-Robert Müller. Covariate shift adaptation by importance weighted cross validation. *Journal of Machine Learning Research*, 8(35):985–1005, 2007.

Javan Tahir, Surya Ganguli, and Grant M. Rotskoff. Features are fate: a theory of transfer learning in high-dimensional regression. In *International Conference on Machine Learning*, volume 267. PMLR, 2025.

Ye Tian and Yang Feng. Transfer learning under high-dimensional generalized linear models. *Journal of the American Statistical Association*, 118(544), 2023.

Nikita Tsoy, Anna Mihalkova, Teodora Todorova, and Nikola Konstantinov. Provable mutual benefits from federated learning in privacy-sensitive domains. In *International Conference on Artificial Intelligence and Statistics*, volume 238. PMLR, 2024.

Enayat Ullah, Michael Menart, Raef Bassily, Cristóbal Guzmán, and Raman Arora. Public-data assisted private stochastic optimization: Power and limitations. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge University Press, 2 edition, 2026.

Sebastian Wagner, Romain Couillet, Mérouane Debbah, and Dirk T. M. Slock. Large system analysis of linear precoding in correlated miso broadcast channels under limited feedback. *IEEE Transactions on Information Theory*, 58(7):4509–4537, 2012.

Zirui Wang, Zihang Dai, Barnabás Póczos, and Jaime Carbonell. Characterizing and avoiding negative transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019.

Linda F. Wightman. Lsac national longitudinal bar passage study. Technical report, Law School Admission Council, Newtown, PA, 1998.

Sen Wu, Hongyang R. Zhang, and Christopher Ré. Understanding and improving information transfer in multi-task learning. In *International Conference on Learning Representations*, 2020.

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), 2025.

Hongzhe Zhang and Hongzhe Li. Transfer learning with random coefficient ridge regression. *arXiv preprint arXiv:2306.15915*, 2023.

A TECHNICAL PRELIMINARIES

In this appendix we provide the privacy and clipping details from Section 3, and derive the bias and variance formulas in (3.7).

A.1 SUFFICIENT CONDITIONS FOR INACTIVE CLIPPING

First, let us recall that

$$\sigma^2_{y,\mathrm{DP},c} = \frac{\min(\Delta^2_{y,c}, K^2_{y,c})}{2\rho_{y,c}}, \quad \sigma^2_{X,\mathrm{DP},c} = \frac{\min(\Delta^2_{X,c}, dK^2_{X,c})}{2\rho_{X,c}}, \tag{A.1}$$

which under Assumption (A4) we can treat as fixed constants independent of n and d.

Let $\mathcal{E}_n$ denote the event that at least one auxiliary observation is clipped:

$$\mathcal{E}_n = \bigcup_{c\in\mathcal{A}} \left\{ \max_{i\le n_c} |y_{c,i}| > K_{y,c}/2 \text{ or } \max_{i\le n_c} \|X_{c,i}\|_2 > \sqrt{d}K_{X,c}/2 \right\}.$$

We give sufficient conditions under which $\sum_n \mathbb{P}(\mathcal{E}_n) < \infty$. The first Borel-Cantelli lemma then implies that, almost surely, for all sufficiently large n,

$$\mathrm{Clip}_{K_{y,c}}(y_c) = y_c, \qquad \mathrm{Clip}_{K_{X,c}}(X_c) = X_c, \qquad c \in \mathcal{A}. \tag{A.2}$$

Note that for label privatization alone, the feature-clipping condition can simply be omitted.

Uniformly bounded samples. Suppose there are deterministic constants $B_{y,c}, B_{X,c} > 0$, independent of n and d, such that

$$|y_{c,i}| \le B_{y,c}/2, \qquad |X_{c,ij}| \le B_{X,c}/2$$

almost surely for every i, j, n, d. Then $\|x_{c,i}\|_2 \le \sqrt{d}\, B_{X,c}/2$, so the choices

$$K_{y,c} = B_{y,c}, \qquad K_{X,c} = B_{X,c}$$

make clipping inactive for every sample size. For fixed positive privacy noise variances, (A.1) then guarantees $\rho_{y,c}$-zCDP under label-only privatization whenever

$$\rho_{y,c} \ge \frac{B^2_{y,c}}{2\sigma^2_{y,\mathrm{DP},c}}. \tag{A.3}$$

Under joint feature and label privatization, composition gives a $(\rho_{y,c} + \rho_{X,c})$-zCDP guarantee, provided that (A.3) holds together with

$$\rho_{X,c} \ge \frac{d\, B^2_{X,c}}{2\sigma^2_{X,\mathrm{DP},c}}.$$

Thus, privacy budgets of constant order for labels and of linear order in d for features suffice to maintain fixed privacy noise variances while keeping clipping inactive, satisfying Assumptions (A4) and (A5).

Uniformly sub-Gaussian samples. Suppose now that, for some constants $a_{y,c}, a_{X,c} > 0$ independent of n, d, i, j, the actual responses and feature coordinates satisfy sub-Gaussian tail bounds

$$\mathbb{P}(|y_{c,i}| > t) \le 2\exp\left(-\frac{t^2}{2a_{y,c}^2}\right), \qquad \mathbb{P}(|X_{c,ij}| > t) \le 2\exp\left(-\frac{t^2}{2a_{X,c}^2}\right), \quad t \ge 0. \tag{A.4}$$

By (Vershynin, 2026, Proposition 2.6.1) we have that (A.4) implies the bounded moments of Assumption (A2), and thus, can be viewed as a specialization of this assumption. Let us now fix $q > 3$ and, for $n \ge 2$, choose

$$K_{y,c} = 2a_{y,c}\sqrt{2q\log n}, \qquad K_{X,c} = 2a_{X,c}\sqrt{2q\log n}. \tag{A.5}$$

A union bound and the sub-Gaussian assumption (A.4) then give

$$\mathbb{P}\left(\max_{i\le n_c}|y_{c,i}| > K_{y,c}/2\right) \le 2n_c n^{-q}.$$

Moreover, a feature row with norm exceeding $\sqrt{d}\,K_{X,c}/2$ must have a coordinate exceeding $K_{X,c}/2$ in magnitude. Hence another union bound gives

$$\mathbb{P}\left(\max_{i\le n_c}\|X_{c,i}\|_2 > \sqrt{d}\,K_{X,c}/2\right) \le 2n_c d\, n^{-q},$$

where $X_{c,i} \in \mathbb{R}^d$ denotes the i-th row of X_c. Since C is fixed and d is proportional to n,

$$\mathbb{P}(\mathcal{E}_n) \le 2\sum_{c\in\mathcal{A}} n_c(1+d)n^{-q} = O(n^{2-q}).$$

The right-hand side is summable because $q > 3$, proving (A.2). Since $\log n$ and $\log d$ are asymptotically equivalent, the clipping constants have orders

$$K_{y,c} \asymp \sqrt{\log d}, \qquad K_{X,c} \asymp \sqrt{\log d}.$$

Thus, from (A.1), we have that constant variances for the privacy noise are compatible with privacy budgets of orders $\log d$ for labels and $d\log d$ for features.

General samples. Assumptions (A2)-(A3) combined with the assumption on bounded noise moments imply that features and labels have bounded 8-th moments. To see this, independence, centering, and the uniformly bounded 8-th moments of the entries of Z_c, as assumed in (A2), imply that, for every deterministic $a \in \mathbb{R}^d$,

$$\mathbb{E}\left|\sum_{j=1}^d a_j Z_{c,ij}\right|^8 \le K_8\left(\sum_{j=1}^d a_j^2\right)^4, \tag{A.6}$$

where K_8 is a constant independent of c, i, n, d and a. Indeed, expanding the 8-th power eliminates every term containing a singleton factor and the remaining terms are bounded by a constant times $\|a\|_2^8$. Taking $a = \Sigma_c^{1/2}e_j$, where e_j is the j-th coordinate vector, gives

$$\mathbb{E}\,|X_{c,ij}|^8 \le K_8(\Sigma_{c,jj})^4 \le K_8\alpha_\Sigma^4.$$

Moreover, Assumption (A2) gives $\sup_{c,n,d}\|\beta_c\|_2 < \infty$ almost surely. Let us condition on this probability one event. Since β_c is independent of the design, (A.6) yields

$$\mathbb{E}\left[\left|x_{c,i}^\top\beta_c\right|^8 \,\middle|\, \beta_c\right] \le K_8\|\Sigma_c^{1/2}\beta_c\|_2^8 \le K_8\alpha_\Sigma^4\|\beta_c\|_2^8.$$

Using $y_{c,i} = X_{c,i}^\top\beta_c + \varepsilon_{c,i}$ and $|u+v|^8 \le 2^7(|u|^8+|v|^8)$, we obtain constants $M_y, M_X < \infty$ such that

$$\sup_{c,i,n,d}\mathbb{E}|y_{c,i}|^8 \le M_y, \qquad \sup_{c,i,j,n,d}\mathbb{E}|X_{c,ij}|^8 \le M_X.$$

For any fixed $\eta > 0$, choose

$$K_{y,c} = 2n_c^{1/4+\eta}, \qquad K_{X,c} = 2\,n_c^{3/8+\eta}. \tag{A.7}$$

Markov's inequality and a union bound give

$$\mathbb{P}\left(\max_{i\le n_c}|y_{c,i}| > K_{y,c}/2\right) \le M_y n_c^{-1-8\eta},$$

$$\mathbb{P}\left(\max_{i\le n_c}\|x_{c,i}\|_2 > \sqrt{d}\,K_{X,c}/2\right) \le \mathbb{P}\left(\max_{i\le n_c,\, j\le d}|X_{c,ij}| > n_c^{3/8+\eta}\right) \le M_X d\, n_c^{-2-8\eta}.$$

Since C is fixed and $n_c \asymp d \asymp n$,

$$\mathbb{P}(\mathcal{E}_n) \le \sum_{c\in\mathcal{A}} \left(M_y n_c^{-1-8\eta} + M_X d\, n_c^{-2-8\eta}\right) = O(n^{-1-8\eta}).$$

This bound is summable, so the first Borel-Cantelli lemma proves (A.2). Hence, constant variances for the privacy noise are compatible with privacy budgets of order $O(n_c^{1/2+2\eta})$ for labels and $O(d\, n_c^{3/4+2\eta})$ for features.

A.2 Proof of the bias-variance identities

For the unclipped estimator, let us set $w_1 = 1$ and $\tilde{\varepsilon}_1 = \varepsilon_1$, $\tilde{\varepsilon}_c = \varepsilon_c + \xi_c$, for $c \in \mathcal{A}$, with $s_1^2 = \sigma_1^2$, $s_c^2 = \sigma_{t,c}^2$, for $c \in \mathcal{A}$. Then

$$\tilde{y}_c = X_c L_c \theta + \tilde{\varepsilon}_c$$

for every client, where $\tilde{y}_1 = y_1$. By assumption, all entries of the vectors $\tilde{\varepsilon}_c$ ($c \in \mathcal{C}$) are independent between themselves and w.r.t. the designs and the signal, they are centered, they have variances s_c^2 and uniformly bounded 8-th moments. Define

$$\hat{u} \coloneqq \frac{1}{n}\sum_{c=1}^{C} w_c L_c^\top X_c^\top \tilde{\varepsilon}_c.$$

Since $\hat{U} = \hat{S}\theta + \hat{u}$, the closed form (3.5) gives

$$\hat{\theta} - \theta = -\hat{Q}\Lambda\theta + \hat{Q}\hat{u}. \tag{A.8}$$

Taking its conditional expectation and covariance yields

$$\mathbb{E}[\hat{\theta} \mid X, \theta] - \theta = -\hat{Q}\Lambda\theta,$$

$$\mathrm{Cov}(\hat{\theta} \mid X, \theta) = \frac{1}{n}\hat{Q}\left(\sigma_1^2\hat{S}_1 + \sum_{c\in\mathcal{A}} w_c^2\sigma_{t,c}^2\hat{S}_c\right)\hat{Q}, \tag{A.9}$$

where $X = (X_1, \dots, X_C)$. Then, we have

$$B(\hat{\theta};\theta) = \|\hat{Q}\Lambda\theta\|_{\hat{S}_1}^2 = \theta^\top\Lambda\hat{Q}S_1\hat{Q}\Lambda\theta.$$

Taking the trace of (A.9) against S_1 similarly gives

$$V(\hat{\theta};\theta) = \frac{1}{n}\,\mathrm{Tr}\left[S_1\hat{Q}\left(\sigma_1^2\hat{S}_1 + \sum_{c\in\mathcal{A}} w_c^2\sigma_{t,c}^2\hat{S}_c\right)\hat{Q}\right],$$

thus proving (3.7). □

B Main theoretical results

B.1 Proof of Theorem 4.1

The proof proceeds in six steps. First, we express the variance and, under Assumption (A6), the bias in terms of the scalar resolvent functional

$$F = \frac{1}{n}\,\mathrm{Tr}(S_1\hat{Q})$$

and its derivatives with respect to the ridge parameters and source weights. Second, a heterogeneous-column resolvent theorem gives a deterministic equivalent for F after introducing a strictly positive auxiliary regularization. Third, the regularization is removed. Fourth, the convergence of F is transferred to all first derivatives entering the bias and variance formulas. Fifth, we combine these results to obtain the variance limit and the bias limit under Assumption (A6). Finally, in the sixth step, we identify the resulting derivative expressions with the quantities g and $\mathcal{K}$ appearing in the theorem, and apply Proposition B.1 to the bias quadratic form to remove Assumption (A6) and obtain the general bias characterization.

Step 1: Reduction to one resolvent trace. As it will play an important role in the proof, we will introduce the notation $\mathcal{F} = e_1$. Let $\mathcal{J} = \{\beta, \delta_2 \dots, \delta_C\}$ and, for $j \in \mathcal{J}$, define the projection $P_j : \mathbb{R}^{Cd} \to \mathbb{R}^{Cd}$ onto the d coordinates of the j-th block, i.e.,

$$P_j v = \begin{cases} \begin{bmatrix} v_1 & \dots & v_d & 0 & \dots & 0 \end{bmatrix}, & \text{if } j = \beta, \\ \begin{bmatrix} 0 & \dots & 0 & v_{(k-1)d+1} & \dots & v_{kd} & 0 & \dots & 0 \end{bmatrix}, & \text{if } j = \delta_k, \end{cases}$$

for $v = \begin{bmatrix} v_1 & v_2 & \dots & v_{Cd} \end{bmatrix} \in \mathbb{R}^{Cd}$. We can rewrite the joint ridge term using these projections as $\Lambda = \sum_{j \in \mathcal{J}} \lambda_j P_j$. For convenience, we recall

$$\hat{S} = \hat{S}_1 + \sum_{c \in \mathcal{A}} w_c \hat{S}_c, \qquad \hat{Q} = (\hat{S} + \Lambda)^{-1},$$

and we refer to Section 3 for the rest of the notation. The central empirical resolvent functional for bias is defined as

$$F(\mathbf{w}, \lambda_\beta, \boldsymbol{\lambda}_\delta) \coloneqq \frac{1}{n} \operatorname{Tr}(S_1 \hat{Q}), \tag{B.1}$$

and the main formula for variance will be obtained from it and its first derivatives. Indeed, using the chain rule for $\vartheta \in \{\lambda_\beta, \lambda_{\delta_2}, \dots, \lambda_{\delta_c}, w_2, \dots, w_C\}$, we get

$$\partial_\vartheta \hat{Q} = -\hat{Q} \left(\partial_\vartheta \hat{Q}^{-1} \right) \hat{Q}.$$

Hence, for $j \in \mathcal{J}$ and $c \in \mathcal{A}$, we have

$$\partial_{\lambda_j} F = -\frac{1}{n} \operatorname{Tr}(S_1 \hat{Q} P_j \hat{Q}), \tag{B.2}$$

$$\partial_{w_c} F = -\frac{1}{n} \operatorname{Tr}(S_1 \hat{Q} \hat{S}_c \hat{Q}). \tag{B.3}$$

We first use these identities to rewrite the variance. Note that

$$\sigma_1^2 \hat{S}_1 + \sum_{c \in \mathcal{A}} w_c^2 \sigma_{t,c}^2 \hat{S}_c = \sigma_1^2 \hat{S} + \sum_{c \in \mathcal{A}} w_c (w_c \sigma_{t,c}^2 - \sigma_1^2) \hat{S}_c, \qquad \hat{Q} \hat{S} \hat{Q} = \hat{Q} - \hat{Q} \Lambda \hat{Q}.$$

Substituting this into (3.7), followed by (B.2) - (B.3) yields

$$V = \sigma_1^2 \left(F + \sum_{j \in \mathcal{J}} \lambda_j \partial_{\lambda_j} F \right) - \sum_{c \in \mathcal{A}} w_c (w_c \sigma_{t,c}^2 - \sigma_1^2) \partial_{w_c} F. \tag{B.4}$$

We next identify a similar representation for the bias. First, introduce $M \coloneqq \Lambda \hat{Q} S_1 \hat{Q} \Lambda$, and partition M into $d \times d$ blocks M_{jk} indexed by $j, k \in \mathcal{J}$. Defining $\theta_\beta = \beta$ and $\theta_{\delta_c} = \delta_c$ for $c \in \mathcal{A}$, we can expand the expression for the bias in (3.7) as

$$B = \sum_{j,k \in \mathcal{J}} \theta_j^\top M_{jk} \theta_k. \tag{B.5}$$

The signal blocks are independent and centered, so, conditionally on all data feature matrices,

$$\begin{aligned} \mathbb{E}_\theta[B \mid X_1, \dots, X_C] &= \sum_{j \in \mathcal{J}} \frac{r_j^2}{d} \operatorname{Tr}(M_{jj}) \\ &= \sum_{j \in \mathcal{J}} \frac{r_j^2 \lambda_j^2}{d} \operatorname{Tr}(S_1 \hat{Q} P_j \hat{Q}) \\ &= -\sum_{j \in \mathcal{J}} \tau_j \lambda_j^2 \partial_{\lambda_j} F, \end{aligned}$$

where last two equalities follow from the definition of M, cyclicity of the trace and (B.2). Therefore

$$\begin{aligned} \mathbb{E}_\theta[B \mid X_1, \dots, X_C] &= -\sum_{j \in \mathcal{J}} \tau_j \lambda_j^2 \partial_{\lambda_j} F \\ &= -\tau_\beta \lambda_\beta^2 \partial_{\lambda_\beta} F - \sum_{c \in \mathcal{A}} \tau_{\delta_c} \lambda_{\delta_c}^2 \partial_{\lambda_{\delta_c}} F. \end{aligned} \tag{B.6}$$

Step 2: Deterministic equivalent for F for strictly positive regularization. The matrix $\hat{S} + \Lambda$, showing up in definition of F, may have zero ridge parameters. We therefore first add an auxiliary scalar regularization, as it is technically simpler to work with. For $\eta > 0$, define

$$\Lambda_\eta \coloneqq \Lambda + \eta I_{Cd}, \qquad \hat{Q}_\eta \coloneqq (\hat{S} + \Lambda_\eta)^{-1}, \qquad F_\eta \coloneqq \frac{1}{n} \operatorname{Tr}(S_1 \hat{Q}_\eta).$$

We will apply the heterogeneous-column resolvent theorem (Wagner et al., 2012, Theorem 1). To this end, set $w_1 \coloneqq 1$, and let

$$\bar{Q}_\eta = \left[\Lambda_\eta + \sum_{c=1}^{C} \frac{\pi_c w_c S_c}{1 + e_{c,\eta}} \right]^{-1}, \tag{B.7}$$

where

$$e_{c,\eta} = \frac{w_c}{n} \operatorname{Tr}(S_c \bar{Q}_\eta), \qquad c = 1, \dots, C. \tag{B.8}$$

Next, notice that all of the assumptions of (Wagner et al., 2012, Theorem 1) are satisfied, as they follow from Assumptions (A1), (A2) and (A3). Thus, there is a unique solution to the fixed point system defined by (B.7)-(B.8) such that, for every deterministic $A_d \in \mathbb{R}^{Cd \times Cd}$ with uniformly bounded operator norm,

$$\frac{1}{n} \operatorname{Tr}\left[A_d (\hat{Q}_\eta - \bar{Q}_\eta) \right] \xrightarrow{\text{a.s.}} 0.$$

Taking $A_d = S_1$ and $\mathcal{F}_\eta \coloneqq \frac{1}{n} \operatorname{Tr}(S_1 \bar{Q}_\eta)$ yields

$$|F_\eta - \mathcal{F}_\eta| \xrightarrow{\text{a.s.}} 0. \tag{B.9}$$

Step 3: Removing the regularization. We prove that (B.9) holds for $\eta = 0$, that is

$$|F - \mathcal{F}| \xrightarrow{\text{a.s.}} 0. \tag{B.10}$$

We will first show that the objects F and $\mathcal{F}$ are indeed well-defined. Note that

$$\hat{\Sigma}_c = \pi_c \Sigma_c^{1/2} \left(\frac{1}{n_c} Z_c^\top Z_c \right) \Sigma_c^{1/2}.$$

By the Bai-Yin theorem (Bai & Silverstein, 2010, Theorem 5.11), there exist constants $0 < a < A < \infty$ such that, almost surely, for all sufficiently large n,

$$a I_d \preceq \hat{\Sigma}_c \preceq A I_d, \qquad c = 1, \dots, C. \tag{B.11}$$

We will now prove that this implies that $\hat{S}$ is positive definite. Let, $u = [u_1^\top, u_2^\top, \dots, u_C^\top]^\top \in \mathbb{R}^{Cd}$ be an arbitrary non-zero vector, where $u_i \in \mathbb{R}^d$, $i \in \mathcal{C}$. Then

$$u^\top \hat{S} u = u_1^\top \hat{\Sigma}_1 u_1 + \sum_{c \in \mathcal{A}} w_c (u_1 + u_c)^\top \hat{\Sigma}_c (u_1 + u_c).$$

On the event (B.11), it holds

$$u^\top \hat{S} u \geq a \|u_1\|_2^2 + a w_- \sum_{c \in \mathcal{A}} \|u_1 + u_c\|_2^2 > 0.$$

Moreover, as $\|u_c\|_2^2 \leq 2\|u_1\|_2^2 + 2\|u_1 + u_c\|_2^2$, it holds

$$\|u\|_2^2 \leq (2C - 1) \left(\|u_1\|_2^2 + \sum_{c \in \mathcal{A}} \|u_1 + u_c\|_2^2 \right).$$

This implies that $\hat{S}$ is positive definite, so there exists a constant $\kappa_{\mathrm{E}} > 0$, independent of n, such that

$$\hat{S} \succeq \kappa_{\mathrm{E}} I_{Cd} \tag{B.12}$$

almost surely for all large enough n. As $\hat{Q}_\eta = (\hat{S} + \Lambda_\eta)^{-1}$, (B.12) implies that almost surely

$$\sup_{\eta \geq 0} \|\hat{Q}_\eta\|_2 \leq \|\hat{S}^{-1}\|_2 \leq \kappa_{\mathrm{E}}^{-1} \tag{B.13}$$

even when some or all ridge parameters are zero. This means that the matrix $\hat{Q}$, and so F as well, is indeed well defined. Next, we turn our attention to $\bar{Q}$ and the corresponding fixed point equations. Fix an arbitrary $\eta > 0$. Then, (B.7) implies that, for $c \in \mathcal{C}$,

$$\bar{Q}_\eta^{-1} \succeq \sum_{c=1}^{C} \frac{\pi_c w_c S_c}{1 + e_{c,\eta}} \succeq \frac{\pi_c w_c S_c}{1 + e_{c,\eta}}.$$

Taking the trace on the range of S_c gives

$$\frac{\pi_c w_c}{1 + e_{c,\eta}} \operatorname{Tr}(S_c \bar{Q}_\eta) \leq \operatorname{rank}(S_c) = d.$$

As $e_{c,\eta}$ satisfies (B.8) and $n\pi_c = n_c$,

$$\frac{e_{c,\eta}}{1 + e_{c,\eta}} \leq \frac{d}{n_c}. \tag{B.14}$$

Let us choose $\gamma_* \in (\max_c \bar{\gamma}_c, 1)$. For all sufficiently large n,

$$\frac{1}{1 + e_{c,\eta}} \geq 1 - \gamma_*, \qquad c = 1, \ldots, C,$$

and therefore

$$\bar{Q}_\eta^{-1} \succeq (1 - \gamma_*) \sum_{c=1}^{C} \pi_c w_c S_c.$$

We can prove in the same way as for $\hat{S}$ that the matrix on the right is uniformly positive definite. Namely, there exists $\kappa_{\mathrm{DE}} > 0$, independent of n, such that

$$\bar{Q}_\eta^{-1} \succeq (1 - \gamma_*) \sum_{c=1}^{C} \pi_c w_c S_c \succeq \kappa_{\mathrm{DE}} I_{Cd} \tag{B.15}$$

uniformly in $\eta > 0$ for all sufficiently large n. This further implies that

$$\sup_{\eta > 0} \|\bar{Q}_\eta\|_2 \leq \kappa_{\mathrm{DE}}^{-1}. \tag{B.16}$$

By (B.14), for all sufficiently large n,

$$0 \leq e_{c,\eta} \leq \frac{\gamma_*}{1 - \gamma_*}, \qquad c \in \mathcal{C}, \quad \eta > 0.$$

Fix such an n and an arbitrary sequence $\eta_k \downarrow 0$. By compactness in $\mathbb{R}^C$, there exists a subsequence η_{k_i} along which the vector $(e_{1,\eta_{k_i}}, \ldots, e_{C,\eta_{k_i}})$ converges. We define

$$e_c \coloneqq \lim_{i \to \infty} e_{c,\eta_{k_i}}. \tag{B.17}$$

Note that by (B.7) - (B.8) it holds

$$\begin{aligned}
\lim_{i \to \infty} e_{c,\eta_{k_i}} &= \lim_{i \to \infty} \frac{w_c}{n} \operatorname{Tr}\left[S_c \left(\Lambda_{\eta_{k_i}} + \sum_{c=1}^{C} \frac{\pi_c w_c S_c}{1 + e_{c,\eta_{k_i}}} \right)^{-1} \right] \\
&= \frac{w_c}{n} \operatorname{Tr}\left[S_c \left(\lim_{i \to \infty} \Lambda_{\eta_{k_i}} + \sum_{c=1}^{C} \frac{\pi_c w_c S_c}{1 + \lim_{i \to \infty} e_{c,\eta_{k_i}}} \right)^{-1} \right] \\
&= \frac{w_c}{n} \operatorname{Tr}\left[S_c \left(\Lambda + \sum_{c=1}^{C} \frac{\pi_c w_c S_c}{1 + e_c} \right)^{-1} \right],
\end{aligned}$$

where the second equality comes from continuity and existence of the limits of a converging sequence. Thus, e_c solves the fixed point equations

$$\bar{Q} = \left[\Lambda + \sum_{c=1}^{C} \frac{\pi_c w_c S_c}{1 + e_c} \right]^{-1}, \qquad e_c = \frac{w_c}{n} \operatorname{Tr}(S_c \bar{Q}), \qquad \text{for } c \in \mathcal{C}.$$

The limiting matrix being inverted is bounded below by $\kappa_{\mathrm{DE}} I$ by (B.15), so $\bar{Q}$ is well-defined and positive definite. Moreover, the fixed point equations imply

$$e_c = \frac{w_c}{n} \operatorname{Tr}(S_c \bar{Q}) > 0, \qquad c \in \mathcal{C},$$

because $w_c > 0$, S_c is positive semidefinite and nonzero, and $\bar{Q}$ is positive definite.

Next, let us now prove that the positive fixed point of this equation is unique. To this end, let us define the fixed point map $\Phi : \mathbb{R}^C \to \mathbb{R}^C$ coordinate-wise as $\Phi \coloneqq (\Phi_1, \ldots, \Phi_C)$ where each coordinate function Φ_c, $c \in \{1, \ldots, C\}$, is defined as

$$\Phi_c(\mathbf{e}) = \frac{w_c}{n} \operatorname{Tr}\left[S_c \left(\Lambda + \sum_{c=1}^{C} \frac{\pi_c w_c S_c}{1 + e_c} \right)^{-1} \right], \tag{B.18}$$

for $\mathbf{e} = (e_1, \ldots, e_C)$. Note that Φ is componentwise increasing and $\Phi(\alpha \mathbf{e}) < \alpha \Phi(\mathbf{e})$, for every $\alpha > 1$. Assume by contradiction that there exist two solutions $\mathbf{e}^{(1)} > 0$ and $\mathbf{e}^{(2)} > 0$ such that

$$\Phi(\mathbf{e}^{(1)}) = \mathbf{e}^{(1)}, \qquad \Phi(\mathbf{e}^{(2)}) = \mathbf{e}^{(2)}.$$

Suppose, without loss of generality, that $\alpha = \max_{c \in \mathcal{C}} \frac{e_c^{(1)}}{e_c^{(2)}} > 1$. Then, for every $c \in \mathcal{C}$, $e_c^{(1)} \leq \alpha e_c^{(2)}$ and, by monotonicity, $\Phi(\mathbf{e}^{(1)}) \leq \Phi(\alpha \mathbf{e}^{(2)})$. As $\mathbf{e}^{(1)}$ is a fixed point, $\mathbf{e}^{(1)} \leq \Phi(\alpha \mathbf{e}^{(2)})$. Next, using that $\Phi(\alpha \mathbf{e}^{(2)}) < \alpha \Phi(\mathbf{e}^{(2)})$ and that $\mathbf{e}^{(2)}$ is a fixed point we get that $\mathbf{e}^{(1)} < \alpha \mathbf{e}^{(2)}$ holds componentwise. However, by definition of α, there exists c s.t. $e_c^{(1)} = \alpha e_c^{(2)}$. This gives a contradiction and proves uniqueness.

Having established existence and uniqueness of the fixed point at $\eta = 0$, fix a deterministic sequence $\eta_m = 1/m$, independent of n, for the remainder of the convergence argument. Using the triangle inequality and taking $\limsup_{n \to \infty}$, we have

$$\limsup_{n \to \infty} |F - \mathcal{F}| \leq \limsup_{n \to \infty} |F - F_{\eta_m}| + \limsup_{n \to \infty} |F_{\eta_m} - \mathcal{F}_{\eta_m}| + \limsup_{n \to \infty} |\mathcal{F}_{\eta_m} - \mathcal{F}|. \tag{B.19}$$

For the first term on the RHS, note that $\hat{Q}_\eta - \hat{Q} = -\eta \hat{Q}_\eta \hat{Q}$, so it follows

$$|F_\eta - F| = \left| \frac{1}{n} \operatorname{Tr}\left[S_1 (\hat{Q}_\eta - \hat{Q}) \right] \right| = \frac{\eta}{n} \left| \operatorname{Tr}\left[S_1 \hat{Q}_\eta \hat{Q} \right] \right|$$

Then, we can give a bound

$$\begin{aligned} |F_{\eta_m} - F| &\leq \frac{\eta_m d}{n} \|S_1\|_2 \, \|\hat{Q}_{\eta_m}\|_2 \, \|\hat{Q}\|_2 \\ &\leq \eta_m \gamma_* \alpha_\Sigma \kappa_{\mathrm{E}}^{-2}, \end{aligned} \tag{B.20}$$

where the last inequality holds almost surely for large enough n. For the second term of RHS, we have that from (B.9) it holds that almost surely

$$\limsup_{n \to \infty} |F_{\eta_m} - \mathcal{F}_{\eta_m}| = 0, \tag{B.21}$$

for each fixed η_m.

Lastly, we give a bound for the third term of the RHS. To this end, let us extend the definition of Φ from (B.18) to account for an arbitrary η parameter. Namely, we introduce $\Phi_\eta : \mathbb{R}^C \to \mathbb{R}^C$ as $\Phi_\eta = (\Phi_{\eta,1}, \ldots, \Phi_{\eta,C})$ where each coordinate function $\Phi_{\eta,c}$, $c \in \{1, \ldots, C\}$, is defined as

$$\Phi_{\eta,c}(\mathbf{e}) \coloneqq \frac{w_c}{n} \operatorname{Tr}\left[S_c \left(\Lambda + \eta I + \sum_{k=1}^{C} \frac{\pi_k w_k S_k}{1 + e_k} \right)^{-1} \right],$$

for $\mathbf{e} = (e_1, \ldots, e_C)$. One can readily verify that Φ_η is componentwise increasing as well as decreasing in η, that is, $\Phi_\eta(\mathbf{e}) \leq \Phi_0(\mathbf{e}) = \Phi(\mathbf{e})$. As before for Φ, we can prove the uniqueness of the fixed point for Φ_η.

Now, let us denote by $\mathbf{e}^{(0)} = (e_1^{(0)}, \dots, e_C^{(0)})$ and $\mathbf{e}_\eta = (e_{1,\eta}, \dots, e_{C,\eta})$ the fixed point of Φ and Φ_η, respectively, such that

$$\Phi_0(\mathbf{e}^{(0)}) = \Phi(\mathbf{e}^{(0)}) = \mathbf{e}^{(0)}, \qquad \Phi_\eta(\mathbf{e}_\eta) = \mathbf{e}_\eta.$$

Since Φ_η is componentwise increasing and decreasing in η, it holds

$$\Phi_\eta(\mathbf{e}^{(0)}) \leq \Phi(\mathbf{e}^{(0)}) = \mathbf{e}^{(0)}.$$

Thus, we define a sequence $\{\mathbf{e}^{(i)}\}_{i=0}^{\infty}$ such that

$$\mathbf{e}^{(i+1)} \coloneqq \Phi_\eta(\mathbf{e}^{(i)}),$$

which is decreasing, due to the componentwise monotonicity of Φ_η. As it is bounded below by 0, the sequence $\mathbf{e}^{(i)}$ converges componentwise to some $\mathbf{e}^* \geq 0$. By continuity of Φ_η, $\mathbf{e}^* = \Phi_\eta(\mathbf{e}^*)$, so uniqueness of the fixed point gives $\mathbf{e}^* = \mathbf{e}_\eta$. Therefore, from the monotonicity of the sequence, it holds

$$0 \leq e_{c,\eta} = \lim_{i\to\infty} e_c^{(i)} \leq e_c^{(0)}, \qquad c \in \mathcal{C}.$$

Let us now define the difference we would like to bound as

$$r_\eta \coloneqq \max_{1\leq c\leq C} \frac{e_c^{(0)} - e_{c,\eta}}{1 + e_c^{(0)}}.$$

Since $e_{c,\eta} \geq 0$, one has $0 \leq r_\eta < 1$. By definition it holds $1 + e_{c,\eta} \geq (1 - r_\eta)(1 + e_c^{(0)})$, and therefore

$$\frac{1}{1 + e_{c,\eta}} \leq \frac{1}{1 - r_\eta} \frac{1}{1 + e_c^{(0)}}.$$

Using the fixed point representations of $\bar{Q}_\eta$ and $\bar{Q}$, we obtain

$$\begin{aligned}
\bar{Q}_\eta^{-1} &= \Lambda + \eta I + \sum_{c=1}^{C} \frac{\pi_c w_c S_c}{1 + e_{c,\eta}} \\
&\preceq \Lambda + \eta I + \frac{1}{1 - r_\eta} \sum_{c=1}^{C} \frac{\pi_c w_c S_c}{1 + e_c^{(0)}} \\
&\preceq \frac{1}{1 - r_\eta} \bar{Q}^{-1} + \eta I.
\end{aligned}$$

By (B.16), and the fact that $\bar{Q}$ is well defined as the limit of $\bar{Q}_{\eta_{k_i}}$ where $\eta_{k_i} \downarrow 0$, it holds $\bar{Q}^{-1} \succeq \kappa_{\mathrm{DE}} I$, for all sufficiently large n. Therefore we get

$$\eta I \preceq \frac{\eta}{\kappa_{\mathrm{DE}}} \bar{Q}^{-1},$$

so that

$$\bar{Q}_\eta^{-1} \preceq \left(\frac{1}{1 - r_\eta} + \frac{\eta}{\kappa_{\mathrm{DE}}} \right) \bar{Q}^{-1}.$$

If we denote by q_η the reciprocal of the bracket coefficient, after inversion we get

$$\bar{Q}_\eta \succeq q_\eta \bar{Q}, \qquad q_\eta \coloneqq \frac{1 - r_\eta}{1 + \frac{\eta}{\kappa_{\mathrm{DE}}}(1 - r_\eta)}. \tag{B.22}$$

Then it directly follows that

$$e_{c,\eta} \geq q_\eta e_c^{(0)},$$

and so

$$\frac{e_c^{(0)} - e_{c,\eta}}{1 + e_c^{(0)}} \leq (1 - q_\eta) \frac{e_c^{(0)}}{1 + e_c^{(0)}}.$$

From (B.14), and the fact that $e_c^{(0)}$ is well defined as the limit of $e_{c,\eta_{k_i}}$ where $\eta_{k_i} \downarrow 0$, for all sufficiently large n,

$$\frac{e_c^{(0)}}{1 + e_c^{(0)}} \leq \frac{d}{n_c} \leq \gamma_* < 1.$$

Taking the maximum over c gives

$$r_\eta \le \gamma_*(1 - q_\eta).$$

Writing $a_\eta \coloneqq \eta/\kappa_{\mathrm{DE}}$, we have

$$1 - q_\eta = \frac{r_\eta + a_\eta(1 - r_\eta)}{1 + a_\eta(1 - r_\eta)} \le r_\eta + a_\eta.$$

Therefore

$$r_\eta \le \gamma_*(r_\eta + a_\eta),$$

which implies

$$r_\eta \le \frac{\gamma_*}{1 - \gamma_*} \frac{\eta}{\kappa_{\mathrm{DE}}}. \tag{B.23}$$

By definition it holds that $\mathcal{F} = e_1^{(0)}$ and $\mathcal{F}_\eta = e_{1,\eta}$. Moreover, (B.14) implies

$$1 + e_1^{(0)} \le \frac{1}{1 - \gamma_*}.$$

Combining this with (B.23) yields

$$0 \le |\mathcal{F} - \mathcal{F}_\eta| \le \eta \, \frac{\gamma_*}{\kappa_{\mathrm{DE}}(1 - \gamma_*)^2}. \tag{B.24}$$

Finally, plugging into (B.19) the bounds from (B.20), (B.21) and (B.24) we get that for each η_m it holds almost surely

$$\limsup_{n\to\infty} |F - \mathcal{F}| \le \eta_m \gamma_* \alpha_\Sigma \kappa_{\mathrm{E}}^{-2} + \eta_m \, \frac{\gamma_*}{\kappa_{\mathrm{DE}}(1 - \gamma_*)^2}.$$

As a countable intersection of probability 1 sets is of probability 1, the previous inequality holds with probability 1 over all η_m simultaneously. Taking $\lim_{m\to\infty}$ finally gives

$$\limsup_{n\to\infty} |F - \mathcal{F}| = \lim_{m\to\infty} \limsup_{n\to\infty} |F - \mathcal{F}| \le \lim_{m\to\infty} \eta_m \left(\gamma_* \alpha_\Sigma \kappa_{\mathrm{E}}^{-2} + \frac{\gamma_*}{\kappa_{\mathrm{DE}}(1 - \gamma_*)^2} \right) = 0,$$

almost surely, proving the desired claim.

Step 4: Deterministic equivalents for first derivatives of F. We now transfer the convergence $F - \mathcal{F} \xrightarrow{\text{a.s.}} 0$ to the first derivatives appearing in (B.4) and (B.6). Let

$$\mathcal{P} \coloneqq \{\lambda_\beta\} \cup \{\lambda_{\delta_c} : c \in \mathcal{A}\} \cup \{w_c : c \in \mathcal{A}\}.$$

We will prove that, for every $\vartheta \in \mathcal{P}$,

$$\partial_\vartheta F - \partial_\vartheta \mathcal{F} \xrightarrow{\text{a.s.}} 0. \tag{B.25}$$

When ϑ is a ridge parameter equal to zero, the derivative is understood as the right derivative.

Take $\vartheta \in \mathcal{P}$ and keep all other parameters fixed. We write t_0 for the value of ϑ at the parameter tuple under consideration, and regard F and $\mathcal{F}$, and any other objects depending on ϑ, as functions of the scalar variable t obtained by replacing ϑ by t. Let us choose an arbitrary $h_0 > 0$ and consider the one-sided interval

$$\mathcal{I}_\vartheta = [t_0, t_0 + h_0].$$

The interval can be chosen sufficiently small that all weights remain bounded above and bounded away from zero.

We first establish a uniform second-derivative bound for the empirical quantity F. On the almost-sure event (B.12), there exists a constant $\kappa_{\mathrm{E}} > 0$, independent of n, such that, almost surely for all sufficiently large n,

$$\hat{S} + \Lambda(t) \succeq \hat{S} \succeq \kappa_{\mathrm{E}} I_{Cd}.$$

Note that the same lower bound holds uniformly for $t \in \mathcal{I}_\vartheta$: increasing a ridge parameter adds a positive semidefinite matrix, while increasing a weight adds a positive multiple of $\hat{S}_c$. Hence,

$$\sup_{t \in \mathcal{I}_\vartheta} \|\hat{Q}(t)\|_2 \le \kappa_{\mathrm{E}}^{-1}$$

almost surely for all sufficiently large n.

Differentiating the resolvent identity gives, for a ridge parameter λ_j and a weight w_c, respectively,

$$\partial^2_{\lambda_j} F = \frac{2}{n} \operatorname{Tr}\Big(S_1 \hat{Q} P_j \hat{Q} P_j \hat{Q}\Big), \qquad \partial^2_{w_c} F = \frac{2}{n} \operatorname{Tr}\Big(S_1 \hat{Q} \hat{S}_c \hat{Q} \hat{S}_c \hat{Q}\Big).$$

As S_1, P_j, $\hat{S}_c$, and $\hat{Q}(t)$ are uniformly bounded in operator norm, there exists a constant $K_1 < \infty$ such that

$$\sup_{t \in \mathcal{I}_\vartheta} \big|\partial_t^2 F(t)\big| \le K_1 \tag{B.26}$$

almost surely for all sufficiently large n.

We next establish an analogous bound for $\mathcal{F}$. Recall that $\mathcal{F} = \mathcal{F}(t) = e_1(t)$, where $\mathbf{e}(t) = (e_1(t), \dots, e_C(t))$ is the unique fixed point of

$$\mathbf{e}(t) = \Phi(\mathbf{e}(t), t),$$

for Φ defined in (B.18) extended to include dependence on parameter t. We will prove that the deterministic fixed point $\mathbf{e}(t)$ is locally smooth in the parameter t. To make this precise, let us make the dependence on t explicit by denoting for an arbitrary $\mathbf{e} = (e_1, \dots, e_C)$

$$Q(\mathbf{e}, t) \coloneqq \left(\Lambda(t) + \sum_{k=1}^{C} \frac{\pi_k w_k(t) S_k}{1 + e_k}\right)^{-1}, \tag{B.27}$$

so $\Phi_c(\mathbf{e}, t) = \frac{w_c(t)}{n} \operatorname{Tr}\left(S_c Q(\mathbf{e}, t)\right)$. Note that, along the fixed point solution $\mathbf{e}(t)$, it holds

$$\bar{Q}(t) = Q(\mathbf{e}(t), t),$$

We will also occasionally drop dependence on t and $\mathbf{e}$ to simplify exposition. Recall that only the parameter corresponding to ϑ varies with t, all remaining ridge parameters and weights are kept fixed. The argument leading to (B.15), applied uniformly over $t \in \mathcal{I}_\vartheta$, gives a constant $\kappa_{\mathrm{DE}} > 0$, independent of n, such that, for all sufficiently large n,

$$\sup_{t \in \mathcal{I}_\vartheta} \|Q(\mathbf{e}(t), t)\|_2 = \sup_{t \in \mathcal{I}_\vartheta} \|\bar{Q}(t)\|_2 \le \kappa_{\mathrm{DE}}^{-1}. \tag{B.28}$$

All bounds on partial derivatives below are evaluated at $(\mathbf{e}, t) = (\mathbf{e}(t), t)$, with $\mathbf{e}$ and t treated as independent variables when taking these derivatives. Moreover, from (B.14),

$$\frac{e_c(t)}{1 + e_c(t)} \le \frac{d}{n_c} \le \gamma_* < 1,$$

uniformly for $t \in \mathcal{I}_\vartheta$ and all sufficiently large n. Hence

$$0 \le e_c(t) \le \frac{\gamma_*}{1 - \gamma_*}, \qquad c = 1, \dots, C. \tag{B.29}$$

In particular, the factors $(1 + e_c(t))^{-1}$ and their first two derivatives with respect to e_c are uniformly bounded.

We next verify that the fixed point map Φ has uniformly bounded partial derivatives up to second order. The basic identity used throughout is

$$\partial_x Q = -Q(\partial_x Q^{-1})Q, \tag{B.30}$$

for any scalar variable x. For example,

$$\partial_{e_k} Q^{-1} = -\frac{\pi_k w_k S_k}{(1 + e_k)^2},$$

and therefore

$$\partial_{e_k} Q = Q \frac{\pi_k w_k S_k}{(1 + e_k)^2} Q.$$

Similarly,

$$\partial_{e_k e_k} Q^{-1} = \frac{2\pi_k w_k S_k}{(1 + e_k)^3}, \qquad \partial_{e_k e_\ell} Q^{-1} = 0, \quad k \ne \ell.$$

If t is a ridge parameter λ_j, then, with $\mathbf{e}$ kept fixed,

$$\partial_t Q^{-1} = P_j, \qquad \partial_{tt} Q^{-1} = 0.$$

If t is a weight w_r, then

$$\partial_t Q^{-1} = \frac{\pi_r S_r}{1 + e_r}, \qquad \partial_{tt} Q^{-1} = 0,$$

again with $\mathbf{e}$ held fixed. Mixed derivatives such as $\partial_{te_k} Q^{-1}$ are either zero or a scalar multiple of S_k with a uniformly bounded coefficient. Repeated use of (B.30) therefore shows that every first or second partial derivative of $Q(\mathbf{e}, t)$ with respect to $(\mathbf{e}, t)$ is a finite sum of products of the forms QA_1QA_2Q and QA_1Q, where each A_i is one of P_j, S_k, or a uniformly bounded scalar multiple thereof. The matrices S_k have uniformly bounded operator norms by (A3), while (B.28) uniformly bounds $Q(\mathbf{e}, t)$. Consequently, all these first and second derivatives of $Q(\mathbf{e}, t)$ have uniformly bounded operator norm. The same is true for the derivatives of Φ. Indeed,

$$\Phi_c(\mathbf{e}, t) = \frac{w_c(t)}{n} \operatorname{Tr}(S_c Q(\mathbf{e}, t)).$$

Since $\operatorname{rank}(S_c) \leq d$, for any matrix A with uniformly bounded operator norm,

$$\frac{1}{n} |\operatorname{Tr}(S_c A)| \leq \frac{d}{n} \|S_c\|_2 \|A\|_2 \leq K_2,$$

for some constant $K_2 < \infty$ independent of n. Thus, for $t \in \mathcal{I}_\vartheta$,

$$\|\partial_t \Phi(\mathbf{e}, t)\|_2 + \|\partial_{tt} \Phi(\mathbf{e}, t)\|_2 + \|\partial_{t\mathbf{e}} \Phi(\mathbf{e}, t)\|_2 + \|\partial_{\mathbf{ee}} \Phi(\mathbf{e}, t)\|_2 \leq K_3, \tag{B.31}$$

for a constant $K_3 < \infty$ independent of n, where $\partial_{\mathbf{e}}$ denotes differentiation with respect to the vector $\mathbf{e} = (e_1, \ldots, e_C)$ applied entrywise to the object that follows.

Define the Jacobian evaluated at the fixed point by

$$J(t) := \partial_{\mathbf{e}} \Phi(\mathbf{e}(t), t).$$

Its entries are

$$J_{ck}(t) = \frac{\pi_k w_c(t) w_k(t)}{n(1 + e_k(t))^2} \operatorname{Tr}\left(S_c \bar{Q}(t) S_k \bar{Q}(t)\right) \geq 0. \tag{B.32}$$

We suppress the dependence on t when no confusion can arise. Let $\mathbf{v} = (v_1, \ldots, v_C) \in \mathbb{R}^C$ be such that

$$v_c \coloneqq 1 + e_c, \qquad c = 1, \ldots, C.$$

Then, we have

$$(Jv)_c = \frac{w_c}{n} \operatorname{Tr}\left[S_c \bar{Q} \left(\sum_{k=1}^{C} \frac{\pi_k w_k S_k}{1 + e_k}\right) \bar{Q}\right] = \frac{w_c}{n} \operatorname{Tr}\left[S_c \bar{Q} \left(\bar{Q}^{-1} - \Lambda\right) \bar{Q}\right] = e_c - \frac{w_c}{n} \operatorname{Tr}(S_c \bar{Q} \Lambda \bar{Q}) \leq e_c. \tag{B.33}$$

Therefore, in the weighted maximum norm $\|x\|_v \coloneqq \max_{1 \leq c \leq C} \frac{|x_c|}{1 + e_c}$, we have that, for all sufficiently large n,

$$\|J(t)\|_v = \max_c \frac{(J(t)v)_c}{1 + e_c} \leq \max_c \frac{e_c(t)}{1 + e_c(t)} \leq \max_c \frac{d}{n_c} \leq \gamma_* < 1,$$

where the bound on $e_c(t)/(1 + e_c(t))$ is due to (B.14). Consequently, $I_C - J(t)$ is invertible, as

$$\|(I_C - J(t))^{-1}\|_v \leq \frac{1}{1 - \gamma_*}, \tag{B.34}$$

for $t \in \mathcal{I}_\vartheta$. By (B.29), $1 \leq v_c(t) \leq (1 - \gamma_*)^{-1}$. Consequently, uniformly in n and t,

$$\|x\|_{v(t)} \leq \|x\|_2 \leq \frac{\sqrt{C}}{1 - \gamma_*} \|x\|_{v(t)}.$$

Combining this with (B.34) yields

$$\sup_{t \in \mathcal{I}_\vartheta} \|(I_C - J(t))^{-1}\|_2 \leq \frac{\sqrt{C}}{(1 - \gamma_*)^2}. \tag{B.35}$$

We can now control the derivatives of the fixed point itself. Let us write

$$G(\mathbf{e}, t) \coloneqq \mathbf{e} - \Phi(\mathbf{e}, t).$$

Then, the fixed point relation is

$$G(\mathbf{e}(t), t) = 0,$$

and its Jacobian with respect to $\mathbf{e}$ is

$$\partial_{\mathbf{e}} G = I_C - J,$$

Fix n sufficiently large and $t_* \in \mathcal{I}_\vartheta$. At $(\mathbf{e}(t_*), t_*)$, all denominators $1 + e_c(t_*)$ are positive and the matrix defining $Q(\mathbf{e}(t_*), t_*)$ is positive definite. Hence Q, Φ, and G are smooth on an open neighborhood of this point in $\mathbb{R}^C \times \mathbb{R}$, including when $t_* = 0$. Note that $\partial_{\mathbf{e}} G$ is invertible, so the implicit function theorem gives a local twice continuously differentiable branch of solutions. Since $\mathbf{e}(t_*)$ is componentwise positive, this branch remains positive after shrinking the neighborhood. Uniqueness of the positive fixed point identifies it with $\mathbf{e}(t)$ for nearby $t \in \mathcal{I}_\vartheta$. Thus, $\mathbf{e}(t)$ is twice continuously differentiable on $\mathcal{I}_\vartheta$, with derivatives at the endpoints interpreted one-sidedly. Differentiating $\mathbf{e}(t) = \Phi(\mathbf{e}(t), t)$ gives

$$(I_C - J)\mathbf{e}'(t) = \partial_t \Phi. \tag{B.36}$$

Therefore,

$$\|\mathbf{e}'(t)\|_2 \le \|(I_C - J)^{-1}\|_2 \|\partial_t \Phi\|_2.$$

The inverse is uniformly bounded by (B.35), while $\partial_t \Phi$ is uniformly bounded by (B.31). Hence there exists $K_4 < \infty$, independent of n, such that

$$\sup_{t \in \mathcal{I}_\vartheta} \|\mathbf{e}'(t)\| \le K_4. \tag{B.37}$$

Differentiating (B.36) once more gives

$$(I_C - J)\mathbf{e}''(t) = \partial_{tt}\Phi + 2\,\partial_{t\mathbf{e}}\Phi\,\mathbf{e}'(t) + \partial_{\mathbf{ee}}\Phi[\mathbf{e}'(t), \mathbf{e}'(t)]. \tag{B.38}$$

Note that here $\partial_{\mathbf{ee}}\Phi[x, y]$ denotes the vector obtained by applying the Hessian of each component of Φ to the pair (x, y). By (B.31) and (B.37), every term on the right-hand side of (B.38) is uniformly bounded. Applying again the uniform bound on $(I_C - J)^{-1}$ coming from (B.35) yields a constant $K_5 < \infty$, independent of n, such that

$$\sup_{t \in \mathcal{I}_\vartheta} \|\mathbf{e}''(t)\| \le K_5. \tag{B.39}$$

Since $\mathcal{F}(t) = e_1(t)$, we conclude that, for some constant $K_6 < \infty$ independent of n,

$$\sup_{t \in \mathcal{I}_\vartheta} \left|\partial_t^2 \mathcal{F}(t)\right| \le K_6. \tag{B.40}$$

We can now transfer the convergence of F and $\mathcal{F}$ to their first derivatives. Let $0 < h < h_0$. By Taylor's theorem, for some $\xi_F, \xi_{\mathcal{F}} \in (t_0, t_0 + h)$, it holds

$$F(t_0 + h) = F(t_0) + h\,\partial_t F(t_0) + \frac{h^2}{2}\partial_t^2 F(\xi_F),$$
$$\mathcal{F}(t_0 + h) = \mathcal{F}(t_0) + h\,\partial_t \mathcal{F}(t_0) + \frac{h^2}{2}\partial_t^2 \mathcal{F}(\xi_{\mathcal{F}}).$$

Dividing by h and using (B.26) and (B.40) gives

$$\left|\partial_t F(t_0) - \frac{F(t_0 + h) - F(t_0)}{h}\right| \le \frac{K_1}{2} h,$$
$$\left|\partial_t \mathcal{F}(t_0) - \frac{\mathcal{F}(t_0 + h) - \mathcal{F}(t_0)}{h}\right| \le \frac{K_6}{2} h.$$

By adding and subtracting the two finite-difference quotients, we obtain

$$|\partial_t F(t_0) - \partial_t \mathcal{F}(t_0)| \le \frac{K_1 + K_6}{2} h + \frac{|F(t_0 + h) - \mathcal{F}(t_0 + h)| + |F(t_0) - \mathcal{F}(t_0)|}{h}.$$

For every fixed $h > 0$, both parameter values t_0 and $t_0 + h$ are fixed. Hence the already established deterministic equivalent convergence gives

$$F(t_0) - \mathcal{F}(t_0) \xrightarrow{\text{a.s.}} 0, \qquad F(t_0 + h) - \mathcal{F}(t_0 + h) \xrightarrow{\text{a.s.}} 0.$$

Therefore, for fixed h,

$$\limsup_{n\to\infty} |\partial_t F(t_0) - \partial_t \mathcal{F}(t_0)| \le \frac{K_1 + K_6}{2} h$$

almost surely.

Finally, let us take a deterministic sequence $h_m \downarrow 0$. Since this sequence is countable, we may intersect the probability-one events corresponding to all parameter values $t_0 + h_m$. On this common event, the preceding inequality holds for every m. Sending $m \to \infty$ then proves (B.25).

Step 5: Completing variance limit and bias limit under Assumption (A6)**.** For the variance, combining (B.4), (B.10) and (B.25) immediately gives

$$\left|V - \bar{\mathcal{V}}\right| \xrightarrow{\text{a.s.}} 0, \tag{B.41}$$

where $\bar{\mathcal{V}} = \sigma_1^2 \left(\mathcal{F} + \sum_{j\in\mathcal{J}} \lambda_j \partial_{\lambda_j} \mathcal{F}\right) - \sum_{c\in\mathcal{A}} w_c (w_c s_c^2 - \sigma_1^2) \partial_{w_c} \mathcal{F}$.

For the bias, recall from (B.5) that

$$B = \sum_{j,k\in\mathcal{J}} \theta_j^\top M_{jk} \theta_k, \qquad M = \Lambda \hat{Q} S_1 \hat{Q} \Lambda,$$

where M_{jk} denotes the (j,k) block of M. By (B.12), $\hat{Q}$ is uniformly bounded almost surely for all sufficiently large n. Since Λ and S_1 are uniformly bounded, there exists a deterministic constant $K_M < \infty$ such that, almost surely for all sufficiently large n,

$$\max_{j,k\in\mathcal{J}} \|M_{jk}\|_2 \le K_M. \tag{B.42}$$

For each $j \in \mathcal{J}$, by the assumption in (4.1), we can write

$$\theta_j = \frac{r_j}{\sqrt{d}} x_j,$$

where the coordinates of x_j are independent, centered, have unit variance, and have uniformly bounded eighth moments. Moreover, the vectors x_j, $j \in \mathcal{J}$, are independent of one another and of the design matrices.

Let us condition on the designs. For a diagonal block $j = k$, define

$$D_j \coloneqq \theta_j^\top M_{jj} \theta_j - \frac{r_j^2}{d} \operatorname{Tr}(M_{jj}).$$

An application of (Bai & Silverstein, 2010, Lemma B.26) gives

$$\mathbb{E}_\theta \left[\left| x_j^\top M_{jj} x_j - \operatorname{Tr}(M_{jj}) \right|^4 \,\middle|\, X_1, \ldots, X_C \right] \le K_1 \left[\operatorname{Tr}(M_{jj} M_{jj}^\top) \right]^2,$$

for a constant K_1. Using (B.42), we get

$$\operatorname{Tr}(M_{jj} M_{jj}^\top) \le d \|M_{jj}\|_2^2 \le d K_M^2.$$

Hence, since the r_j are fixed,

$$\mathbb{E}_\theta \left[|D_j|^4 \,\middle|\, X_1, \ldots, X_C \right] \le \frac{K_2}{d^2} \tag{B.43}$$

almost surely for all sufficiently large n for a constant K_2. For two distinct blocks $j \neq k$, we have

$$\mathbb{E}_\theta \left[\left| x_j^\top M_{jk} x_k \right|^4 \,\middle|\, X_1, \ldots, X_C \right] \le K_3 \left[\operatorname{Tr}(M_{jk} M_{jk}^\top) \right]^2.$$

Therefore,

$$\mathbb{E}_\theta \left[\left| \theta_j^\top M_{jk} \theta_k \right|^4 \,\middle|\, X_1, \ldots, X_C \right] \le \frac{K_4}{d^2}, \qquad j \neq k. \tag{B.44}$$

Now, by the independence and centering of the signal blocks,

$$\mathbb{E}_\theta[B \mid X_1, \dots, X_C] = \sum_{j \in \mathcal{J}} \frac{r_j^2}{d} \operatorname{Tr}(M_{jj}),$$

and consequently

$$B - \mathbb{E}_\theta[B \mid X_1, \dots, X_C] = \sum_{j \in \mathcal{J}} D_j + \sum_{\substack{j,k \in \mathcal{J} \\ j \neq k}} \theta_j^\top M_{jk} \theta_k.$$

Since C is fixed, the number of terms in the two sums is uniformly bounded. Thus, using (B.43) and (B.44),

$$\mathbb{E}_\theta \left[|B - \mathbb{E}_\theta[B \mid X_1, \dots, X_C]|^4 \,\middle|\, X_1, \dots, X_C \right] \leq \frac{K_5}{d^2}$$

almost surely for all sufficiently large n.

Let us fix $\varepsilon > 0$. Markov's inequality then gives

$$\mathbb{P}_\theta \left(|B - \mathbb{E}_\theta[B \mid X_1, \dots, X_C]| > \varepsilon \,|\, X_1, \dots, X_C \right) \leq \frac{K_6}{\varepsilon^4 d^2}.$$

As $\sum_{n=1}^{\infty} \frac{1}{d^2} < \infty$, for almost every realization of the design sequence, the conditional Borel-Cantelli lemma gives

$$B - \mathbb{E}_\theta[B \mid X_1, \dots, X_C] \xrightarrow{\text{a.s.}} 0. \tag{B.45}$$

Since the signal vectors are independent of the designs, this conditional almost-sure statement is also an almost-sure statement with respect to the joint probability space generating both the designs and the signal vectors.

Finally, from (B.6),

$$\mathbb{E}_\theta[B \mid X_1, \dots, X_C] = -\sum_{j \in \mathcal{J}} \tau_j \lambda_j^2 \partial_{\lambda_j} F.$$

By (B.25), we obtain that

$$\mathbb{E}_\theta[B \mid X_1, \dots, X_C] - \bar{\mathcal{B}} \xrightarrow{\text{a.s.}} 0, \tag{B.46}$$

where $\bar{\mathcal{B}} = -\sum_{j \in \mathcal{J}} \tau_j \lambda_j^2 \partial_{\lambda_j} \mathcal{F}$. Combining (B.45) and (B.46) yields

$$\left| B - \bar{\mathcal{B}} \right| \leq |B - \mathbb{E}_\theta[B \mid X_1, \dots, X_C]| + \left| \mathbb{E}_\theta[B \mid X_1, \dots, X_C] - \bar{\mathcal{B}} \right| \xrightarrow{\text{a.s.}} 0,$$

which, together with (B.41), gives the deterministic equivalent formulas for bias and variance under Assumption (A6). In the last step that follows, we simplify deterministic equivalent formulas and extend the bias characterization to the general signal setting without Assumption (A6).

Step 6: Extension to general signals and identification of the deterministic equivalents. Let $\bar{Q}_{S_1}(t)$ and $\mathbf{e}(t) = (e_{c,S_1}(t))_{c=1}^C$ be the deterministic quantities in Proposition B.1 with $A = S_1$. At $t = 0$, these are exactly $\bar{Q}$ and $\mathbf{e} = (e_c)_{c=1}^C$ from (4.2)-(4.3). Differentiating the fixed point equations (B.55) gives

$$e'_{c,S_1}(0) = -\frac{w_c}{n} \operatorname{Tr}(S_c \bar{Q} S_1 \bar{Q}) + \sum_{k=1}^{C} \frac{\pi_k w_c w_k}{n(1+e_k)^2} \operatorname{Tr}(S_c \bar{Q} S_k \bar{Q}) e'_{k,S_1}(0).$$

Thus, $(I_C - J)\mathbf{e}'(0) = -b$, with J and b as in (4.4). By the invertibility established in (B.35) in Step 4,

$$\mathbf{e}'(0) = -(I_C - J)^{-1} b = -g.$$

Differentiating the matrix inverse $\bar{Q}_{S_1}(t)$ therefore yields

$$-\frac{d}{dt} \bar{Q}_{S_1}(t) \Big|_{t=0} = \bar{Q} \left[S_1 + \sum_{c=1}^{C} \frac{\pi_c w_c g_c}{(1+e_c)^2} S_c \right] \bar{Q} = \mathcal{K}, \qquad g_c = \frac{w_c}{n} \operatorname{Tr}(S_c \mathcal{K}). \tag{B.47}$$

Let us first fix a deterministic, uniformly bounded signal sequence θ, and set $a = \Lambda\theta$. The uniform bounds on Λ and θ give $\sup_n \|a\|_2 < \infty$. Moreover, $S_1 \succeq 0$ is deterministic and uniformly bounded. Thus, we can apply Proposition B.1 with $a = \Lambda\theta$ and $A = S_1$. Namely, for

$$H_{S_1}(t) = a^\top(\hat{S} + \Lambda + tS_1)^{-1}a, \qquad \mathcal{H}_{S_1}(t) = a^\top \bar{Q}_{S_1}(t)a,$$

it follows from (B.57) that

$$\frac{d}{dt}H_{S_1}(t)\Big|_{t=0} - \frac{d}{dt}\mathcal{H}_{S_1}(t)\Big|_{t=0} \xrightarrow{\text{a.s.}} 0. \tag{B.48}$$

On the other hand, the exact bias identity (3.7) and (B.47) give

$$B = -\frac{d}{dt}H_{S_1}(t)\Big|_{t=0}, \qquad -\frac{d}{dt}\mathcal{H}_{S_1}(t)\Big|_{t=0} = \theta^\top \Lambda\mathcal{K}\Lambda\theta, \tag{B.49}$$

and thus, we have proved the convergence of bias in (4.7). Recall that θ is almost surely uniformly bounded by Assumption (A2). Hence, we can apply Proposition B.1 for almost every realization, which yields that (B.48) holds under the law of θ.

It remains to identify the scalar deterministic equivalents obtained in the preceding five steps with the simplified expressions in the theorem. Define

$$D = \operatorname{diag}\left(\frac{\pi_c}{(1+e_c)^2}\right)_{c=1}^{C}, \qquad T_{ck} = \frac{w_c w_k}{n}\operatorname{Tr}(S_c\bar{Q}S_k\bar{Q}),$$

and let $\mathbf{u}_c$ be the c-th coordinate vector of $\mathbb{R}^C$. Then $T = T^\top$, $J = TD$, and $b = T\mathbf{u}_1$, since $w_1 = 1$. The identity

$$(I_C - TD)^{-1}T = T(I_C - DT)^{-1}$$

shows that $(I_C - J)^{-1}T$ is symmetric. Its first column is g, so

$$\mathbf{u}_1^\top(I_C - J)^{-1} = \mathbf{u}_1^\top + g^\top D. \tag{B.50}$$

For $j \in \mathcal{J}$, set $q_c^{(j)} = w_c \operatorname{Tr}(S_c\bar{Q}P_j\bar{Q})/n$. Differentiating the fixed point equations with respect to a ridge parameter and an auxiliary weight, respectively, gives

$$(I_C - J)\partial_{\lambda_j}\mathbf{e} = -q^{(j)}, \qquad (I_C - J)\partial_{w_c}\mathbf{e} = \frac{1}{w_c}\left[e_c\mathbf{u}_c - (1+e_c)J\mathbf{u}_c\right], \quad c \in \mathcal{A}.$$

Taking the first coordinate and using (B.50), cyclicity of the trace, and (4.5), we obtain

$$-\partial_{\lambda_j}\mathcal{F} = (\mathbf{u}_1^\top + g^\top D)q^{(j)} = \frac{1}{n}\operatorname{Tr}(P_j\mathcal{K}), \qquad j \in \mathcal{J}, \tag{B.51}$$

$$-\partial_{w_c}\mathcal{F} = \frac{[(I_C - J)^{-1}]_{1c}}{w_c} = \frac{\pi_c g_c}{w_c(1+e_c)^2}, \qquad c \in \mathcal{A}. \tag{B.52}$$

Consequently, under Assumption (A6), the scalar bias equivalent established in Step 5 is

$$\bar{\mathcal{B}} = -\sum_{j\in\mathcal{J}}\tau_j\lambda_j^2\partial_{\lambda_j}\mathcal{F} = \frac{1}{n}\sum_{j\in\mathcal{J}}\tau_j\lambda_j^2\operatorname{Tr}(P_j\mathcal{K}) = \mathcal{B}, \tag{B.53}$$

which is exactly the bias formula (4.8) in the theorem.

Finally, evaluating $n^{-1}\operatorname{Tr}(\bar{Q}^{-1}\mathcal{K})$ using, respectively, (4.5) and (4.2), together with (B.47), gives

$$\frac{1}{n}\operatorname{Tr}(\bar{Q}^{-1}\mathcal{K}) = \mathcal{F} + \sum_{c=1}^{C}\frac{\pi_c e_c g_c}{(1+e_c)^2} = \frac{1}{n}\operatorname{Tr}(\Lambda\mathcal{K}) + \sum_{c=1}^{C}\frac{\pi_c g_c}{1+e_c}.$$

Hence, by (B.51),

$$\mathcal{F} + \sum_{j\in\mathcal{J}}\lambda_j\partial_{\lambda_j}\mathcal{F} = \mathcal{F} - \frac{1}{n}\operatorname{Tr}(\Lambda\mathcal{K}) = \sum_{c=1}^{C}\frac{\pi_c g_c}{(1+e_c)^2}.$$

Substituting this identity and (B.52) into the variance equivalent from Step 5 gives

$$\bar{\mathcal{V}} = \sigma_1^2\left(\mathcal{F} + \sum_{j\in\mathcal{J}}\lambda_j\partial_{\lambda_j}\mathcal{F}\right) - \sum_{c\in\mathcal{A}}w_c(w_c s_c^2 - \sigma_1^2)\partial_{w_c}\mathcal{F} = \sum_{c=1}^{C}\frac{\pi_c w_c s_c^2 g_c}{(1+e_c)^2} = \mathcal{V},$$

matching the formula in (4.7). Since the variance argument does not use assumption (A6), this proves the claimed variance limit of (4.6) in full generality. Together with (B.48)-(B.49) and, when applicable, (B.53), this completes the proof of Theorem 4.1. □

B.2 PROOF OF THE FULL BIAS TERM

We state an auxiliary result that allows us to concentrate the bias term for general fixed hidden signals. Recall from (3.7) that $B(\hat{\theta}, \theta) = \theta^\top \Lambda \hat{Q} S_1 \hat{Q} \Lambda \theta$. We will first rewrite the bias term in terms of matrices. Namely, let us set $a_\theta = \Lambda\theta$, and introduce

$$\hat{Q}_{S_1}(t) = (\hat{S} + \Lambda + tS_1)^{-1}, \qquad H_{S_1}(t) = a_\theta^\top \hat{Q}_{S_1}(t) a_\theta. \tag{B.54}$$

The vector a_θ is deterministic, uniformly bounded, and held fixed when t varies. Differentiating the inverse gives

$$\frac{d}{dt} H_{S_1}(t) = -a_\theta^\top \hat{Q}_{S_1}(t) S_1 \hat{Q}_{S_1}(t) a_\theta,$$

so the exact bias identity is

$$B = -\frac{d}{dt} H_{S_1}(t)\Big|_{t=0}.$$

Thus, we state and prove the following proposition which gives deterministic equivalents for $\frac{d}{dt} H_{S_1}(t)$.

Proposition B.1. *Under the design and parameter assumptions of Theorem 4.1, let $a \in \mathbb{R}^{Cd}$ and $A \succeq 0$ belong to deterministic sequences with $\sup_n \|a\|_2 < \infty$ and $\sup_n \|A\|_2 < \infty$. Let us choose an arbitrary $h_0 > 0$ and for $t \in [0, h_0]$, define empirical matrix $\hat{Q}_A(t) = (\hat{S} + \Lambda + tA)^{-1}$ and its corresponding deterministic equivalent*

$$\bar{Q}_A(t) = \left[\Lambda + tA + \sum_{c=1}^{C} \frac{\pi_c w_c S_c}{1 + e_{c,A}(t)} \right]^{-1}, \qquad e_{c,A}(t) = \frac{w_c}{n} \operatorname{Tr}(S_c \bar{Q}_A(t)). \tag{B.55}$$

Set $H_A(t) = a^\top \hat{Q}_A(t) a$ and $\mathcal{H}_A(t) = a^\top \bar{Q}_A(t) a$. Then for every fixed $t \in [0, h_0]$,

$$H_A(t) - \mathcal{H}_A(t) \xrightarrow{a.s.} 0, \tag{B.56}$$

and

$$\frac{d}{dt} H_A(t)\Big|_{t=0} - \frac{d}{dt} \mathcal{H}_A(t)\Big|_{t=0} \xrightarrow{a.s.} 0. \tag{B.57}$$

Proof. We extend the existing scalar argument to one deterministic quadratic form. In the proof of this proposition, where obvious, we omit the subscript A.

To begin with, we note that the matrix in (B.55) is indeed well-defined with unique positive solutions $e_{c,A} > 0$ as proved Step 2 and Step 3 of the argument. Namely, this follows since every positivity, existence, uniqueness, and comparison argument in Step 3 remains valid with Λ replaced by $\Lambda + tA$, because $tA \succeq 0$.

The main idea is to generalize the approach of using the resolvent theorem treating average direction (Wagner et al., 2012, Theorem 1) to using a more fine-grained result (Louart, 2026, Theorem 1.6) allowing per-direction control.

To begin with, we fix $t \geq 0$ and $\eta > 0$ and set

$$D_\eta = \Lambda + tA + \eta I, \qquad \hat{Q}_\eta(t) = (\hat{S} + D_\eta)^{-1}.$$

As before, we introduce η as it will allow us to make the argument technically simpler, with the aim to later obtain the results for $\eta = 0$. We can rewrite the $\hat{Q}_\eta(t)$ in terms of a resolvent function,

$$\hat{Q}_\eta(t) = D_\eta^{-1/2} R(-1) D_\eta^{-1/2}, \tag{B.58}$$

where $R(z) = (D_\eta^{-1/2} \hat{S} D_\eta^{-1/2} - zI)^{-1}$. We can view $D_\eta^{-1/2} \hat{S} D_\eta^{-1/2}$ as the empirical covariance matrix of the whitened n samples $\breve{x}_{c,i} \in \mathbb{R}^{Cd}$ defined as

$$\breve{x}_{c,i} = \sqrt{w_c}\, D_\eta^{-1/2} L_c^\top \Sigma_c^{1/2} Z_{c,i}^\top, \quad c \in \mathcal{C},\ i \in [n_c],$$

where $Z_{c,i}$ denotes the transpose of row i of Z_c. Moreover, we have that the population covariance matrix of each sample belonging to client c is

$$\breve{\Sigma}_c := \mathbb{E}[\breve{x}_{c,i} \breve{x}_{c,i}^\top] = w_c D_\eta^{-1/2} S_c D_\eta^{-1/2}.$$

Let us introduce the quantity $M_{\mathrm{F}}^{(q)}$ as

$$M_{\mathrm{F}}^{(q)} \coloneqq \sup_{\substack{c\in\mathcal{C}\\ i\in[n_c]\\ \|B\|_{\mathrm{F}}\le 1}} \left(\mathbb{E}\left|\breve{x}_{c,i}^{\top} B \breve{x}_{c,i} - \operatorname{Tr}(\breve{\Sigma}_c B)\right|^q\right)^{1/q},$$

where $\|\cdot\|_{\mathrm{F}}$ is the Frobenius norm, such that $\|B\|_{\mathrm{F}} = \sqrt{\operatorname{Tr}(BB^*)}$, where B^* is the complex conjugate of B. By Assumption (A3), it holds

$$\|\breve{\Sigma}_c\|_{\mathrm{F}} \le K_{c,\eta}\sqrt{d},$$

where the constant $K_{c,\eta}$ depends on the fixed η. By Assumption (A1), we have $\sqrt{d}\|\breve{\Sigma}_c\|_{\mathrm{F}} = O(n)$, and thus the assumption of Theorem 1.6 of (Louart, 2026) is satisfied. We then apply this theorem with $q = 4$, to get the deterministic equivalent of the resolvent function $R(z)$. Namely, for $v = D_\eta^{-1/2}a$, we have that, at every fixed complex $z \in \mathbb{C}\setminus[0,+\infty[$

$$\left(\mathbb{E}\left|v^{\top}(R(z)-\bar{R}(z))v\right|^4\right)^{\frac{1}{4}} \le K_7\|vv^{\top}\|_{\mathrm{F}}\left[\frac{M_{\mathrm{F}}^{(4)}}{\sqrt{n}} + (1+\max_c\|\breve{\Sigma}_c\|_2)\frac{\max_c\|\breve{\Sigma}_c\|_{\mathrm{F}} + (M_{\mathrm{F}}^{(2)})^2\sqrt{Cd}}{n}\right], \tag{B.59}$$

where K_7 is a constant independent of n and d, and $\bar{R}$ is the deterministic resolvent defined by equations

$$\bar{R}(z) = -\frac{1}{z}\left(\sum_{c=1}^{C} d_c(z)\pi_c\breve{\Sigma}_c + I\right)^{-1}, \qquad d_c(z) = -\frac{1}{z\left(1+\frac{1}{n}\operatorname{Tr}\left(\breve{\Sigma}_c\bar{R}(z)\right)\right)}. \tag{B.60}$$

Note that the objects in the above equations (B.60) are well-defined and have unique guaranteed solutions for every $z \in \mathbb{C}\setminus[0,+\infty[$, by (Louart, 2026, Theorem 1.1).

Let us bound the values on the LHS of (B.59). By previous calculations, we get

$$\max_c\|\breve{\Sigma}_c\|_{\mathrm{F}} \le K_\eta\sqrt{d},$$

for some constant K_η depending on η. Moreover, by Assumption (A3), it holds

$$\max_c\|\breve{\Sigma}_c\|_2 \le K_\eta\alpha_\Sigma.$$

We now give bounds for $M_{\mathrm{F}}^{(4)}$ and $M_{\mathrm{F}}^{(2)}$. For a deterministic matrix B, the fourth-moment inequality for quadratic forms of independent coordinates with bounded eighth moments (Bai & Silverstein, 2010, Lemma B.26) gives

$$\left(\mathbb{E}\left|\breve{x}_{c,i}^{\top} B \breve{x}_{c,i} - \operatorname{Tr}(\breve{\Sigma}_c B)\right|^4\right)^{1/4} \le K_\eta\|B\|_F, \tag{B.61}$$

which follows by writing $\breve{x}_{c,i} = \breve{U}_c Z_{c,i}^{\top}$ and the assumption (A2) on the bounded eighth moments of coordinates of Z_c. As the supremum is taken over $\|B\|_F \le 1$, this in turn directly implies that $M_{\mathrm{F}}^{(4)} \le K_\eta$. Similarly, by applying (Bai & Silverstein, 2010, Lemma B.26) for second moments, we can get $M_{\mathrm{F}}^{(2)} \le K_\eta$ using the fact that bounded eighth moments of coordinates of Z_c imply bounded fourth moments.

Plugging these bounds back into (B.59) we get

$$\left(\mathbb{E}\left|v^{\top}(R(z)-\bar{R}(z))v\right|^4\right)^{\frac{1}{4}} \le \frac{K_\eta}{\sqrt{n}}.$$

By applying Markov's inequality we have

$$\mathbb{P}\left(\left|v^{\top}(R(z)-\bar{R}(z))v\right| \ge t\right) \le \frac{K_\eta}{t^4n^2}.$$

Thus, by picking $t = n^{-1/20}$ and applying Borel-Cantelli, we get almost sure convergence to 0, as the right hand side is summable in n. Namely, we get for $z \in \mathbb{C} \setminus [0, +\infty[$

$$\left| v^\top (R(z) - \bar{R}(z)) v \right| \xrightarrow{\text{a.s.}} 0. \tag{B.62}$$

From the definition of v and the identity in (B.58), we have $v^\top R(-1) v = a^\top \hat{Q}_\eta(t) a$. Moreover, we get

$$D_\eta^{-1/2} \bar{R}(-1) D_\eta^{-1/2} = \bar{Q}_\eta(t), \qquad d_c(-1) = 1/(1 + e_{c,\eta}),$$

where $\bar{Q}_\eta(t)$ and $e_{c,\eta}$ are the solutions to (B.55), with $\Lambda + tA$ replaced by D_η. Thus, by taking $z = -1$ in (B.62) we get

$$a^\top (\hat{Q}_\eta(t) - \bar{Q}_\eta(t)) a \xrightarrow{\text{a.s.}} 0. \tag{B.63}$$

We move onto removing η from the previous identity. To this end, note that every positivity, existence, uniqueness, and comparison argument in Step 3 of the proof of Theorem 4.1 remains valid with Λ replaced by $\Lambda + tA$. This is because $tA \succeq 0$. In particular, the bounds (B.12)-(B.23) hold uniformly for $0 \le t \le h_0$. Thus, arguing as in the derivations of (B.20) and (B.23), we get

$$\sup_{0 \le t \le h_0} \|\hat{Q}_\eta(t) - \hat{Q}(t)\|_2 \le K\eta, \tag{B.64}$$

$$\sup_{0 \le t \le h_0} \max_c |e_{c,\eta}(t) - e_c(t)| \le K\eta. \tag{B.65}$$

On the other hand, subtracting the deterministic inverse matrices gives

$$\bar{Q}_\eta(t)^{-1} - \bar{Q}(t)^{-1} = \eta I + \sum_{c=1}^{C} \pi_c w_c S_c \left(\frac{1}{1 + e_{c,\eta}(t)} - \frac{1}{1 + e_c(t)} \right).$$

Thus, using (B.65), and uniform boundedness of π_c, w_c and S_c, it holds

$$\sup_{0 \le t \le h_0} \|\bar{Q}_\eta(t)^{-1} - \bar{Q}(t)^{-1}\|_2 \le K\eta.$$

Another inverse matrix identity and the bounds on $\|\bar{Q}_\eta(t)\|_2$ and $\|\bar{Q}(t)\|_2$, derived in the same way as (B.16) and (B.22), give

$$\sup_{0 \le t \le h_0} \|\bar{Q}_\eta(t) - \bar{Q}(t)\|_2 \le K\eta. \tag{B.66}$$

Then, it holds

$$\begin{aligned} |H(t) - \mathcal{H}(t)| &= \left| a^\top (\hat{Q}(t) - \bar{Q}(t)) a \right| \\ &\le \left| a^\top (\hat{Q}(t) - \hat{Q}_\eta(t)) a \right| + \left| a^\top (\hat{Q}_\eta(t) - \bar{Q}_\eta(t)) a \right| + \left| a^\top (\bar{Q}_\eta(t) - \bar{Q}(t)) a \right| \\ &\le K\eta \|a\|_2^2 + \left| a^\top (\hat{Q}_\eta(t) - \bar{Q}_\eta(t)) a \right|, \end{aligned}$$

where we have used (B.64) and (B.66) in the last inequality. We then apply (B.63), and take a countable sequence $\eta \downarrow 0$ as done in Step 3 of the previous section, thus proving (B.56) for every fixed t.

We next turn to the derivative convergence. We can rely on Step 4 of the previous section, specialized to the case of $\Lambda(t) = \Lambda + tA$. The goal is to give bounds on second derivatives of $H_A(t)$ and $\mathcal{H}_A(t)$, so that we can use Taylor's theorem to transfer the convergence to the derivatives.

Firstly, expanding the second derivative of the empirical term, we have

$$\frac{d^2}{dt^2} H_A(t) = 2a^\top \hat{Q}_A(t) A \hat{Q}_A(t) A \hat{Q}_A(t) a.$$

Using the bounds on $\hat{Q}_A(t)$ from previous section, we get eventually almost surely,

$$\sup_{0 \le t \le h_0} \left| \frac{d^2}{dt^2} H_A(t) \right| \le 2\|a\|_2^2 \|A\|_2^2 \kappa_{\mathrm{E}}^{-3} \le K_8.$$

We next aim to obtain a bound for the second derivative of $\mathcal{H}_A(t)$. To this end, we use the same map $\Phi_c(\mathbf{e}, t) = w_c \operatorname{Tr}(S_c Q(\mathbf{e}, t))/n$ as in (B.27) of Step 4 of the previous section, and all the

accompanying derivations. Note that, in the case we are considering, we can calculate the partial derivatives t explicitly as

$$\partial_t Q(\mathbf{e}, t)^{-1} = A, \qquad \partial_{tt} Q(\mathbf{e}, t)^{-1} = 0.$$

The Jacobian has exactly the form (B.32) and moreover by (B.33) it holds

$$(J(\mathbf{1} + \mathbf{e}))_c = e_c(t) - \frac{w_c}{n} \operatorname{Tr}\big(S_c \bar{Q}_A(t)(\Lambda + tA)\bar{Q}_A(t)\big) \leq e_c(t),$$

where $\mathbf{1}$ denotes the all-ones vector. Since A is bounded, the partial derivative bounds of Step 4 and the two implicit function identities (B.36) and (B.38) still apply. Namely, through (B.37) and (B.39), they imply

$$\sup_{0 \leq t \leq h_0} (\|\mathbf{e}'(t)\|_2 + \|\mathbf{e}''(t)\|_2) < K. \tag{B.67}$$

For completeness, let $M(t) = \bar{Q}_A(t)^{-1}$. Then

$$\begin{aligned}
M'(t) &= A - \sum_c \frac{\pi_c w_c e_c'(t)}{(1 + e_c(t))^2} S_c, \\
M''(t) &= -\sum_c \pi_c w_c \left[\frac{e_c''(t)}{(1 + e_c(t))^2} - \frac{2(e_c'(t))^2}{(1 + e_c(t))^3} \right] S_c, \\
\bar{Q}_A''(t) &= 2\bar{Q}_A(t)M'(t)\bar{Q}_A(t)M'(t)\bar{Q}_A(t) - \bar{Q}_A(t)M''(t)\bar{Q}_A(t).
\end{aligned}$$

By (B.67), all these matrices are uniformly bounded in operator norm. Therefore

$$\sup_{0 \leq t \leq h_0} \left| \frac{d^2}{dt^2} \mathcal{H}_A(t) \right| \leq \|a\|_2^2 \sup_{0 \leq t \leq h_0} \|\bar{Q}_A''(t)\|_2 \leq K_9. \tag{B.68}$$

Lastly, for every fixed $0 < h < h_0$, Taylor's theorem gives precisely the bound used in the original Step 4

$$\left| \frac{d}{dt} H_A(t)\Big|_{t=0} - \frac{d}{dt} \mathcal{H}_A(t)\Big|_{t=0} \right| \leq \frac{K_8 + K_9}{2} h + \frac{|H_A(h) - \mathcal{H}_A(h)| + |H_A(0) - \mathcal{H}_A(0)|}{h}.$$

Since h is fixed, the already established deterministic equivalent convergence of (B.56) gives

$$H(0) - \mathcal{H}(0) \xrightarrow{\text{a.s.}} 0, \qquad H(h) - \mathcal{H}(h) \xrightarrow{\text{a.s.}} 0.$$

Therefore, for fixed h, it holds almost surely

$$\limsup_{n \to \infty} \left| \frac{d}{dt} H_A(t)\Big|_{t=0} - \frac{d}{dt} \mathcal{H}_A(t)\Big|_{t=0} \right| \leq \frac{K_8 + K_9}{2} h.$$

Finally, let us take a deterministic sequence $h_m \downarrow 0$. Since this sequence is countable, we may intersect the probability-one events corresponding to all parameter values h_m. On this common event, the preceding inequality holds for every m. Sending $m \to \infty$ then proves (B.57). □

B.3 Concentration of the realized risk

We strengthen the convergence statement of Theorem 4.1. To this end, let us define

$$\hat{R}(\hat{\beta}; \beta) \coloneqq \|\hat{\theta} - \theta\|_{S_1}^2. \tag{B.69}$$

We will prove the slightly stronger statement that $\hat{R}(\hat{\beta}; \beta)$ almost surely concentrates around $B(\hat{\theta}; \theta) + V(\hat{\theta}; \theta)$, and thus around $\mathcal{B} + \mathcal{V}$.

By Assumption (A5), clipping is almost surely inactive for all sufficiently large n, and (A.8) therefore gives

$$\|\hat{\theta} - \theta\|_{S_1}^2 = \theta^\top \Lambda \hat{Q} S_1 \hat{Q} \Lambda \theta - 2\theta^\top \Lambda \hat{Q} S_1 \hat{Q} \hat{u} + \hat{u}^\top \hat{Q} S_1 \hat{Q} \hat{u}, \tag{B.70}$$

where the first term is precisely $B(\hat{\theta}; \theta)$. We next show that the second term vanishes almost surely and that the third concentrates around the variance expression in (3.7). Let

$$\tilde{X} \coloneqq \begin{bmatrix} X_1 L_1 \\ w_2 X_2 L_2 \\ \vdots \\ w_C X_C L_C \end{bmatrix}, \qquad \tilde{\varepsilon} \coloneqq \begin{bmatrix} \tilde{\varepsilon}_1 \\ \tilde{\varepsilon}_2 \\ \vdots \\ \tilde{\varepsilon}_C \end{bmatrix},$$

so that $\hat{u} = n^{-1}\tilde{X}^\top\tilde{\varepsilon}$. Note that under the spectral bounds established in the proof of Theorem 4.1,

$$\|\hat{Q}\|_2 = O(1), \qquad \|S_1\|_2 = O(1), \qquad \left\|\frac{1}{n}\tilde{X}^\top\tilde{X}\right\|_2 = O(1)$$

almost surely for all sufficiently large n. Moreover, by Assumption (A2) it holds that $\|\theta\|_2$ is uniformly bounded almost surely.

Consider now the first the cross term in (B.70) and define $a_n \coloneqq \frac{2}{n}\tilde{X}\hat{Q}S_1\hat{Q}\Lambda\theta$, so that

$$2\theta^\top\Lambda\hat{Q}S_1\hat{Q}\hat{u} = \tilde{\varepsilon}^\top a_n.$$

The preceding spectral bounds imply, for some constant K,

$$\|a_n\|_2^2 \le \frac{K}{n^2}\theta^\top\Lambda\hat{Q}S_1\hat{Q}\tilde{X}^\top\tilde{X}\hat{Q}S_1\hat{Q}\Lambda\theta = O(n^{-1})$$

almost surely. Since the entries of $\tilde{\varepsilon}$ are independent, centered, and have uniformly bounded 4-th moments, the standard fourth-moment bound for linear forms gives

$$\mathbb{E}\left[|\tilde{\varepsilon}^\top a_n|^4 \,\middle|\, X, \theta\right] \le K\|a_n\|_2^4 = O(n^{-2}).$$

Hence, by Markov's inequality and the Borel-Cantelli lemma,

$$\theta^\top\Lambda\hat{Q}S_1\hat{Q}\hat{u} \xrightarrow{\text{a.s.}} 0. \tag{B.71}$$

We now consider the quadratic noise term. Namely we express it as

$$\hat{u}^\top\hat{Q}S_1\hat{Q}\hat{u} = \tilde{\varepsilon}^\top M\tilde{\varepsilon},$$

for the matrix term $M \coloneqq \frac{1}{n^2}\tilde{X}\hat{Q}S_1\hat{Q}\tilde{X}^\top$. Moreover, let $\Omega \coloneqq \operatorname{diag}\left(\sigma_1^2 I_{n_1}, \sigma_{t,2}^2 I_{n_2}, \ldots, \sigma_{t,C}^2 I_{n_C}\right)$ denote the covariance of $\tilde{\varepsilon}$. Then, the deterministic centering term associated with this quadratic form is

$$\operatorname{Tr}(M\Omega) = V(\hat{\theta}; \theta).$$

Furthermore,

$$\|M\|_2 \le \frac{\|\tilde{X}\|_2^2}{n^2}\|\hat{Q}S_1\hat{Q}\|_2 = O(n^{-1}),$$

and, since $\operatorname{rank}(M) \le Cd = O(n)$,

$$\|M\|_F^2 \le \operatorname{rank}(M)\|M\|_2^2 = O(n^{-1}).$$

For independent, centered entries with uniformly bounded 8-th moments, the fourth-moment bound for quadratic forms yields

$$\mathbb{E}\left[\left|\tilde{\varepsilon}^\top M\tilde{\varepsilon} - \operatorname{Tr}(M\Omega)\right|^4 \,\middle|\, X, \theta\right] \le K\|M\|_F^4 = O(n^{-2}).$$

Another application of Markov's inequality and Borel-Cantelli therefore gives

$$\hat{u}^\top\hat{Q}S_1\hat{Q}\hat{u} - V(\hat{\theta}; \theta) \xrightarrow{\text{a.s.}} 0. \tag{B.72}$$

Combining (B.70), (B.71), and (B.72), we conclude that

$$\|\hat{\theta} - \theta\|_{S_1}^2 - \Big(B(\hat{\theta}; \theta) + V(\hat{\theta}; \theta)\Big) \xrightarrow{\text{a.s.}} 0. \tag{B.73}$$

Finally, using Theorem 4.1 we get

$$\left|\hat{R}(\hat{\beta}, \beta) - (\mathcal{B} + \mathcal{V})\right| \xrightarrow{\text{a.s.}} 0.$$

B.4 Proof of Proposition 4.3

We prove that the choice of weights and ridge parameters in (4.10) makes the estimator in (3.3) equal to the posterior mean, and that its target component minimizes the prediction risk. Throughout the proof, we condition on the design matrices $X_1, \ldots, X_C$, which are therefore treated as fixed.

Recall that, under label privatization and Gaussian observation noise assumption, the target observations satisfy

$$y_1 = X_1 L_1 \theta + \varepsilon_1, \qquad \varepsilon_1 \sim \mathcal{N}(0, \sigma_1^2 I_{n_1}),$$

while, for each $c \in \mathcal{A}$,

$$\tilde{y}_c = X_c L_c \theta + \tilde{\varepsilon}_c, \qquad \tilde{\varepsilon}_c \coloneqq \varepsilon_c + \xi_c \sim \mathcal{N}(0, \sigma_{t,c}^2 I_{n_c}),$$

For simplicity let us introduce $\sigma_{t,1} \coloneqq \sigma_1$ and $\tilde{y}_1 = y_1$. Thus, conditional on θ and the designs, we have

$$\tilde{y}_c \mid \theta, X_c \sim \mathcal{N}\left(X_c L_c \theta,\, \sigma_{t,c}^2 I_{n_c}\right).$$

Since the observation and privacy noises are independent across clients, the responses are conditionally independent given θ and the designs. Hence the likelihood factorizes as

$$\mathcal{L}(\theta) \coloneqq p\left(\tilde{y}_1, \ldots, \tilde{y}_C | \theta, X_1, \ldots, X_C\right) = \prod_{c \in \mathcal{C}} \frac{1}{(2\pi\sigma_{t,c}^2)^{n_c/2}} \exp\left(-\frac{1}{2\sigma_{t,c}^2} \|\tilde{y}_c - X_c L_c \theta\|_2^2\right). \tag{B.74}$$

Suppressing factors that do not depend on θ, we get

$$\mathcal{L}(\theta) \propto \exp\left(-\sum_{c \in \mathcal{C}} \frac{1}{2\sigma_{t,c}^2} \|\tilde{y}_c - X_c L_c \theta\|_2^2\right).$$

Thus, up to an additive constant independent of θ, the negative log-likelihood is

$$-\log \mathcal{L}(\theta) = \sum_{c \in \mathcal{C}} \frac{1}{2\sigma_{t,c}^2} \|\tilde{y}_c - X_c L_c \theta\|_2^2 + \text{const.} \tag{B.75}$$

On the other hand, by Assumption (4.9), the parameter θ has the Gaussian prior

$$\theta \sim \mathcal{N}(0, \Gamma), \text{ where } \quad \Gamma = \frac{r_\beta^2}{d} P_\beta + \sum_{c \in \mathcal{A}} \frac{r_{\delta_c}^2}{d} P_{\delta_c}.$$

Its precision matrix is therefore

$$\Gamma^{-1} = \frac{d}{r_\beta^2} P_\beta + \sum_{c \in \mathcal{A}} \frac{d}{r_{\delta_c}^2} P_{\delta_c}. \tag{B.76}$$

Let $u = \begin{bmatrix} b^\top & h_2^\top & \cdots & h_C^\top \end{bmatrix}^\top$ denote a generic value of the stacked parameter. By Bayes' rule, the posterior density of u, conditional on the designs and responses, is proportional to the product of its prior density and the likelihood. Thus, up to an additive constant independent of u, the negative log-posterior is

$$-\log p\left(u \mid X_1, \ldots, X_C, \tilde{y}_1, \ldots, \tilde{y}_C\right) = \sum_{c \in \mathcal{C}} \frac{1}{2\sigma_{t,c}^2} \|\tilde{y}_c - X_c L_c u\|_2^2 + \frac{1}{2} u^\top \Gamma^{-1} u + \text{const.} \tag{B.77}$$

Using (B.76), the prior term can equivalently be written as

$$\frac{1}{2} u^\top \Gamma^{-1} u = \frac{d}{2r_\beta^2} \|b\|_2^2 + \sum_{c \in \mathcal{A}} \frac{d}{2r_{\delta_c}^2} \|h_c\|_2^2.$$

We next identify the posterior distribution explicitly. Expanding the quadratic terms in (B.77), and discarding terms that do not depend on u, gives

$$-\log p(u \mid X_1, \ldots, X_C, \tilde{y}_1, \ldots, \tilde{y}_C) = \frac{1}{2} u^\top \Omega u - u^\top q + \text{const},$$

where

$$\Omega = \Gamma^{-1} + \sum_{c \in \mathcal{C}} \frac{1}{\sigma_{t,c}^2} L_c^\top X_c^\top X_c L_c, \tag{B.78}$$

$$q = \sum_{c \in \mathcal{C}} \frac{1}{\sigma_{t,c}^2} L_c^\top X_c^\top \tilde{y}_c. \tag{B.79}$$

Since Γ^{-1} is positive definite, Ω is also positive definite and hence invertible. Setting $m \coloneqq \Omega^{-1} q$, we complete the square:

$$\frac{1}{2} u^\top \Omega u - u^\top q = \frac{1}{2}(u-m)^\top \Omega (u-m) - \frac{1}{2} q^\top \Omega^{-1} q.$$

Thus it holds that

$$p\left(u \mid X_1, \ldots, X_C, \tilde{y}_1, \ldots, \tilde{y}_C\right) \propto \exp\left(-\frac{1}{2}(u-m)^\top \Omega (u-m)\right),$$

from which it directly follows that

$$\theta \mid X_1, \ldots, X_C, \tilde{y}_1, \ldots, \tilde{y}_C \sim \mathcal{N}(m, \Omega^{-1}),$$

and therefore

$$\mathbb{E}\left[\theta | X_1, \ldots, X_C, \tilde{y}_1, \ldots, \tilde{y}_C\right] = m = \Omega^{-1} q. \tag{B.80}$$

We now show that this posterior mean is precisely the estimator (3.3) evaluated at the parameters in (4.10). Recall that

$$\hat{S}_c = \frac{1}{n} L_c^\top X_c^\top X_c L_c, \qquad \hat{U}_c = \frac{1}{n} L_c^\top X_c^\top \tilde{y}_c.$$

Let us introduce the ridge matrix as

$$\Lambda^\circ \coloneqq \lambda_\beta^\circ P_\beta + \sum_{c \in \mathcal{A}} \lambda_{\delta_c}^\circ P_{\delta_c},$$

with λ_β°, $\lambda_{\delta_c}^\circ$ defined in (4.10). From the definition of the ridge parameters, it holds

$$\frac{n}{\sigma_1^2} \Lambda^\circ = \frac{d}{r_\beta^2} P_\beta + \sum_{c \in \mathcal{A}} \frac{d}{r_{\delta_c}^2} P_{\delta_c} = \Gamma^{-1}. \tag{B.81}$$

Similarly, using w_c° defined in (4.10), we get

$$\frac{n}{\sigma_1^2}\left(\hat{S}_1 + \sum_{c \in \mathcal{A}} w_c^\circ \hat{S}_c\right) = \frac{1}{\sigma_1^2} L_1^\top X_1^\top X_1 L_1 + \sum_{c \in \mathcal{A}} \frac{1}{\sigma_{t,c}^2} L_c^\top X_c^\top X_c L_c.$$

Combining this identity with (B.78) and (B.81) yields

$$\Omega = \frac{n}{\sigma_1^2}\left(\hat{S}_1 + \sum_{c \in \mathcal{A}} w_c^\circ \hat{S}_c + \Lambda^\circ\right). \tag{B.82}$$

Likewise, from (B.79),

$$q = \frac{n}{\sigma_1^2}\left(\hat{U}_1 + \sum_{c \in \mathcal{A}} w_c^\circ \hat{U}_c\right). \tag{B.83}$$

Substituting (B.82) and (B.83) into (B.80) gives

$$\mathbb{E}\left[\theta \mid X_1, \ldots, X_C, y_1, \tilde{y}_2, \ldots, \tilde{y}_C\right] = \left(\hat{S}_1 + \sum_{c \in \mathcal{A}} w_c^\circ \hat{S}_c + \Lambda^\circ\right)^{-1} \left(\hat{U}_1 + \sum_{c \in \mathcal{A}} w_c^\circ \hat{U}_c\right).$$

The right-hand side is exactly the closed-form expression for the estimator in (3.3), as expressed in (3.5) with parameters $(\mathbf{w}^\circ, \lambda_\beta^\circ, \boldsymbol{\lambda}_\delta^\circ)$. Hence if we denote this estimator by $\hat{\theta}^\circ$, we have

$$\hat{\theta}^\circ = \mathbb{E}\left[\theta \mid X_1, \ldots, X_C, \tilde{y}_1, \ldots, \tilde{y}_C\right].$$

Taking the first d-dimensional block gives

$$\hat{\beta}^{\circ} = \mathbb{E}\left[\beta \mid X_1, \ldots, X_C, \tilde{y}_1, \ldots, \tilde{y}_C\right].$$

It remains to prove that $\hat{\beta}^{\circ}$ minimizes $\mathbb{E}_{\theta}[R(\hat{\beta}; \beta)]$. Let us introduce the sigma algebra $\mathcal{D} \coloneqq \sigma\left(X_1, \ldots, X_C, \tilde{y}_1, \ldots, \tilde{y}_C\right)$, generated by the observed data, and let

$$m_{\beta} \coloneqq \mathbb{E}[\beta \mid \mathcal{D}] = \hat{\beta}^{\circ}.$$

Consider any other $\mathcal{D}$-measurable estimator $\tilde{\beta}$. We have the decomposition

$$\tilde{\beta} - \beta = (\tilde{\beta} - m_{\beta}) + (m_{\beta} - \beta).$$

Therefore,

$$\|\tilde{\beta} - \beta\|_{\Sigma_1}^2 = \|\tilde{\beta} - m_{\beta}\|_{\Sigma_1}^2 + \|m_{\beta} - \beta\|_{\Sigma_1}^2 + 2(\tilde{\beta} - m_{\beta})^{\top}\Sigma_1(m_{\beta} - \beta). \tag{B.84}$$

Taking the conditional expectation given $\mathcal{D}$, the first term is unchanged because both $\tilde{\beta}$ and m_{β} are $\mathcal{D}$-measurable. For the cross term,

$$\mathbb{E}[m_{\beta} - \beta \mid \mathcal{D}] = m_{\beta} - \mathbb{E}[\beta \mid \mathcal{D}] = 0.$$

Hence (B.84) gives

$$\mathbb{E}\left[\|\tilde{\beta} - \beta\|_{\Sigma_1}^2 \mid \mathcal{D}\right] = \|\tilde{\beta} - \hat{\beta}^{\circ}\|_{\Sigma_1}^2 + \mathbb{E}\left[\|\hat{\beta}^{\circ} - \beta\|_{\Sigma_1}^2 \mid \mathcal{D}\right]. \tag{B.85}$$

Since $\Sigma_1 \succ 0$, the first term on the right-hand side is nonnegative, and it is equal to zero if and only if $\tilde{\beta} = \hat{\beta}^{\circ}$. Therefore, for every realization of the observed data,

$$\hat{\beta}^{\circ} = \mathbb{E}[\beta \mid \mathcal{D}]$$

uniquely minimizes the posterior expected target prediction loss.

Finally, recall that

$$\mathbb{E}_{\theta}[R(\tilde{\beta}; \beta)] = \mathbb{E}\left[\|\tilde{\beta} - \beta\|_{\Sigma_1}^2 \mid X_1, \ldots, X_C\right].$$

Applying the tower property to (B.85) gives

$$\mathbb{E}_{\theta}[R(\tilde{\beta}; \beta)] = \mathbb{E}_{\theta}[R(\hat{\beta}^{\circ}; \beta)] + \mathbb{E}\left[\|\tilde{\beta} - \hat{\beta}^{\circ}\|_{\Sigma_1}^2 \mid X_1, \ldots, X_C\right].$$

The second term is nonnegative, which proves that $\hat{\beta}^{\circ}$ is the desired minimizer and concludes the argument. □

C Experimental details and additional numerical results

C.1 Experimental details

Dataset preprocessing. For *Student Performance*, we use 649 observations from the two schools and 38 encoded features after removing the intermediate grades G1 and G2 and constant columns. For *LSAC*, removing observations with missing outcomes, client identifiers, or selected predictors leaves 20,970 observations across 6 law-school tiers and 10 encoded features. For *Lianjia*, we use 300,284 observations across 9 Beijing districts and the 6 selected features. Missing or invalid numerical values are imputed with medians.

Estimation of distributional quantities. The deterministic equivalents (4.7) and (4.8) depend on the unknown distributional quantities $\{\Sigma_c, \sigma_c\}_{c=1}^{C}$, together with either the relevant inner products between $\{\beta_c\}_{c=1}^{C}$ or the magnitudes τ_{β} and $\{\tau_{\delta_c}\}_{c=2}^{C}$, respectively. In both experiments of Figure 2 and Figure 3, using designated non-private data, we estimate the relevant regression coefficients and noise variances by five-fold cross-fitting. To stabilize covariance estimation, we apply 10% isotropic shrinkage to the covariance estimate, i.e. we mix the regular empirical covariance estimate with an equal variance isotropic covariance matrix, outputting $0.9\hat{\Sigma}_c + 0.1(\mathrm{Tr}(\hat{\Sigma}_c)/d)I$, where $\hat{\Sigma}_c$ is the standard empirical covariance matrix as in (3.4). The two experiments differ, however, in which

non-private observations are available for estimating these quantities. For the experiment on the prediction of the risk change in Figure 2, our goal is to test how accurately the deterministic equivalent predicts empirical changes in target risk with data from real datasets, when its distributional quantities are estimated as accurately as possible. We therefore estimate the distributional quantities in each repetition using all de-privatized training and validation observations available for each client. We then evaluate the simplified deterministic equivalent (4.8), which averages over signal and model shift directions. On the other hand, for the dataset selection experiment in Figure 3, we instead simulate a setting in which expected dataset benefit must be predicted before accessing the observations used for model fitting or evaluation. We therefore split each of the datasets into two roughly same-sized partitions, one designated for estimating distributional quantities, and the other for the actual experiment. The distributional quantities are thus estimated once from this fixed non-private reference partition that is disjoint from all training, validation, and test observations. These estimates are held fixed across different privacy noise levels and experimental repetitions. Client selection uses the general, direction-aware deterministic equivalent (4.7). In both experiments, the used training sample sizes and the auxiliary privacy noise variances of all clients are assumed to be known when evaluating the deterministic equivalent.

Hyperparameter tuning. The experiments in Figures 2 and 3 also use different hyperparameter-selection procedures. For the experiment on the prediction of the risk change in Figure 2, in order to select hyperparameters, we first do a pilot experiment with 100 repetitions. Each pilot repetition uses a randomized target training-validation split and all available auxiliary observations, with the corresponding privacy noise added to the auxiliary responses. For each client pair and privacy level, we jointly select the ridge penalties and auxiliary weight that minimize the average target-validation MSE across these pilot repetitions. The selected parameters are then fixed across all auxiliary sample sizes and all main experiment 100 evaluation repetitions, with identical values used by the empirical estimator and the deterministic-equivalent prediction. This allows us to isolate how accurately the deterministic equivalent predicts empirical changes in target risk with data from real datasets. On the other hand, for the dataset selection experiment in Figure 3, the empirical ridge penalties and auxiliary weights are chosen using target-validation MSE in each evaluation repetition, privacy condition and candidate client subset. Hyperparameter tuning uses that repetition's sampled training observations, including privatized auxiliary responses. All selection methods share the same hyperparameters for a given client subset and repetition. To rank candidate clients, the deterministic equivalent used by ***Ours*** instead uses the population-based hyperparameter formulas from Proposition 4.3, evaluated using estimates from the independent non-private reference partition. These selection hyperparameters remain fixed across repetitions for each privacy level and may differ from those used by the empirically evaluated estimators. This reflects a setting in which the expected benefit of auxiliary data must be predicted before accessing the observations used for model fitting or evaluation.

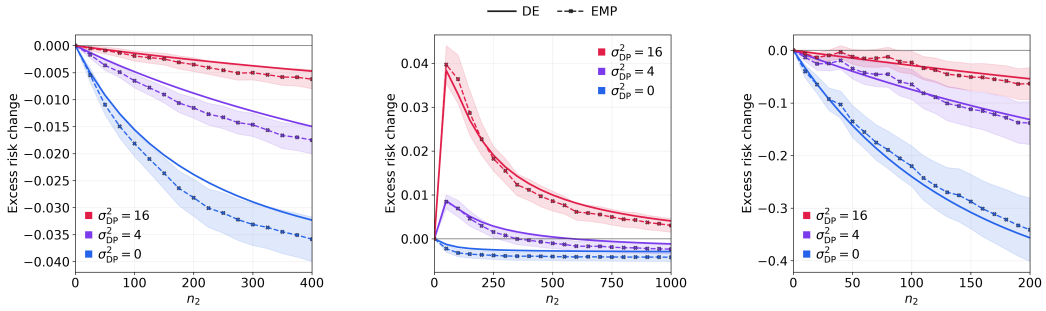
C.2 Semi-synthetic experiment

To isolate the accuracy of the deterministic equivalent from the error in estimating population-level quantities, we construct a semi-synthetic version of the experiment in Figure 2. We use the original covariates and client definitions, but replace the responses with data generated from planted linear models. Namely, using all data, we first estimate all hidden signals β_c and noise variances σ_c^2 by five-fold cross-fitting. We then use these estimates, denoted by β_c^* and σ_c^{*2}, to generate responses

$$y_c^* = X_c\beta_c^* + \varepsilon_c^*, \qquad \varepsilon_c^* \sim \mathcal{N}(0, \sigma_c^{*2} I).$$

The covariates and planted signals remain fixed, while observation noise, data partitions, and privacy noise are resampled in each repetition.

As in the original experiment from Figure 2, we fix the target training size at $n_1 = 100$ and vary the auxiliary sample size up to 200, 400, and 1000 for Student Performance, LSAC, and Lianjia, respectively. We then add to the auxiliary responses independent Gaussian noise with variance $\sigma_{\mathrm{DP}}^2 \in \{0, 4, 16\}$. We then repeat the hyperparameter tuning setup of the experiment in Figure 2. Namely, we run a pilot experiment in which the ridge penalties and auxiliary weights are jointly selected by minimizing the average target-validation MSE over 100 repetitions using all available auxiliary observations. The selected hyperparameters are then fixed across auxiliary sample sizes



(a) LSAC **(b)** Lianjia **(c)** Student Performance

Figure 4: Change in target excess risk as auxiliary data are added in two-client semi-synthetic regression experiments. Lower values correspond to a better risk. Original covariates are retained, while responses are generated from planted linear models calibrated to each client. **(a) LSAC:** Tier 3 as the target client, and Tier 4 as the auxiliary client. **(b) Lianjia:** District 8 as the target client, and District 7 as the auxiliary client. **(c) Student Performance:** Mousinho da Silveira (MS) as the target school, and Gabriel Pereira (GP) as the auxiliary school. Solid lines show direction-aware deterministic equivalent predictions (DE) using the known generating parameters; dashed lines with crosses show empirical results (EMP).

and the 100 evaluation repetitions, and are used identically by the empirical estimator and the deterministic equivalent.

The deterministic equivalent is evaluated using the known planted signals, noise variances, and empirical covariance computed with 10% isotropic shrinkage, as in the original experiment. As we have exact access to planted signals, we use the direction-aware deterministic equivalent expression of (4.7). Figure 4 reports the change in target risk relative to the target-only estimator. Dashed lines with crosses show empirical means and pointwise 90% confidence intervals, while solid lines show the deterministic-equivalent predictions.

Removing the error in estimating population-level quantities and imposing a correctly specified linear response model improves the accuracy of theoretical predictions for the datasets Student Performance and Lianjia, as can be seen by comparing plots **(b)-(c)** in Figure 2 and Figure 4. For the LSAC dataset, the discrepancy between empirical behaviour and theoretical predictions changes less between the two figures, as it is driven primarily by the variance term, rather than by estimation error or model misspecification. This is consistent with LSAC covariates having large fourth moments, which cause stronger finite sample fluctuations in the empirical covariance matrices than are captured by the asymptotic approximation via the deterministic equivalent.

C.3 Dataset selection varying privacy levels

In Tables 1-3, we present results of dataset selection experiments for different privacy noise levels. As before, the values shown are changes in the target excess risk, expressed as a percentage of each repetition's target-only test risk, for varying auxiliary privacy noise variance σ^2_{DP}. Each entry is the mean $\pm$ 90% confidence interval computed over 100 repetitions, in percentage points. Lower is better. Bold marks the lowest mean in each column and dataset.

Dataset	Method	1 Client	2 Clients	3 Clients
LSAC	Ours	$\mathbf{-2.41 \pm 0.32}$	$\mathbf{-3.05 \pm 0.33}$	$\mathbf{-3.05 \pm 0.33}$
	Random	-1.35 ± 0.18	-2.20 ± 0.26	-2.75 ± 0.29
	Largest sample	-1.32 ± 0.26	-2.81 ± 0.32	-2.88 ± 0.32
	Best local risk	-0.94 ± 0.20	-1.93 ± 0.25	-2.72 ± 0.30
	Lower bound (oracle)	-2.94 ± 0.31	-3.39 ± 0.32	-3.52 ± 0.33
Lianjia	Ours	$\mathbf{-2.15 \pm 0.70}$	$\mathbf{-3.97 \pm 0.83}$	$\mathbf{-4.40 \pm 0.85}$
	Random	-1.69 ± 0.46	-2.64 ± 0.64	-3.29 ± 0.74
	Largest sample	-1.70 ± 0.57	-2.85 ± 0.76	-3.11 ± 0.77
	Best local risk	-1.65 ± 0.54	-2.51 ± 0.66	-3.17 ± 0.72
	Lower bound (oracle)	-3.42 ± 0.69	-4.22 ± 0.83	-4.51 ± 0.84

Table 1: Change in target excess risk for auxiliary privacy noise variance $\sigma^2_{\mathrm{DP}} = 0$.

Dataset	Method	1 Client	2 Clients	3 Clients
LSAC	Ours	$\mathbf{-1.02 \pm 0.20}$	$\mathbf{-1.93 \pm 0.29}$	$\mathbf{-2.31 \pm 0.29}$
	Random	-0.78 ± 0.12	-1.42 ± 0.18	-1.94 ± 0.23
	Largest sample	-0.69 ± 0.17	-1.45 ± 0.22	-1.66 ± 0.23
	Best local risk	-0.69 ± 0.18	-1.50 ± 0.26	-2.10 ± 0.26
	Lower bound (oracle)	-2.12 ± 0.24	-2.59 ± 0.28	-2.77 ± 0.30
Lianjia	Ours	$\mathbf{-1.87 \pm 0.55}$	$\mathbf{-3.14 \pm 0.77}$	$\mathbf{-3.67 \pm 0.78}$
	Random	-1.37 ± 0.37	-2.28 ± 0.56	-2.92 ± 0.67
	Largest sample	-1.52 ± 0.53	-2.69 ± 0.75	-3.01 ± 0.75
	Best local risk	-1.63 ± 0.57	-2.48 ± 0.67	-3.06 ± 0.72
	Lower bound (oracle)	-3.19 ± 0.67	-3.99 ± 0.77	-4.21 ± 0.81

Table 2: Change in target excess risk for auxiliary privacy noise variance $\sigma^2_{\mathrm{DP}} = 4$.

Dataset	Method	1 Client	2 Clients	3 Clients
LSAC	Ours	-0.60 ± 0.16	$\mathbf{-1.49 \pm 0.27}$	-1.90 ± 0.28
	Random	-0.62 ± 0.10	-1.17 ± 0.17	-1.66 ± 0.22
	Largest sample	-0.52 ± 0.16	-1.05 ± 0.20	-1.32 ± 0.21
	Best local risk	$\mathbf{-0.66 \pm 0.19}$	-1.28 ± 0.25	$\mathbf{-1.93 \pm 0.28}$
	Lower bound (oracle)	-1.87 ± 0.24	-2.30 ± 0.27	-2.48 ± 0.29
Lianjia	Ours	$\mathbf{-1.40 \pm 0.54}$	-2.43 ± 0.67	$\mathbf{-3.17 \pm 0.68}$
	Random	-1.06 ± 0.28	-1.92 ± 0.47	-2.57 ± 0.60
	Largest sample	-1.28 ± 0.46	-2.27 ± 0.66	-2.74 ± 0.70
	Best local risk	-1.36 ± 0.45	$\mathbf{-2.46 \pm 0.70}$	-2.97 ± 0.75
	Lower bound (oracle)	-3.09 ± 0.67	-3.83 ± 0.77	-4.08 ± 0.80

Table 3: Change in target excess risk for auxiliary privacy noise variance $\sigma^2_{\mathrm{DP}} = 16$.