Subsample Ridge Ensembles: Equivalences and Generalized Cross-Validation

Jin-Hong Du [*1] **Pratik Patil** [*2] **Arun Kumar Kuchibhotla** [1]

Abstract

We study subsampling-based ridge ensembles in the proportional asymptotics regime, where the feature size grows proportionally with the sample size such that their ratio converges to a constant. By analyzing the squared prediction risk of ridge ensembles as a function of the explicit penalty λ and the limiting subsample aspect ratio ϕ_s (the ratio of the feature size to the subsample size), we characterize contours in the (λ, ϕ_s)-plane at any achievable risk. As a consequence, we prove that the risk of the optimal full ridgeless ensemble (fitted on all possible subsamples) matches that of the optimal ridge predictor. In addition, we prove strong uniform consistency of generalized cross-validation (GCV) over the subsample sizes for estimating the prediction risk of ridge ensembles. This allows for GCV-based tuning of full ridgeless ensembles without sample splitting and yields a predictor whose risk matches optimal ridge risk.

1. Introduction

Ensemble methods (Breiman, 1996) are widely used in various real-world applications in statistics and machine learning. They combine a collection of weak predictors to produce more stable and accurate predictions. One notable example of an ensemble method is bagging (**b**ootstrap **agg**regat**ing**) (Breiman, 1996; Bühlmann & Yu, 2002). Bagging involves averaging base predictors that are fitted on different subsampled datasets and has been shown to stabilize the prediction and reduce the predictive variance (Bühlmann & Yu, 2002). In this paper, we study such a class of ensemble methods that fit each base predictor independently using a different subsampled dataset of the full training data. As a prototypical base predictor, we focus on *ridge regression* (Hoerl & Kennard, 1970a;b), one of the most popular statistical methods. We refer readers to the "ridgefest" by Hastie (2020) for the history and review of ridge regression.

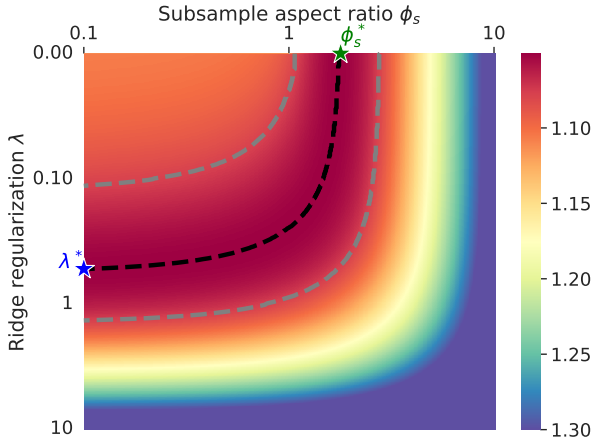

Figure 1. Heat map of the asymptotic prediction risk landscape of full ridge ensembles as the number of observation n, the subsample size k, and the feature dimension p tend to infinity, for varying regularization parameters λ and limiting subsample aspect ratio $\phi_s = \lim p/k$. The data $(\boldsymbol{x}, y) \in \mathbb{R}^p \times \mathbb{R}$ is generated from a non-isotropic linear model $y = \boldsymbol{x}^\top \boldsymbol{\beta}_0 + \epsilon$ with $\phi = \lim p/n = 0.1$, where the features, the coefficients, and the residuals are distributed as $\boldsymbol{x} \sim \mathcal{N}(0, \boldsymbol{\Sigma}_{\mathrm{AR1}})$, $\boldsymbol{\beta}_0 = \frac{1}{5}\sum_{j=1}^{5} \boldsymbol{w}_{(j)}$, and $\epsilon \sim \mathcal{N}(0, 1)$, respectively. Here, the covariance matrix $(\boldsymbol{\Sigma}_{\mathrm{AR1}})_{ij} = 0.5^{|i-j|}$, $\boldsymbol{w}_{(j)}$ is the top jth eigenvector of $\boldsymbol{\Sigma}_{\mathrm{AR1}}$. The green and blue stars denote the risk of the optimal full-ensemble ridgeless predictor and the optimal ridge predictor without subsampling, respectively. The black dashed line denotes the set of (λ, ϕ_s) pairs that yield the same risk as (λ^*, ϕ) and $(0, \phi_s^*)$, while the gray dashed lines indicate the set of pairs that all result in the same sub-optimal risk.

Ridge regression has recently attracted great interest, particularly the limiting case of zero regularization (where the regularization parameter tends to zero), termed "ridgeless" regression. In the underparameterized regime, the ridgeless predictor is ordinary least squares. However, in the overparameterized regime, it interpolates the training data and exhibits a peculiar risk behavior (Belkin et al., 2020; Bartlett et al., 2020; Hastie et al., 2022; Muthukumar et al., 2020). LeJeune et al. (2020); Patil et al. (2022a) have recently analyzed the statistical properties of the ensemble ridge and ridgeless predictors under proportional asymp-

[*]Equal contribution [1]Department of Statistics and Data Science, Carnegie Mellon University, Pittsburgh, PA 15213, USA. [2]Department of Statistics, University of California, Berkeley, CA 94720, USA. Correspondence to: Jin-Hong <jinhongd@andrew.cmu.edu>, Pratik <pratikpatil@berkeley.edu>.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023.

totics. Under a linear model with the isotropic Gaussian covariate distribution, LeJeune et al. (2020) prove that the *full ensemble* (ensemble fitted on all possible subsampled datasets) of least squares predictors with optimal subsample size has the same risk as that of ridge predictor with optimal regularization. Under a more general but still isotropic covariate distribution, Patil et al. (2022a) prove similar risk equivalence of the optimized full ridgeless ensemble and the optimized ridge predictor.

These findings inspire two natural avenues to investigate.

(1) Understanding the extent of risk equivalences. As a curious experiment, one can empirically observe that a similar phenomenon to the one just mentioned appears to hold under quite general non-isotopic data models, as illustrated in Figure 1. We observe that the optimal ridgeless in the full ensemble (the green star) has the same prediction risk as the optimal ridge on the full data (the blue star). Furthermore, any pair of (λ, ϕ_s) on the black line achieves the same optimal risk. Such a relationship also extends to any other attainable risk value. For example, see the grey lines for (λ, ϕ_s) pairs that all achieve the same sub-optimal risk. This inspires our first investigation to establish risk equivalences between subsampling and ridge regression under general settings.

(2) Overcoming limitations of split cross-validation. Apart from its theoretical interest, the risk equivalences also suggest an alternative practical way to tune the ridge regularization parameter by tuning the subsample size. In terms of practical tuning of the ridge and ridgeless ensembles, Patil et al. (2022a) provide a split cross-validation method to estimate the prediction risk of ensembles with a fixed (finite) number of ensemble sizes and further prove that the split cross-validation consistently selects the best subsample size. The split cross-validation procedure has two disadvantages: (a) sample splitting introduces additional external randomness in the predictor; and (b) the reduced sample size, although asymptotically negligible, has significant finite sample effects, especially near the interpolation thresholds. This inspires our second investigation to address these limitations by considering generalized cross-validation (GCV) that does not require any sample splitting. The consideration of GCV as a viable estimator of the prediction risk for ridge ensembles stems from the observation that the ridge ensembles are also in fact linear smoothers.

1.1. Summary of Contributions

Below we provide a brief overview of our main results.

- **General risk equivalences.** We establish general equivalences between the subsample-optimized ridgeless ensemble, the optimal ridge predictor, and the optimal subsample ridge ensemble (see Theorem 2.3). In addition, for any $\tau \geq 0$, we provide an exact characterization of the sets $\mathcal{C}_\tau$ of pairs (λ, ϕ_s) (the regularization parameter and the limiting subsample aspect ratio) such that the risk of the full ridge ensemble with ridge regularization λ and subsample aspect ratio ϕ_s is equal to the risk of the ridge predictor with ridge regularization τ. In essence, this amounts to showing that the implicit regularization of subsampling is the same as additional explicit ridge regularization.

- **Uniform consistency of GCV.** We establish the uniform consistency of GCV across all possible subsample sizes for full ridge ensembles with fixed regularization parameters (see Theorem 3.1). Notably, this result is also applicable to zero explicit regularization and covers the case of ridgeless regression. This finding enables tuning over the subsample size in a data-dependent manner, and in conjunction with Theorem 2.3, it implies that GCV tuning leads to a predictor with the same risk as the optimal ridge predictor (see Corollary 3.2).

- **Finite-ensemble surprises.** Even though GCV is consistent for the non-ensemble ridge and full-ensemble ridge predictors, interestingly, this is the first paper that proves GCV *can* be inconsistent even for ridge ensembles when the ensemble size is two (see Proposition 3.3). This finding is in contrast to other known results of GCV for ridge (see Section 1.2 for more details). Nevertheless, experiments on synthetic data and real-world single-cell multiomic datasets demonstrate the applicability of GCV for tuning subsample sizes, even with moderate ensemble sizes (roughly of order 10).

1.2. Related Work

Ensembles and risk analysis. Ensemble methods are effective in combining weak predictors to build strong predictors in both regression and classification settings (Hastie et al., 2009). Early work on ensemble methods includes classical papers by Breiman (1996); Bühlmann & Yu (2002). There has been further work on the ensembles of smooth weak predictors (Buja & Stuetzle, 2006; Friedman & Hall, 2007), non-parametric estimators (Bühlmann & Yu, 2002; Loureiro et al., 2022), and classifiers (Hall & Samworth, 2005; Samworth, 2012). Under proportional asymptotics, d'Ascoli et al. (2020); Adlam & Pennington (2020a); Loureiro et al. (2022) study ensemble learning under random feature models. For ridge ensembles, Sollich & Krogh (1995); Krogh & Sollich (1997) derive risk asymptotics under Gaussian features. LeJeune et al. (2020) consider least squares ensembles obtained by subsampling such that the final subsampled dataset has more observations than the number of features. The asymptotic risk characterization for general data models has been derived by Patil et al. (2022a). Both of these works show the equivalence between the subsample optimized full ridgeless ensemble and the optimal ridge under isotropic

models. Our work significantly extends the scope of these results by characterizing risk equivalences for both optimal and suboptimal risks and for arbitrary feature covariance and signal structures. See the remarks after Theorem 2.3 for a detailed comparison,

Cross-validation and consistency. Cross-validation (CV) is arguably the most popular class of methods for model assessment and selection. Classical work on CV include: Allen (1974); Stone (1974; 1977); Geisser (1975), among others. We refer the reader to Arlot & Celisse (2010); Zhang & Yang (2015) for comprehensive surveys of different CV variants. In practice, k-fold CV is widely used with typical k being 5 or 10 (Hastie et al., 2009; Györfi et al., 2006), but such small values of k suffer from bias in high dimensions (Rad & Maleki, 2020). The extreme case of leave-one-out cross-validation (LOOCV) (when $k = n$) alleviates the bias issues in risk estimation, and various statistical consistency properties of LOOCV have been analyzed in recent years; see, e.g., Kale et al. (2011); Kumar et al. (2013); Obuchi & Kabashima (2016); Rad et al. (2020). Except for special cases, LOOCV is computationally expensive, and consequently, various approximations and their theoretical properties have been studied; see, e.g, Wang et al. (2018); Rad & Maleki (2020); Rad et al. (2020); Xu et al. (2019). Generalized cross-validation (GCV) is a sort of approximation for the "shortcut" leave-one-out formula (Hastie et al., 2009), originally studied for the fixed-X design setting for linear smoothers by Golub et al. (1979); Craven & Wahba (1979). The consistency of GCV in such a setting has been investigated in Li (1985; 1986; 1987). More recently, in the random-X setting, GCV has received considerable attention. In particular, consistency of GCV for ridge regression has been established in Adlam & Pennington (2020b); Hastie (2020); Patil et al. (2021; 2022c); Wei et al. (2022) under various data settings. Our work contributes to this body of work by analyzing GCV for subsampled ensemble ridge regression.

2. Subsample and Ridge Equivalences

We consider the standard supervised regression setting. Let $\mathcal{D}_n = \{(\boldsymbol{x}_j, y_j) : j \in [n]\}$ denote a dataset containing i.i.d. random vectors in $\mathbb{R}^p \times \mathbb{R}$, $\boldsymbol{X} \in \mathbb{R}^{n\times p}$ denote the feature matrix whose j-th row contains $\boldsymbol{x}_j^\top$, and $\boldsymbol{y} \in \mathbb{R}^n$ denote the response vector whose j-th entry contains y_j. For an index set $I \subseteq [n]$ of size k, let $\mathcal{D}_I = \{(\boldsymbol{x}_j, y_j) : j \in I\}$ be a subsampled dataset and let $\boldsymbol{L}_I \in \mathbb{R}^{n\times n}$ denote a diagonal matrix such that its jth diagonal entry is 1 if $j \in I$ and 0 otherwise. Noting that the feature matrix and response vector associated with $\mathcal{D}_I$ are $\boldsymbol{L}_I\boldsymbol{X}$ and $\boldsymbol{L}_I\boldsymbol{y}$, respectively, the *ridge* estimator $\widehat{\boldsymbol{\beta}}_k^\lambda(\mathcal{D}_I)$ fitted on $\mathcal{D}_I$ (containing k samples) with regularization parameter $\lambda > 0$ can be expressed as:

$$\widehat{\boldsymbol{\beta}}_k^\lambda(\mathcal{D}_I) = \operatorname*{argmin}_{\boldsymbol{\beta}\in\mathbb{R}^p} \sum_{j\in I}(y_j - \boldsymbol{x}_j^\top\boldsymbol{\beta})^2/k + \lambda\|\boldsymbol{\beta}\|_2^2$$
$$= (\boldsymbol{X}^\top\boldsymbol{L}_I\boldsymbol{X}/k + \lambda\boldsymbol{I}_p)^{-1}\boldsymbol{X}^\top\boldsymbol{L}_I\boldsymbol{y}/k. \quad (1)$$

Letting $\lambda \to 0^+$, $\widehat{\boldsymbol{\beta}}_k^0(\mathcal{D}_I) := (\boldsymbol{X}^\top\boldsymbol{L}_I\boldsymbol{X}/k)^+\boldsymbol{X}^\top\boldsymbol{L}_I\boldsymbol{y}/k$ becomes the so-called *ridgeless* estimator, where $\boldsymbol{A}^+$ denotes the Moore-Penrose inverse of matrix $\boldsymbol{A}$.

Ensemble estimator. To introduce the ensemble estimator, it helps to define the set of all k distinct elements from $[n]$ to be $\mathcal{I}_k := \{\{i_1, i_2, \ldots, i_k\} : 1 \le i_1 < i_2 < \ldots < i_k \le n\}$. Note that the cardinality of $\mathcal{I}_k$ is $\binom{n}{k}$. For $\lambda \ge 0$, the ensemble estimator is then defined as:

$$\widetilde{\boldsymbol{\beta}}_{k,M}^\lambda(\mathcal{D}_n; \{I_\ell\}_{\ell=1}^M) := \frac{1}{M}\sum_{\ell\in[M]} \widehat{\boldsymbol{\beta}}_k^\lambda(\mathcal{D}_{I_\ell}), \quad (2)$$

where $I_1, \ldots, I_M$ are simple random samples from $\mathcal{I}_k$. The *full-ensemble* ridge estimator is the average of predictors fitted on all possible subsampled datasets:

$$\widetilde{\boldsymbol{\beta}}_{k,\infty}^\lambda(\mathcal{D}_n) := \frac{1}{|\mathcal{I}_k|}\sum_{I\in\mathcal{I}_k} \widehat{\boldsymbol{\beta}}_k^\lambda(\mathcal{D}_I) = \mathbb{E}[\widehat{\boldsymbol{\beta}}_k^\lambda(\mathcal{D}_I) \,|\, \mathcal{D}_n], \quad (3)$$

where the conditional expectation is taken with respect to the randomness of sampling from $\mathcal{I}_k$. Lemma A.1 shows that $\widetilde{\boldsymbol{\beta}}_{k,\infty}^\lambda(\mathcal{D}_n)$ is also almost surely equivalent to letting the ensemble size M tend to infinity in (2) conditioning on the full dataset $\mathcal{D}_n$, thus justifying the notation in (3). For simplicity, we drop the dependency on $\mathcal{D}_n$, $\{I_\ell\}_{\ell=1}^M$ and only write $\widetilde{\boldsymbol{\beta}}_{k,M}^\lambda$, $\widetilde{\boldsymbol{\beta}}_{k,\infty}^\lambda$, when it is clear from the context.

Prediction risk. We assess the performance of an M-ensemble predictor via conditional squared prediction risk:

$$R_{k,M}^\lambda := \mathbb{E}_{(\boldsymbol{x},y)}[(y - \boldsymbol{x}^\top\widetilde{\boldsymbol{\beta}}_{k,M}^\lambda)^2 \mid \mathcal{D}_n, \{I_\ell\}_{\ell=1}^M], \quad (4)$$

where $(\boldsymbol{x}, y)$ is an independent test point sampled from the distribution as $\mathcal{D}_n$. Note that the conditional risk $R_{k,M}^\lambda$ is a random variable that depends on both the dataset $\mathcal{D}_n$ and the random samples I_ℓ, $\ell = 1, \ldots, M$. For the full ensemble estimator $\widetilde{\boldsymbol{\beta}}_{k,\infty}^\lambda$, the conditional prediction risk is defined analogously, except the risk now only depends on $\mathcal{D}_n$:

$$R_{k,\infty}^\lambda := \mathbb{E}_{(\boldsymbol{x},y)}[(y - \boldsymbol{x}^\top\widetilde{\boldsymbol{\beta}}_{k,\infty}^\lambda)^2 \mid \mathcal{D}_n]. \quad (5)$$

2.1. Data Assumptions

For our theoretical results, we work under a proportional asymptotics regime, in which the original *data aspect ratio* (p/n) converges to $\phi \in (0, \infty)$ as $n, p \to \infty$, and the *subsample aspect ratio* (p/k) converges to ϕ_s as $k, p \to \infty$. Note that because $k \le n$, ϕ_s always lie in $[\phi, \infty]$. In addition, we impose two structural assumptions on the feature matrix and response vector as summarized in Assumptions 2.1 to 2.2, respectively.

Assumption 2.1 (Feature model)**.** The feature matrix decomposes as $\boldsymbol{X} = \boldsymbol{Z}\boldsymbol{\Sigma}^{1/2}$, where $\boldsymbol{Z} \in \mathbb{R}^{n\times p}$ contains i.i.d. entries with mean 0, variance 1, bounded moments of order $4+\delta$ for some $\delta > 0$, and $\boldsymbol{\Sigma} \in \mathbb{R}^{p\times p}$ is deterministic and symmetric with eigenvalues uniformly bounded between $r_{\min} > 0$ and $r_{\max} < \infty$. Let $\boldsymbol{\Sigma} = \sum_{j=1}^{p} r_j \boldsymbol{w}_j \boldsymbol{w}_j^\top$ denote the eigenvalue decomposition, where $(r_j, \boldsymbol{w}_j), j \in [p]$, are pairs of associated eigenvalue and normalized eigenvector. We assume there exists a deterministic distribution H such that the empirical spectral distribution of $\boldsymbol{\Sigma}$, $H_p(r) := p^{-1}\sum_{i=1}^{p} \mathbb{1}_{\{r_i \le r\}}$, weakly converges to H, almost surely (with respect to $\boldsymbol{X}$).

Assumption 2.2 (Response model)**.** The response vector decomposes as $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta}_0 + \boldsymbol{\epsilon}$, where $\boldsymbol{\beta}_0 \in \mathbb{R}^p$ is an unknown signal vector with ℓ_2-norm uniformly bounded and $\lim_{p\to\infty} \|\boldsymbol{\beta}_0\|_2^2 = \rho^2$, and $\boldsymbol{\epsilon}$ is an unobserved error vector independent of $\boldsymbol{X}$ with mean 0, variance σ^2, and bounded moment of order $4+\delta$ for some $\delta > 0$. We assume there exists a deterministic distribution G such that the empirical distribution of $\boldsymbol{\beta}_0$'s (squared) projection onto $\boldsymbol{\Sigma}$'s eigenspace, $G_p(r) := \|\boldsymbol{\beta}_0\|_2^{-2}\sum_{i=1}^{p}(\boldsymbol{\beta}_0^\top \boldsymbol{w}_i)^2\, \mathbb{1}_{\{r_i \le r\}}$, weakly converges to G, almost surely (with respect to $\boldsymbol{X}$).

Assumptions 2.1 and 2.2 are standard in the study of the ridge and ridgeless regression under proportional asymptotics; see, e.g., Hastie et al. (2022); Patil et al. (2022b;a). It is possible to further relax both of these assumptions. Specifically, one can incorporate other feature models, e.g., random features (Mei & Montanari, 2022), and can allow for certain non-linearities in the regression function (Bartlett et al., 2021) for the response model. We leave these for future work.

2.2. Risk Equivalences

Under the above assumptions, Lemma A.2 from Patil et al. (2022a) implies that for every $M \ge 1$, the prediction risk $R^{\lambda}_{k,M}$ of the ridge and ridgeless predictors in the full ensemble converges to some deterministic limit $\mathscr{R}^{\lambda}_{M}(\phi, \phi_s)$ as $k, n, p \to \infty$, $p/n \to \phi$ and $p/k \to \phi_s$. When $\phi_s = \phi$ (e.g., $k = n$), the asymptotic risk $\mathscr{R}^{\lambda}_{M}(\phi, \phi)$ is equal to $\mathscr{R}^{\lambda}_{1}(\phi, \phi)$ of the ridge predictor on the full dataset $\mathcal{D}_n$ for all $M \ge 1$, and we denote this risk simply by $\mathscr{R}^{\lambda}_{\infty}(\phi, \phi)$. To facilitate our discussion and for simplicity, Table 1 provides pointers to definitions of all important quantities used in the paper.

From a practical point of view, it is important to understand the least attainable risk that could be attained in the full ensemble. For the full ridge ensembles, we found that the explicit ridge regularization is unnecessary when considering optimal bagging and that the implicit regularization of ridgeless and subsampling suffices. The result below formalizes this empirical observation.

Theorem 2.3 (Optimal ridgeless ensemble vs optimal ridge)**.** *Under Assumptions 2.1 and 2.2, for all $\phi \in (0, \infty)$, we have*

$$\underbrace{\min_{\phi_s \ge \phi} \mathscr{R}^{0}_{\infty}(\phi, \phi_s)}_{\substack{\text{opt. ensemble}\\ \text{and no ridge}}} \overset{(a)}{=} \underbrace{\min_{\lambda \ge 0} \mathscr{R}^{\lambda}_{\infty}(\phi, \phi)}_{\substack{\text{no ensemble}\\ \text{and opt. ridge}}} \overset{(b)}{=} \underbrace{\min_{\substack{\phi_s \ge \phi,\\ \lambda \ge 0}} \mathscr{R}^{\lambda}_{\infty}(\phi, \phi_s)}_{\substack{\text{opt. ensemble}\\ \text{and opt. ridge}}}.$$

Further, if ϕ_s^* is the optimal subsample aspect ratio for ridgeless, and λ^* is the optimal ridge regularization with no subsampling, then for any $\theta \in [0, \lambda^*]$, full ridge ensemble with penalty parameter $\lambda = \lambda^* - \theta$ and subsample aspect ratio of $\phi_s = \phi + \theta(\phi_s^* - \phi)/\lambda^*$ also attains the optimal prediction risk.

In words, Theorem 2.3 says that optimizing subsample size (i.e. k) with the full ridgeless ensemble attains the same prediction risk as just optimizing the explicit regularization parameter (i.e., λ) of the ridge predictor. Further, both of them are the same as optimizing both k and λ. If one uses a lesser ridge penalty than needed for optimal prediction (i.e., uses $\lambda < \lambda^*$), then a full ensemble at a specific subsample aspect ratio $\phi_s = \phi + (1 - \lambda/\lambda^*)(\phi_s^* - \phi) > \phi$ can recover the remaining ridge regularization. In this sense, the implicit regularization provided by the ensemble amounts to adding more explicit ridge regularization. Similarly, one can supplement a sub-optimal implicit regularization of subsampling by adding explicit ridge regularization.

A special case of equivalence of (a) in Theorem 2.3 was previously formalized in LeJeune et al. (2020); Patil et al. (2022a) for isotropic covariates. Working with isotropic design helps their proof significantly, as the spectral distributions are the same for all p, n, and the closed-form expression of the asymptotic prediction risk can be derived analytically. However, in the general non-isotropic design, the asymptotic risk does not admit a closed-form expression, and one needs to account for this carefully.

General risk equivalences. Theorem 2.3 proves the risk equivalence of the ridge and full ensemble ridgeless when they attain minimum risk. Appendix A.3 shows a further risk equivalence in the full range, i.e., for any $\bar{\phi}_s \in [\phi, +\infty]$, there exists a $\bar{\lambda} \ge 0$ such that $\mathscr{R}^{0}_{\infty}(\phi, \bar{\phi}_s) = \mathscr{R}^{\bar{\lambda}}_{\infty}(\phi, \phi)$. Further, $\mathscr{R}^{\lambda}_{\infty}(\phi, \phi_s)$ remains constant as (λ, ϕ_s) varies on the line segment $(1-\theta)\cdot(\bar{\lambda}, \phi) + \theta\cdot(0, \bar{\phi}_s)$, for all $\theta \in [0, 1]$.

A remarkable implication of Theorem 2.3 is that for a fixed dataset $\mathcal{D}_n$, one does not need to tune both the subsample size (i.e., k) and the ridge regularization parameter (i.e., λ), but it suffices to fix for example $\lambda = 0$ and only tune ϕ_s. Alternatively, one can also fix $k = n$ and just tune $\lambda \ge 0$, which was considered in Patil et al. (2021). Performing tuning over $\lambda \ge 0$ requires one to discretize an infinite interval, while tuning the subsample size for a fixed λ only requires searching over a finite grid varying from $k = 1$ to $k = n$. For this reason, we fix λ and focus on tuning over k

Variable	M-ensemble				Full ensemble			
	Finite-sample		Asymptotic		Finite-sample		Asymptotic	
Prediction risk	$R^\lambda_{k,M}$	(4)	$\mathscr{R}^\lambda_M(\phi,\phi_s)$	(21)	$R^\lambda_{k,\infty}$	(5)	$\mathscr{R}^\lambda_\infty(\phi,\phi_s)$	(21)
Test error	$\overline{R}^\lambda_{k,M}$	(7)	$\mathscr{R}^\lambda_M(\phi,\phi_s)$	(21)				
Training error	$T^\lambda_{k,M}$	(6)	$\mathscr{T}^\lambda_M(\phi,\phi_s)$	(32)	$T^\lambda_{k,\infty}$	(8)	$\mathscr{T}^\lambda_\infty(\phi,\phi_s)$	(15)
GCV denominator	$D^\lambda_{k,M}$	(12)			$D^\lambda_{k,\infty}$	(13)	$\mathscr{D}^\lambda_\infty(\phi,\phi_s)$	(26)
GCV estimator	$\mathrm{gcv}^\lambda_{k,M}$	(11)			$\mathrm{gcv}^\lambda_{k,\infty}$	(11)	$\mathscr{G}^\lambda_\infty(\phi,\phi_s)$	(16)

Table 1. Summary of notations and pointers to definitions of important empirical quantities used in this paper and their asymptotic limits.

in this paper. In the next section, we investigate the problem of tuning the subsample size in the full ensemble to achieve the minimum oracle risk via generalized cross-validation.

3. Generalized Cross-Validation

Suppose $\widehat{f}(\cdot;\mathcal{D}_n):\mathbb{R}^p\to\mathbb{R}$ is a predictor trained on $\mathcal{D}_n$. We call $\widehat{f}(\cdot;\mathcal{D}_n)$ a linear smoother if $\widehat{f}(\boldsymbol{x};\mathcal{D}_n)=\boldsymbol{a}_{\boldsymbol{x}}^\top\boldsymbol{y}$ for some vector $\boldsymbol{a}_{\boldsymbol{x}}$ that only depends on the design $\boldsymbol{X}$ (and $\boldsymbol{x}$). Define the smoothing matrix $\boldsymbol{S}\in\mathbb{R}^{n\times n}$ with rows $\boldsymbol{a}_{\boldsymbol{x}_1}^\top,\dots,\boldsymbol{a}_{\boldsymbol{x}_n}^\top$, which in turn is only a function of $\boldsymbol{X}$. For any linear smoother, the generalized cross-validation (GCV) estimator of the prediction risk is defined to be $n^{-1}\|\boldsymbol{y}-\boldsymbol{S}\boldsymbol{y}\|_2^2/(1-n^{-1}\operatorname{tr}(\boldsymbol{S}))^2$; see, e.g., Wasserman (2006, Section 5.3). The numerator of GCV is the training error, which typically is biased downwards, and the denominator attempts to account for such optimism of the predictor.

Ensemble GCV. Before we analyze GCV for the ridge ensemble, we first introduce some notations. Let $I_{1:M}:=\cup_{\ell=1}^M I_\ell$ and $I^c_{1:M}:=[n]\setminus I_{1:M}$. We define the *in-sample* training error and the *out-of-sample* test error of $\widetilde{\boldsymbol{\beta}}^\lambda_{k,M}$ as:

$$T^\lambda_{k,M}:=\frac{1}{|I_{1:M}|}\sum_{i\in I_{1:M}}(y_i-\boldsymbol{x}_i^\top\widetilde{\boldsymbol{\beta}}^\lambda_{k,M})^2, \tag{6}$$

$$\overline{R}^\lambda_{k,M}:=\frac{1}{|I^c_{1:M}|}\sum_{i\in I^c_{1:M}}(y_i-\boldsymbol{x}_i^\top\widetilde{\boldsymbol{\beta}}^\lambda_{k,M})^2. \tag{7}$$

Since the full ensemble estimator $\widetilde{\boldsymbol{\beta}}^\lambda_{k,\infty}$ uses all the data $\mathcal{D}_n$, its training error, denoted by $T^\lambda_{k,\infty}$, is simply:

$$T^\lambda_{k,\infty}:=\frac{1}{n}\sum_{i\in[n]}(y_i-\boldsymbol{x}_i^\top\widetilde{\boldsymbol{\beta}}^\lambda_{k,\infty})^2. \tag{8}$$

Since $I_{1:M}\xrightarrow{\text{a.s.}}[n]$ for any $n\in\mathbb{N}$ as $M\to\infty$, the notation $T^\lambda_{k,\infty}$ in (8) is justified as a limiting case of (6) (see Appendix A.1 for more details). Now, observe that a ridge ensemble is a linear smoother because $\boldsymbol{X}_{I_{1:M}}\widetilde{\boldsymbol{\beta}}^\lambda_{k,M}=\boldsymbol{S}^\lambda_{k,M}\boldsymbol{y}_{I_{1:M}}$, where the smoothing matrix $\boldsymbol{S}^\lambda_{k,M}$ is given by:

$$\boldsymbol{S}^\lambda_{k,M}=\frac{1}{M}\sum_{\ell=1}^M\boldsymbol{X}_{I_\ell}(\boldsymbol{X}_{I_\ell}^\top\boldsymbol{X}_{I_\ell}/k+\lambda\boldsymbol{I}_p)^+\boldsymbol{X}_{I_\ell}^\top/k. \tag{9}$$

Analogously, the smoothing matrix for $\widetilde{\boldsymbol{\beta}}^\lambda_{k,\infty}$ is given by:

$$\boldsymbol{S}^\lambda_{k,\infty}=\frac{1}{|\mathcal{I}_k|}\sum_{I\in\mathcal{I}_k}\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{L}_I\boldsymbol{X}/k+\lambda\boldsymbol{I}_p)^+\boldsymbol{X}^\top\boldsymbol{L}_I/k. \tag{10}$$

Thus, the GCV estimates for ridge predictors in the finite and full ensemble case are respectively given by:

$$\mathrm{gcv}^\lambda_{k,M}=\frac{T^\lambda_{k,M}}{D^\lambda_{k,M}},\qquad \mathrm{gcv}^\lambda_{k,\infty}=\frac{T^\lambda_{k,\infty}}{D^\lambda_{k,\infty}}, \tag{11}$$

where the denominators $D^\lambda_{k,M}$ and $D^\lambda_{k,\infty}$ are as follows:

$$D^\lambda_{k,M}:=(1-|I_{1:M}|^{-1}\operatorname{tr}(\boldsymbol{S}^\lambda_{k,M}))^2, \tag{12}$$

$$D^\lambda_{k,\infty}:=(1-n^{-1}\operatorname{tr}(\boldsymbol{S}^\lambda_{k,\infty}))^2. \tag{13}$$

3.1. Full-Ensemble Uniform Consistency

Let $\mathcal{K}_n\subset\{0,1,\dots,n\}$ be a grid of subsample sizes that covers the full range of $[0,n]$ asymptotically in the sense that $\{k/n:\ k\in\mathcal{K}_n\}$ "converges" to the set $[0,1]$ as $n\to\infty$. One simple choice is to set

$$\mathcal{K}_n=\{0,k_0,2k_0,\dots.\lfloor n/k_0\rfloor k_0\},$$

where the increment is $k_0=\lfloor n^\nu\rfloor$ for some $\nu\in(0,1)$. Here, we adopt the convention that when $k=0$, the predictor reduces to a null predictor that always returns zero. Based on the definition above, we now present the uniform consistency results of the GCV estimator (11) for full ensembles when the ridge regularization parameter λ is fixed.

Theorem 3.1 (Uniform consistency of GCV)**.** *Suppose Assumptions 2.1 and 2.2 hold. Then, for all $\lambda\geq 0$, we have*

$$\max_{k\in\mathcal{K}_n}|\mathrm{gcv}^\lambda_{k,\infty}-R^\lambda_{k,\infty}|\xrightarrow{\text{a.s.}}0,$$

as $n,p\to\infty$ such that $p/n\to\phi\in(0,\infty)$.

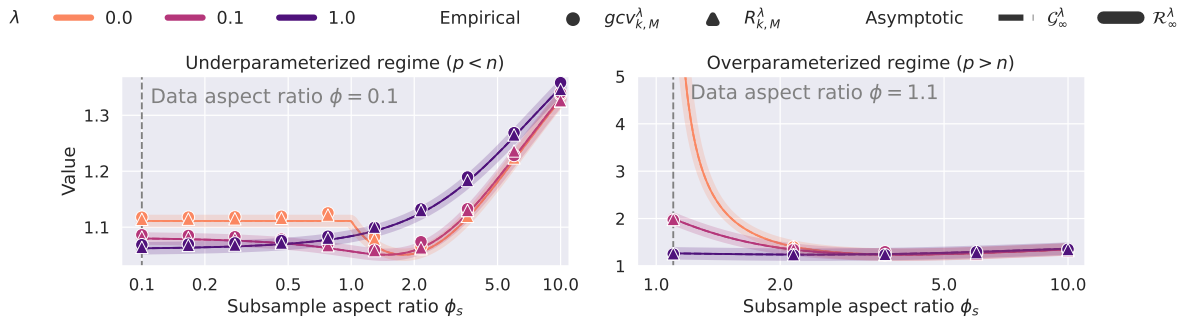

Figure 2. Asymptotic risk and GCV curves for full ridge ensembles, under model (M-AR1) when $\rho_{\mathrm{AR1}} = 0.5$ and $\sigma^2 = 1$ with varying regularization parameters $\lambda \in \{0, 0.1, 1\}$ and subsample sizes $k = \lfloor p/\phi_s \rfloor$. The points denote finite-sample risks averaged over 50 dataset repetitions with an ensemble size of $M = 500$, with $n = \lfloor p/\phi \rfloor$ and $p = 500$. The left and the right panels illustrate the underparameterized and overparameterized cases with the limiting data aspect ratio $\phi = 0.1$ and $\phi = 1.1$, respectively.

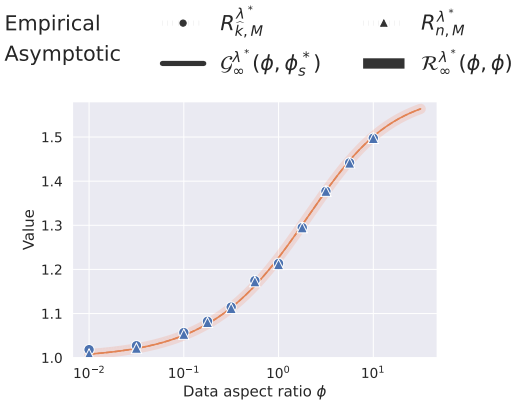

Figure 3. Asymptotic prediction risk curves with optimal tuned parameters λ^* and ϕ_s^*, under model (M-AR1) when $\rho_{\mathrm{AR1}} = 0.5$, $\sigma^2 = 1$, for varying data aspect ratio ϕ. The curves represent the theoretical asymptotic GCV estimate in the full ensemble and the asymptotic risk of the optimal ridge predictors. The points represent the finite-sample risks of the best 500-ensemble ridgeless and the best ridge predictor averaged over 50 dataset repetitions, with $n = \lfloor p/\phi \rfloor$ and $p = 500$.

Theorem 3.1 shows the uniform consistency of GCV in the full ensemble for fixed subsample size k and ridge regularization parameter λ. The almost sure qualification in Theorem 3.1 is with respect the entire training data $(\boldsymbol{X}, \boldsymbol{y})$. An implication of Theorem 3.1 is that one can select the optimal subsample size in a data-dependent manner, i.e., selecting $\widehat{k}^{\lambda} \in \mathrm{argmin}_{k \in \mathcal{K}_n} \mathrm{gcv}^{\lambda}_{k,\infty}$ guarantees to track the minimum prediction risk $\min_{k \in [n]} R^{\lambda}_{k,\infty}$ asymptotically.

We first provide numerical illustrations for Theorem 3.1 under the non-isotropic AR(1) data model, which is the same as the one used for Figure 1; see Appendix I for model details. Figure 2 shows both the GCV estimate and the asymptotic risk for the full ridge ensemble. We observe a close match of the theoretical curves and the GCV estimates.

Combining Theorem 2.3 and Theorem 3.1, we can obtain the following corollary regarding GCV subsample tuning.

Corollary 3.2 (Ridge tuning by GCV subsample tuning)**.** *Suppose Assumptions 2.1 and 2.2 hold. Then, we have*

$$\mathrm{gcv}^{0}_{\widehat{k}^0,\infty} \xrightarrow{\text{a.s.}} \min_{\phi_s \geq \phi, \lambda \geq 0} \mathscr{R}^{\lambda}_{\infty}(\phi, \phi_s),$$

as $k, n, p \to \infty$ such that $p/n \to \phi \in (0, \infty)$.

Corollary 3.2 certifies the validity of GCV tuning for achieving the optimal risk over all possible regularization parameters and subsample sizes. In practice, tuning for the ridge parameter λ requires one to determine a grid of λ's for cross-validation. However, the maximum value for the grid is generally chosen by some ad hoc criteria. For example, there is no default maximum value for ridge tuning in the widely-used package `glmnet` (Friedman et al., 2010). From Theorem 2.3, when the signal-noise ratio ρ^2/σ^2 is small, the subsample size should be small enough (so that ϕ_s is large), and the range of λ's grid should be large enough to cover its optimal value. On the contrary, the GCV-based method does not need such an upper bound for the grid $\mathcal{K}_n$ of subsample sizes because the sample size provides a natural grid in finite samples, informed by the dataset.

In Corollary 3.2, we fix the ridge regularization parameter λ to be zero. But, one can also use other value of $\lambda < \lambda^*$ and the similar statement still holds with $\mathrm{gcv}^{0}_{\widehat{k}^0,\infty}$ replaced by $\mathrm{gcv}^{\lambda}_{\widehat{k}^{\lambda},\infty}$ based on Theorem 2.3. Furthermore, one can construct the estimator of λ^* as $\widehat{\lambda} = \lambda(n - \widehat{k}^0)/(\widehat{k}^{\lambda} - \widehat{k}^0)$ by extrapolating the line segment between $(0, \widehat{k}^0)$ and $(\lambda, \widehat{k}^{\lambda})$.

In Figure 3, we numerically compare the optimal subsampled ridgeless ensemble with the optimal ridge predictor to verify Corollary 3.2. As we can see, their theoretical curves exactly match, and the empirical estimates in finite samples are also close to their asymptotic limits.

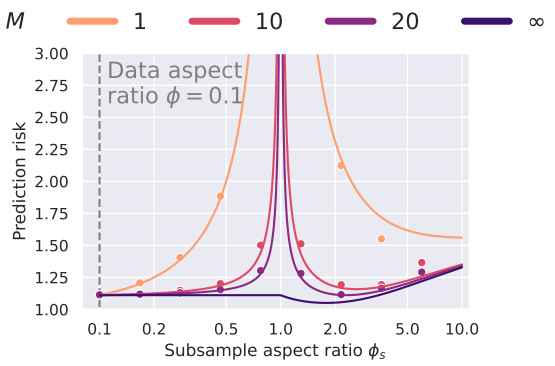

Figure 4. Asymptotic prediction risk curves of ridgeless ensembles, under model (M-AR1) when $\rho_{\mathrm{AR1}} = 0.5$, $\sigma^2 = 1$, and $\phi = 0.1$. The points denote the finite-sample GCV estimates of ridgeless ensembles for varying ensemble sizes $M \in \{1, 10, 20\}$ averaged over 50 dataset repetitions, with $n = \lfloor p/\phi \rfloor$ and $p = 500$.

3.2. A Finite-Ensemble Inconsistency Result

While deriving GCV asymptotics in the proof of Theorem 3.1 for the full ensemble, we also obtain as a byproduct the asymptotic limit of the GCV estimate for finite ensembles. From related work (Patil et al., 2021) and Theorem 3.1, we already know that $\mathrm{gcv}^{\lambda}_{k,1}$ and $\mathrm{gcv}^{\lambda}_{k,\infty}$ are consistent estimators of the non-ensemble risk ($R^{\lambda}_{k,1}$) and the full ensemble risk ($R^{\lambda}_{k,\infty}$), respectively. However, $\mathrm{gcv}^{\lambda}_{k,M}$ for $1 < M < \infty$ may not be consistent, which is somewhat surprising. As an example, GCV for $M = 2$ is not a consistent estimator for the prediction risk $R^{\lambda}_{k,2}$, as shown in the following proposition.

Proposition 3.3 (GCV inconsistency for ridgeless, $M = 2$)**.** *Suppose Assumptions 2.1 and 2.2 hold with $\rho^2, \sigma^2 \in (0, \infty)$. Then, for any $\phi \in (0, \infty)$, we have*

$$|\mathrm{gcv}^{0}_{k,2} - R^{0}_{k,2}| \overset{\mathrm{p}}{\not\to} 0,$$

as $k, n, p \to \infty$, $p/n \to \phi$, $p/k \to \phi_s \in (1, \infty) \cap (\phi, \infty)$.

Intuitively, the inconsistency for a finite in large part happens because, for a finite M, the residuals computed using the bagged predictor contain non-negligible fractions of out-of-sample and in-sample, and all of them are treated equally. As a result, the GCV estimate for finite ensembles indirectly relates to the original data through the aspect ratios (ϕ, ϕ_s), even though the GCV estimate is computed only using the training observations. See Section 5 about possible approaches for the corrected GCV estimate for arbitrary ensemble sizes. Though in practice, the correction may not be crucial for a moderate M, because the GCV estimate is close to the underlying target as shown in Figure 4.

3.3. Proof Outline of Theorem 3.1

There are three key steps are involved to prove Theorem 3.1. (1) Deriving the asymptotic limit of the prediction risk $R^{\lambda}_{k,M}$. (2) Deriving the asymptotic limit of GCV estimate $\mathrm{gcv}^{\lambda}_{k,\infty}$. (3) Showing pointwise consistency in k by matching the two limits and then lifting to uniform convergence in k. We briefly explain key ideas for showing the three steps below.

(1) Asymptotic limit of risk. We build upon prior results on the risk analysis of ridge ensembles. Under Assumptions 2.1 and 2.2, Lemma A.2 adapted from Patil et al. (2022a) implies that the conditional prediction risks under proportional asymptotics converge to certain deterministic limits:

$$R^{\lambda}_{k,M} \xrightarrow{\text{a.s.}} \mathscr{R}^{\lambda}_{M}(\phi, \phi_s), \qquad R^{\lambda}_{k,\infty} \xrightarrow{\text{a.s.}} \mathscr{R}^{\lambda}_{\infty}(\phi, \phi_s), \quad (14)$$

where $\mathscr{R}^{\lambda}_{M}(\phi, \phi_s)$, $\mathscr{R}^{\lambda}_{\infty}(\phi, \phi_s)$ are as defined in (21).

(2) Asymptotic limit of GCV. To analyze the asymptotic behavior of the GCV estimates, we obtain the asymptotics of the denominator and the numerator of GCV separately. We first show the regular cases when $\phi_s < \infty$ and $\lambda > 0$, and then incorporate boundary cases of $\phi_s = \infty$ and $\lambda = 0$. Our analysis begins with the following lemma that provides asymptotics for the denominator $D^{\lambda}_{k,\infty}$ (as in (13)) of GCV:

Lemma 3.4 (Asymptotics of the GCV denominator)**.** *Suppose Assumption 2.1 holds. Then, for all $\lambda > 0$,*

$$D^{\lambda}_{k,\infty} \xrightarrow{\text{a.s.}} \mathscr{D}^{\lambda}_{\infty}(\phi, \phi_s),$$

as $k, n, p \to \infty$, $p/n \to \phi \in (0, \infty)$, $p/k \to \phi_s \in [\phi, \infty)$.

It is worth noting that Lemma 3.4 does not require Assumption 2.2 because the smoothing matrix only concerns the design matrix $\boldsymbol{X}$ and does not depend on the response $\boldsymbol{y}$.

Towards obtaining asymptotics for the numerator $T^{\lambda}_{k,\infty}$ (as in (8)) of GCV, we first decompose $T^{\lambda}_{k,\infty}$ into simpler components via Lemma D.1. Specifically, the full mean squared training error admits the following decomposition:

$$T^{\lambda}_{k,\infty} - \sum_{m=1}^{2} (c_m T^{\lambda}_{k,m} + (1 - c_m) \overline{R}^{\lambda}_{k,m}) \xrightarrow{\text{a.s.}} 0,$$

where $c_1 = \phi/\phi_s$ and $c_2 = 2\phi(2\phi_s - \phi)/\phi_s^2$. Here, $T^{\lambda}_{k,m}$ and $\overline{R}^{\lambda}_{k,m}$ are the in-sample training and out-of-sample test errors of the m-ensemble for $m = 1$ and 2, as defined in (6) and (7). This decomposition implies that the full training error is asymptotically simply a linear combination of training and test errors. Therefore, it suffices to obtain the asymptotics of each of these components. As analyzed in Lemma D.2, it is easy to show that the test errors converge to $\mathscr{R}^{\lambda}_{m}$ for $m = 1, 2$. On the other hand, it is more challenging to derive the asymptotic limits for the training errors $T^{\lambda}_{k,m}$. We first split the $T^{\lambda}_{k,m}$ into finer components via a

bias-variance decomposition of $T^{\lambda}_{k,m}$. By developing novel deterministic equivalents of resolvents arising from the decomposition of in-sample errors in Lemma F.8, we are able to show the convergence of the bias and variance components (see Lemma D.3). Combining Lemmas D.1 to D.3 yields the convergence of $T^{\lambda}_{k,m}$ to a deterministic limit $\mathscr{T}^{\lambda}_{m}$ as summarized in the following lemma:

Lemma 3.5 (Asymptotics of the GCV numerator)**.** *Suppose Assumptions 2.1 and 2.2 hold. Then, for all $\lambda > 0$,*

$$T^{\lambda}_{k,\infty} \xrightarrow{\text{a.s.}} \mathscr{T}^{\lambda}_{\infty} = \sum_{m=1}^{2} (c_m \mathscr{T}^{\lambda}_{m} + (1 - c_m)\mathscr{R}^{\lambda}_{m}), \quad (15)$$

as $k, n, p \to \infty$, $p/n \to \phi \in (0, \infty)$, $p/k \to \phi_s \in [\phi, \infty)$, where $c_1 = \phi/\phi_s$ and $c_2 = 2\phi(2\phi_s - \phi)/\phi_s^2$.

Finally, the boundary cases when $\phi_s = +\infty$ and $\lambda = 0$ are taken care of in succession by Proposition E.1 and Proposition E.2, respectively. Combining the above results provides the asymptotics for the GCV estimate in the full ensemble:

Proposition 3.6 (Asymptotics of GCV for full ensemble)**.** *Suppose Assumptions 2.1 and 2.2 hold. Then, for all $\lambda \geq 0$,*

$$\text{gcv}^{\lambda}_{k,\infty} \xrightarrow{\text{a.s.}} \mathscr{G}^{\lambda}_{\infty}(\phi, \phi_s) := \frac{\mathscr{T}^{\lambda}_{\infty}(\phi, \phi_s)}{\mathscr{D}^{\lambda}_{\infty}(\phi, \phi_s)}, \quad (16)$$

as $k, n, p \to \infty$, $p/n \to \phi \in (0, \infty)$, $p/k \to \phi_s \in [\phi, \infty]$.

(3) Asymptotics matching and uniform convergence. The asymptotic limits obtained in the first steps can be shown to match with each other by algebraic manipulations. This shows the pointwise consistency in Theorem 3.1. The uniform convergence then follows by applying a certain Cesàro-type mean convergence lemma (see Lemma G.5).

4. Real Data Example: Single-Cell Multiomics

We compare tuning subsample size in the full ridgeless ensemble with tuning the ridge parameter on the full data in a real-world data example from multiomics. This single-cell CITE-seq dataset from Hao et al. (2021) consists of 50,781 human peripheral blood mononuclear cells (PBMCs) originating from eight volunteers post-vaccination (day 3) of an HIV vaccine, which simultaneously measures 20,729 genes and 228 proteins in individual cells.

We follow the standard preprocessing procedure in single-cell data analysis (Hao et al., 2021; Du et al., 2022) to select the top 5,000 highly variable genes and the top 50 highly abundant surface proteins, which exhibit high cell-to-cell variations in the dataset. The gene expression and protein abundance counts for each cell are then divided by the total counts for that cell and multiplied by 10^4 and log-normalized. We randomly hold out half of the cells in each cell type as a test set. The top 500 principal components

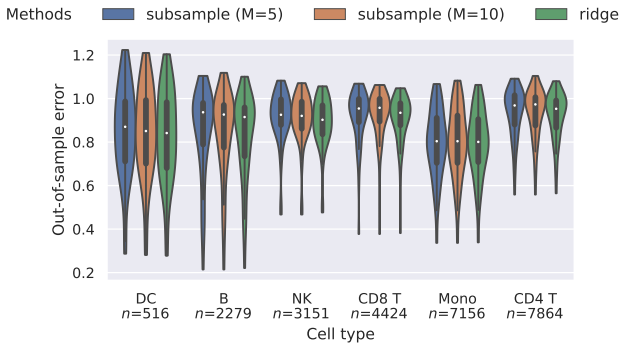

Figure 5. Violin plots of mean squared errors on the randomly held out test sets of different tuning methods for predicting the abundances of 50 proteins in the single-cell CITE-seq dataset. The sizes of the test sets are the same as the sizes of the training sets.

of the standardized gene expressions are used as features to predict protein abundances. The results of using ensembles with subsample size (k) tuning and ridge tuning (λ) without subsampling based on the GCV estimates are compared in Figure 5. For the former, we search over the grid of 25 k's from n^{ν} to n spaced evenly on the log scale, with $\nu = 0.5$ and sample size n ranges from 516 to 7864 for different cell types. For the latter, we search over the grid of 100 λ's from 10^{-2} to 10^{2} spaced evenly on log scale. Since different cell types have different sample sizes, this results in different data aspect ratios, presented in increasing order in Figure 5.

From Figure 5, we see that using a moderate ensemble size ($M = 5$ or 10) and tuning the subsample size have a very similar performance to only tuning the ridge regularization parameter in the full dataset. This suggests that the results of Corollary 3.2 also hold even on real data for different data aspect ratios. As discussed after Corollary 3.2, subsample tuning is easier to implement because the dataset provides a natural lower and upper bound of the subsample size. On the other hand, ridge tuning requires one to heuristically pick the upper regularization threshold for the search grid.

5. Discussion and Future Directions

In this work, we provide the risk characterization for the full ridge ensemble and establish the oracle risk equivalences between the full ridgeless ensemble and ridge regression. At a high level, these equivalences show that implicit regularization induced by subsampling matches explicit ridge regularization, i.e., a subsampled ridge predictor with penalty λ_1 has the same risk as another ridge predictor with penalty $\lambda_2 \geq \lambda_1$. Additionally, we prove the uniform consistency of generalized cross-validation for full ridge ensembles, which implies the validity of GCV tuning (that does not require sample splitting) for optimal predictive performance. We describe next some avenues for future work moving forward.

Bias correction for finite ensembles. In Proposition 3.3, we show that the GCV estimate can be inconsistent in the finite ridge and ridgeless ensembles. The inconsistency for $M = 2$ occurs because sampling from the whole dataset induces extra randomness beyond those of the training observations used to compute the GCV estimates. Our analysis of GCV for the full ensemble suggests a way to correct the bias of the GCV estimate. In Appendix H, we outline a possible correction strategy for finite ensembles based on out-of-bag estimates. An intriguing next research direction is to investigate the implementation and uniform consistency of the corrected GCV for finite ensembles in detail.

Extensions to other error metrics. In this paper, we focus on the in-distribution squared prediction risk. It is of interest to extend the equivalences for other error metrics, such as squared estimation risk, general prediction risks, and other functionals of the out-of-sample error distribution, like the quantiles of the error distribution. Additionally, for the purposes of tuning, it is also of interest to extend the GCV analysis to estimate such functionals of the out-of-sample error distribution. Such functional estimation could be valuable in constructing prediction intervals for the unknown response. The technical tools introduced in Patil et al. (2022c) involving leave-one-out perturbation techniques could prove useful for such an extension. Furthermore, this extension would also allow for extending the results presented in this paper to hold under a general non-linear response model.

Extensions to other base predictors. Finally, the focus of this paper is the base ridge predictor. A natural extension of the current work is to consider kernel ridge regression. Going further, it is of much interest to consider other regularized predictors, such as lasso. Whether optimal subsampled lassoless regression still matches with the optimal lasso is an interesting question. There is already empirical evidence along the lines of Figure 1 for such a connection. The results proved in the current paper make us believe that there is a general story quantifying the effect of implicit regularization by subsampling and that provided by explicit regularization. Whether the general story unfolds as neatly as presented here for ridge regression remains an exciting next question!

Acknowledgements

We are grateful to Ryan Tibshirani, Alessandro Rinaldo, Yuting Wei, Matey Neykov, Daniel LeJeune, Shamindra Shrotriya for many helpful conversations surrounding ridge regression, subsampling, and generalized cross-validation. Many thanks are also due to the anonymous reviewers for their encouraging remarks and insightful questions that have informed several new directions for follow-up future work. In particular, special thanks to the reviewer "2t75" for a wonderful review and highlighting other related works on subsampling that have made their way into the manuscript.

References

Adlam, B. and Pennington, J. Understanding double descent requires a fine-grained bias-variance decomposition. *Advances in neural information processing systems*, 33: 11022–11032, 2020a.

Adlam, B. and Pennington, J. The neural tangent kernel in high dimensions: Triple descent and a multi-scale theory of generalization. In *International Conference on Machine Learning*. PMLR, 2020b.

Allen, D. M. The relationship between variable selection and data augmentation and a method for prediction. *Technometrics*, 16(1):125–127, 1974.

Arlot, S. and Celisse, A. A survey of cross-validation procedures for model selection. *Statistics surveys*, 4:40–79, 2010.

Bai, Z. and Silverstein, J. W. *Spectral Analysis of Large Dimensional Random Matrices*. Springer, 2010. Second edition.

Bartlett, P. L., Long, P. M., Lugosi, G., and Tsigler, A. Benign overfitting in linear regression. *Proceedings of the National Academy of Sciences*, 117(48):30063–30070, 2020.

Bartlett, P. L., Montanari, A., and Rakhlin, A. Deep learning: a statistical viewpoint. *Acta numerica*, 30:87–201, 2021.

Belkin, M., Hsu, D., and Xu, J. Two models of double descent for weak features. *SIAM Journal on Mathematics of Data Science*, 2(4):1167–1180, 2020.

Bloemendal, A., Knowles, A., Yau, H.-T., and Yin, J. On the principal components of sample covariance matrices. *Probability theory and Related Fields*, 164(1):459–552, 2016.

Breiman, L. Bagging predictors. *Machine Learning*, 24(2): 123–140, 1996.

Bühlmann, P. and Yu, B. Analyzing bagging. *The Annals of Statistics*, 30(4):927–961, 2002.

Buja, A. and Stuetzle, W. Observations on bagging. *Statistica Sinica*, pp. 323–351, 2006.

Craven, P. and Wahba, G. Estimating the correct degree of smoothing by the method of generalized cross-validation. *Numerische Mathematik*, 31:377–403, 1979.

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

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

Du, J.-H., Cai, Z., and Roeder, K. Robust probabilistic modeling for single-cell multimodal mosaic integration and imputation via scvaeit. *Proceedings of the National Academy of Sciences*, 119(49):e2214414119, 2022.

d'Ascoli, S., Refinetti, M., Biroli, G., and Krzakala, F. Double trouble in double descent: Bias and variance (s) in the lazy regime. In *International Conference on Machine Learning*, pp. 2280–2290. PMLR, 2020.

El Karoui, N. Asymptotic behavior of unregularized and ridge-regularized high-dimensional robust regression estimators: rigorous results. *arXiv preprint arXiv:1311.2445*, 2013.

El Karoui, N. On the impact of predictor geometry on the performance on high-dimensional ridge-regularized generalized robust regression estimators. *Probability Theory and Related Fields*, 170(1):95–175, 2018.

Erdős, L. and Yau, H.-T. *A Dynamical Approach to Random Matrix Theory*. American Mathematical Society, 2017.

Friedman, J., Hastie, T., and Tibshirani, R. Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software*, 33(1):1, 2010.

Friedman, J. H. and Hall, P. On bagging and nonlinear estimation. *Journal of Statistical Planning and Inference*, 137(3):669–683, 2007.

Geisser, S. The predictive sample reuse method with applications. *Journal of the American statistical Association*, 70(350):320–328, 1975.

Golub, G. H., Heath, M., and Wahba, G. Generalized cross-validation as a method for choosing a good ridge parameter. *Technometrics*, 21(2):215–223, 1979.

Greene, E. and Wellner, J. A. Exponential bounds for the hypergeometric distribution. *Bernoulli*, 23(3):1911, 2017.

Grenander, U. and Szegö, G. *Toeplitz Forms and Their Applications*. University of California Press, 1958. First edition.

Gut, A. *Probability: A Graduate Course*. Springer, New York, 2005.

Györfi, L., Kohler, M., Krzyzak, A., and Walk, H. *A Distribution-free Theory of Nonparametric Regression*. Springer Science & Business Media, 2006.

Hall, P. and Samworth, R. J. Properties of bagged nearest neighbour classifiers. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(3):363–379, 2005.

Hao, Y., Hao, S., Andersen-Nissen, E., Mauck III, W. M., Zheng, S., Butler, A., Lee, M. J., Wilk, A. J., Darby, C., Zager, M., et al. Integrated analysis of multimodal single-cell data. *Cell*, 2021.

Hastie, T. Ridge regularization: An essential concept in data science. *Technometrics*, 62(4):426–433, 2020.

Hastie, T., Tibshirani, R., and Friedman, J. *The Elements of Statistical Learning*. Springer Series in Statistics, 2009. Second edition.

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

Hoerl, A. E. and Kennard, R. W. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970a.

Hoerl, A. E. and Kennard, R. W. Ridge regression: applications to nonorthogonal problems. *Technometrics*, 12(1): 69–82, 1970b.

Kale, S., Kumar, R., and Vassilvitskii, S. Cross-validation and mean-square stability. In *In Proceedings of the Second Symposium on Innovations in Computer Science*, 2011.

Krogh, A. and Sollich, P. Statistical mechanics of ensemble learning. *Physical Review E*, 55(1):811, 1997.

Kumar, R., Lokshtanov, D., Vassilvitskii, S., and Vattani, A. Near-optimal bounds for cross-validation via loss stability. In *International Conference on Machine Learning*, 2013.

LeJeune, D., Javadi, H., and Baraniuk, R. The implicit regularization of ordinary least squares ensembles. In *International Conference on Artificial Intelligence and Statistics*, 2020.

Li, K.-C. From Stein's unbiased risk estimates to the method of generalized cross validation. *The Annals of Statistics*, pp. 1352–1377, 1985.

Li, K.-C. Asymptotic optimality of c_l and generalized cross-validation in ridge regression with application to spline smoothing. *The Annals of Statistics*, 14(3):1101–1112, 1986.

Li, K.-C. Asymptotic optimality for c_p, c_l, cross-validation and generalized cross-validation: Discrete index set. *The Annals of Statistics*, 15(3):958–975, 1987.

Loureiro, B., Gerbelot, C., Refinetti, M., Sicuro, G., and Krzakala, F. Fluctuations, bias, variance & ensemble of learners: Exact asymptotics for convex losses in high-dimension. In *International Conference on Machine Learning*, pp. 14283–14314. PMLR, 2022.

Mei, S. and Montanari, A. The generalization error of random features regression: Precise asymptotics and the double descent curve. *Communications on Pure and Applied Mathematics*, 75(4):667–766, 2022.

Miolane, L. and Montanari, A. The distribution of the lasso: Uniform control over sparse balls and adaptive parameter tuning. *The Annals of Statistics*, 49(4):2313–2335, 2021.

Muthukumar, V., Vodrahalli, K., Subramanian, V., and Sahai, A. Harmless interpolation of noisy data in regression. *IEEE Journal on Selected Areas in Information Theory*, 1 (1):67–83, 2020.

Obuchi, T. and Kabashima, Y. Cross validation in LASSO and its acceleration. *Journal of Statistical Mechanics: Theory and Experiment*, 2016.

Patil, P., Wei, Y., Rinaldo, A., and Tibshirani, R. Uniform consistency of cross-validation estimators for high-dimensional ridge regression. In *International Conference on Artificial Intelligence and Statistics*. PMLR, 2021.

Patil, P., Du, J.-H., and Kuchibhotla, A. K. Bagging in overparameterized learning: Risk characterization and risk monotonization. *arXiv preprint arXiv:2210.11445*, 2022a.

Patil, P., Kuchibhotla, A. K., Wei, Y., and Rinaldo, A. Mitigating multiple descents: A model-agnostic framework for risk monotonization. *arXiv preprint arXiv:2205.12937*, 2022b.

Patil, P., Rinaldo, A., and Tibshirani, R. Estimating functionals of the out-of-sample error distribution in high-dimensional ridge regression. In *International Conference on Artificial Intelligence and Statistics*. PMLR, 2022c.

Rad, K. R. and Maleki, A. A scalable estimate of the out-of-sample prediction error via approximate leave-one-out cross-validation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 82(4):965–996, 2020.

Rad, K. R., Zhou, W., and Maleki, A. Error bounds in estimating the out-of-sample prediction error using leave-one-out cross validation in high-dimensions. In *International Conference on Artificial Intelligence and Statistics*. PMLR, 2020.

Rubio, F. and Mestre, X. Spectral convergence for a general class of random matrices. *Statistics & probability letters*, 81(5):592–602, 2011.

Rudin, W. *Principles of Mathematical Analysis*. McGraw-Hill New York, 1976.

Samworth, R. J. Optimal weighted nearest neighbour classifiers. *The Annals of Statistics*, 40(5):2733–2763, 2012.

Sollich, P. and Krogh, A. Learning with ensembles: How overfitting can be useful. *Advances in neural information processing systems*, 8, 1995.

Stone, M. Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society: Series B*, 36(2):111–133, 1974.

Stone, M. Asymptotics for and against cross-validation. *Biometrika*, 64(1):29–35, 1977.

Sur, P., Chen, Y., and Candès, E. J. The likelihood ratio test in high-dimensional logistic regression is asymptotically a rescaled chi-square. *Probability Theory and Related Fields*, 175(1):487–558, 2019.

Thrampoulidis, C., Oymak, S., and Hassibi, B. Regularized linear regression: A precise analysis of the estimation error. In *Conference on Learning Theory*, pp. 1683–1709. PMLR, 2015.

Thrampoulidis, C., Abbasi, E., and Hassibi, B. Precise error analysis of regularized M-estimators in high dimensions. *IEEE Transactions on Information Theory*, 64(8):5592–5628, 2018.

Wang, S., Zhou, W., Lu, H., Maleki, A., and Mirrokni, V. Approximate leave-one-out for fast parameter tuning in high dimensions. *arXiv preprint arXiv:1807.02694*, 2018.

Wasserman, L. *Olive Nonparametric Statistics*. Springer, 2006.

Wei, A., Hu, W., and Steinhardt, J. More than a toy: Random matrix models predict how real-world neural representations generalize. *arXiv preprint arXiv:2203.06176*, 2022.

Xu, J., Maleki, A., and Rad, K. R. Consistent risk estimation in high-dimensional linear regression. *arXiv preprint arXiv:1902.01753*, 2019.

Zhang, Y. and Yang, Y. Cross-validation for selecting a model selection procedure. *Journal of Econometrics*, 187 (1):95–112, 2015.

Title in Wasserman (2006) may seem like a typo, but it is not. If you have a moment to chuckle, peek at column 2 of page 266!

Appendix

This serves as an appendix to the paper "Subsample Ridge Ensembles: Equivalences and Generalized Cross-Validation." Below we provide an outline for the appendix along with a summary of the notation used in the main paper and the appendix.

Organization

The content of the appendix is organized as follows.

- Appendix A presents proofs of results in Section 2.
  - Appendix A.1 relates the ridge estimator in the full ensemble to the M-ensemble ridge estimator as the ensemble size M tends to infinity. Specifically, we provide proof of the fact [mentioned in Section 2] that the estimator $\widetilde{\boldsymbol{\beta}}^{\lambda}_{k,\infty}$ as defined in (3) is almost surely equivalent to letting the ensemble size $M \to \infty$ for the estimator $\widetilde{\boldsymbol{\beta}}^{\lambda}_{k,M}$ as defined in (2). The main ingredients are results in Appendix G.
  - Appendix A.2 gathers known results from Patil et al. (2022a) in the form of Lemma A.2 that characterize the asymptotic prediction risks of ridge ensembles used in the remaining sections.
  - Appendix A.3 proves Theorem 2.3. The main ingredients are Lemma A.2 and results in Appendix F. See Figure 6.

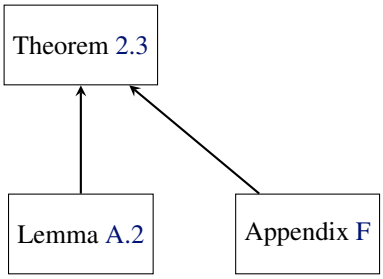

Figure 6. Schematic for the proof of Theorem 2.3.

- Appendix B presents proofs of results in Section 3.
  - Appendix B.1 proves Theorem 3.1. The main ingredients are Lemma B.1 (proved in Appendix B) and results in Appendix G. The main ingredients that prove Lemma B.1 are Proposition 3.6 (proved in Appendix E) and Lemma A.2. See Figure 7.

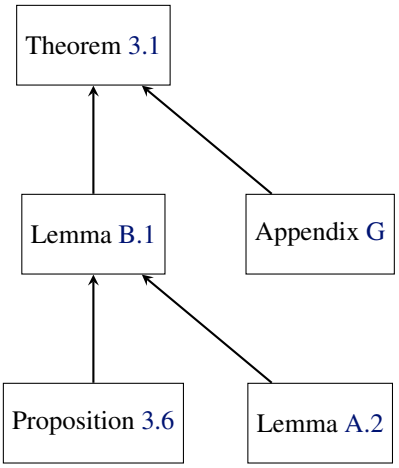

Figure 7. Schematic for the proof of Theorem 3.1.

- Appendix B.2 proves Corollary 3.2. The main ingredient in Theorem 3.1.
- Appendix B.3 proves Proposition 3.3. The main ingredients are Lemma 3.4 and results in Appendix G. See Figure 8.

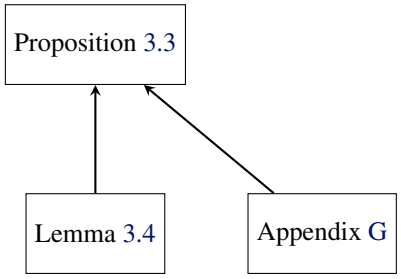

Figure 8. Schematic for the proof of Proposition 3.3.

- Appendix C proves Lemma 3.4. The main ingredient is Lemma C.1 (proved in Appendix C). See Figure 9.

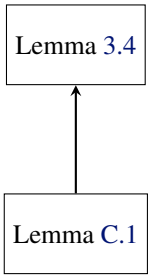

Figure 9. Schematic for the proof of Lemma 3.4.

- Appendix D proves Lemma 3.5. The main ingredients are a series of lemmas, Lemmas D.1 to D.3 (proved in Appendices D.1 to D.3). These lemmas provide structural decompositions for the ensemble train error and obtain the limiting behaviors of the terms in the decompositions. See Figure 10.
  - Appendix D.1 proves Lemma D.1 that shows a certain decomposition of the ensemble train error into in-sample train and out-of-sample test error components.
  - Appendix D.2 proves Lemma D.2 on the convergence of out-of-sample test error components. The main ingredient is Lemma A.2.
  - Appendix D.3 proves Lemma D.3 on the convergence of in-sample train error components. The main ingredients are Lemmas D.4 and D.5 (proved in Appendix D.4) and Lemmas D.6 and D.7 (proved in Appendix D.5).
  - Appendix D.4 proves Lemmas D.4 and D.5 on component concentrations of certain cross and variance terms arising in the aforementioned decompositions. The main ingredients are results in Appendix G.
  - Appendix D.5 proves Lemmas D.6 and D.7 on component deterministic approximations for the concentrated bias and variance functionals in the steps above. The main ingredients are results in Appendix F.
- Appendix E proves Proposition 3.6. The main ingredients are components proved in Lemma 3.4 and Lemma 3.5, Propositions E.1 and E.2 that handle certain boundary cases not covered by Lemma 3.4 and Lemma 3.5 (proved in Appendices E.1 and E.2), and results in Appendix G. See Figure 11.
  - Appendix E.1 proves Proposition E.1 that considers the boundary case as the subsampling ratio $\phi_s \to \infty$ for ridge regression ($\lambda > 0$).
  - Proposition E.2 proves Proposition E.2 that handles the limiting case of ridgeless regression ($\lambda = 0$), by justifying and taking a suitable limit as $\lambda \to 0^+$ of the corresponding results for ridge regression.
- Appendix F summarizes auxiliary asymptotic equivalency results used in the proofs throughout.

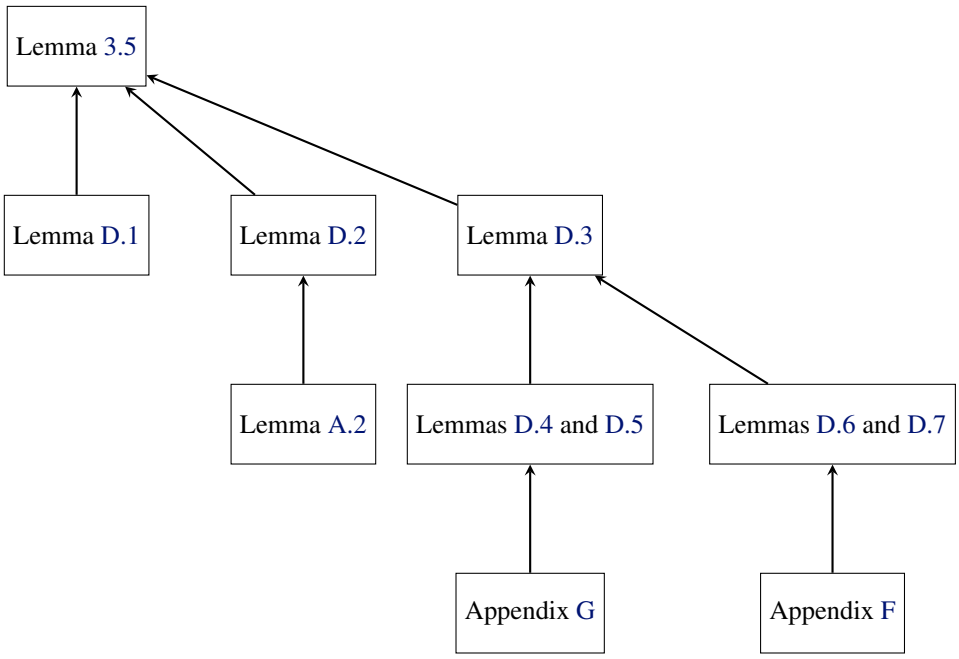

Figure 10. Schematic for the proof of Lemma 3.5.

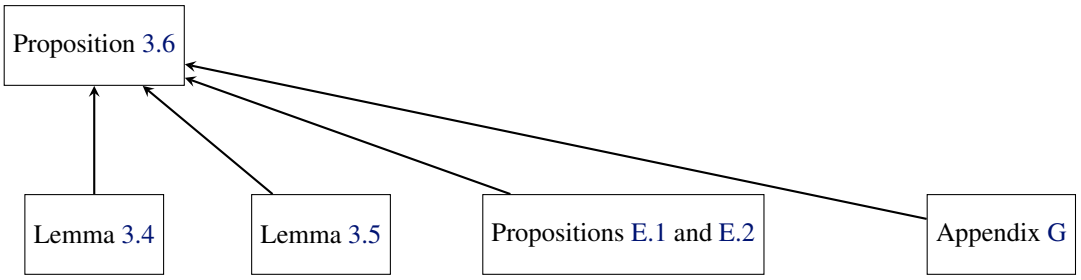

Figure 11. Schematic for the proof of Proposition 3.6.

  - Appendix F.1 provides background on the notion of asymptotic matrix equivalents and various calculus rules that such notion of equivalency obeys.
  - Appendix F.2 gathers some known asymptotic matrix equivalents and derives some novel asymptotic matrix equivalents that arise in our analysis.
  - Appendix F.3 gathers various known analytic properties of certain fixed-point equations and proves some additional properties that arise in our analysis.

- Appendix G collects several helper concentration results used in the proofs throughout.

  - Appendix G.1 provides lemmas deriving the asymptotic proportion of shared observations when subsampling.
  - Appendix G.2 provides lemmas establishing concentrations for linear and quadratic forms of random vectors.
  - Appendix G.3 provides lemmas for lifting original convergences to converges of Cesàro-type mean and max for triangular arrays.

- Appendix H discusses the bias correction of GCV for finite ensembles [mentioned in Section 5].

- Appendix I describes additional numerical details for experiments [mentioned in Section 3].

Notation

An overview of some general notation used in the main paper and the appendix is as follows.

1. General notation: We denote scalars in non-bold lower or upper case (e.g., n, λ, C), vectors in bold lower case (e.g., $\boldsymbol{x}$, $\boldsymbol{\beta}$), and matrices in bold upper case (e.g., $\boldsymbol{X}$). For a real number x, $(x)_+$ denotes its positive part, $\lfloor x \rfloor$ its floor, and $\lceil x \rceil$ its ceiling. For a vector $\boldsymbol{\beta}$, $\|\boldsymbol{\beta}\|_2$ denotes its ℓ_2 norm. For a pair of vectors $\boldsymbol{v}$ and $\boldsymbol{w}$, $\langle \boldsymbol{v}, \boldsymbol{w} \rangle$ denotes their inner product. For an event A, $\mathbb{1}_A$ denotes the associated indicator random variable. We denote convergence in probability by "$\xrightarrow{\text{p}}$", almost sure convergence by "$\xrightarrow{\text{a.s.}}$", and convergence in distribution by "$\xrightarrow{\text{d}}$".

2. Set notation: We denote sets using calligraphic letters (e.g., $\mathcal{D}$), and use blackboard letters to denote some special sets: $\mathbb{N}$ denotes the set of positive integers, $\mathbb{R}$ denotes the set of real numbers, $\mathbb{R}_{\geq 0}$ denotes the set of non-negative real numbers, $\mathbb{R}_{>0}$ denotes the set of positive real numbers, $\mathbb{C}$ denotes the set of complex numbers, $\mathbb{C}^+$ denotes the set of complex numbers with positive imaginary part, and $\mathbb{C}^-$ denotes the set of complex numbers with negative imaginary part. For a natural number n, we use $[n]$ to denote the set $\{1, \ldots, n\}$.

3. Matrix notation: For a matrix $\boldsymbol{X} \in \mathbb{R}^{n \times p}$, $\boldsymbol{X}^\top \in \mathbb{R}^{p \times n}$ denotes its transpose, and $\boldsymbol{X}^+ \in \mathbb{R}^{p \times n}$ denote its Moore-Penrose inverse. For a square matrix $\boldsymbol{A} \in \mathbb{R}^{p \times p}$, $\mathrm{tr}[\boldsymbol{A}]$ denotes its trace, and $\boldsymbol{A}^{-1} \in \mathbb{R}^{p \times p}$ denotes its inverse, provided it is invertible. For a positive semidefinite matrix $\boldsymbol{\Sigma}$, $\boldsymbol{\Sigma}^{1/2}$ denotes its principal square root. A $p \times p$ identity matrix is denoted $\boldsymbol{I}_p$, or simply by $\boldsymbol{I}$, when it is clear from the context.

   For a real matrix $\boldsymbol{X}$, its operator norm (or spectral norm) with respect to ℓ_2 vector norm is denoted by $\|\boldsymbol{X}\|_{\mathrm{op}}$, and its trace norm (or nuclear norm) is denoted by $\|\boldsymbol{X}\|_{\mathrm{tr}}$ (recall that $\|\boldsymbol{X}\|_{\mathrm{tr}} = \mathrm{tr}[(\boldsymbol{X}^\top \boldsymbol{X})^{1/2}]$). For a positive semidefinite matrix $\boldsymbol{A} \in \mathbb{R}^{p \times p}$ with eigenvalue decomposition $\boldsymbol{A} = \boldsymbol{V}\boldsymbol{R}\boldsymbol{V}^{-1}$ for an orthonormal matrix $\boldsymbol{V} \in \mathbb{R}^{p \times p}$ and a diagonal matrix $\boldsymbol{R} \in \mathbb{R}^{p \times p}$ with non-negative entries, and a function $f : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$, we denote by $f(\boldsymbol{A})$ the $p \times p$ positive semidefinite matrix $\boldsymbol{V} f(\boldsymbol{R}) \boldsymbol{V}^{-1}$. Here, $f(\boldsymbol{R})$ is a $p \times p$ diagonal matrix obtained by applying the function f to each diagonal entry of $\boldsymbol{R}$.

   For symmetric matrices $\boldsymbol{A}$ and $\boldsymbol{B}$, $\boldsymbol{A} \preceq \boldsymbol{B}$ denotes the Loewner ordering. For sequences of matrices $\boldsymbol{A}_n$ and $\boldsymbol{B}_n$, $\boldsymbol{A}_n \simeq \boldsymbol{B}_n$ denotes a certain notion of asymptotic equivalence (see Appendix F).

Finally, in what follows, we will prove the results for n, k, p being a sequence of integers $\{n_m\}_{m=1}^\infty$, $\{k_m\}_{m=1}^\infty$, $\{p_m\}_{m=1}^\infty$. One can also view k and p as sequences k_n and p_n that are indexed by n. For simplicity, we drop the subscript when it is clear from the context.

A. Proofs of results in Section 2

A.1. Full-ensemble versus limiting M-ensemble

Lemma A.1 (Almost sure equivalence of full-ensemble and limiting M-ensemble)**.** *For k, n fixed, for the ensemble estimator defined in (2), it holds that*

$$\widetilde{\boldsymbol{\beta}}^\lambda_{k,M}(\mathcal{D}_n; \{I_\ell\}_{\ell=1}^M) \xrightarrow{\text{a.s.}} \mathbb{E}[\widehat{\boldsymbol{\beta}}^\lambda_k(\mathcal{D}_I) \mid \mathcal{D}_n] = \frac{1}{|\mathcal{I}_k|} \sum_{I \in \mathcal{I}_k} \widehat{\boldsymbol{\beta}}^\lambda_k(\mathcal{D}_I),$$

as $M \to \infty$.

Proof of Lemma A.1. Note that for k, n fixed, the cardinality of $\mathcal{I}_k$ is $\binom{n}{k}$. Thus, we have

$$\widetilde{\boldsymbol{\beta}}^\lambda_{k,M}(\mathcal{D}_n; \{I_\ell\}_{\ell=1}^M) = \sum_{I \in \mathcal{I}_k} \frac{n_{M,I}}{M} \widehat{\boldsymbol{\beta}}^\lambda_k(\mathcal{D}_I)$$

for random variables $n_{M,i}$'s. Since when sampling with replacement $n_{M,I} \sim \text{Binomial}(M, 1/\binom{n}{k})$ with mean $M/\binom{n}{k}$, from the strong law of large numbers, we have that as $M \to \infty$,

$$\frac{n_{M,I}}{M} \xrightarrow{\text{a.s.}} \frac{1}{\binom{n}{k}}, \qquad \forall\, I \in \mathcal{I}_k. \tag{17}$$

For sampling without replacement, $n_{M,I} \sim \text{Hypergeometric}(M, 1, \binom{n}{k})$ for $M \leq \binom{n}{k}$ (see Definition G.1) with mean $M/\binom{n}{k}$. When $M = \binom{n}{k}$, $n_{M,I}/M = 1/\binom{n}{k}$. In both cases, we have

$$\widetilde{\boldsymbol{\beta}}^{\lambda}_{k,M}(\mathcal{D}_n; \{I_\ell\}_{\ell=1}^{\infty}) := \lim_{M\to\infty} \widetilde{\boldsymbol{\beta}}^{\lambda}_{k,M}(\mathcal{D}_n; \{I_\ell\}_{\ell=1}^{M}) \overset{\text{a.s.}}{=} \frac{1}{\binom{n}{k}} \sum_{I\in\mathcal{I}_k} \widehat{\boldsymbol{\beta}}^{\lambda}_{k}(\mathcal{D}_I),$$

which concludes the proof. □

A.2. Risk characterization of ridge ensembles

In the study of ridge ensembles under proportional asymptotics, a key quantity that appears is the solution of a fixed-point equation. For any $\lambda > 0$ and $\theta > 0$, define $v(-\lambda;\theta)$ as the unique nonnegative solution to the fixed-point equation

$$v(-\lambda;\theta)^{-1} = \lambda + \theta \int r(1 + v(-\lambda;\theta)r)^{-1}\,\mathrm{d}H(r). \tag{18}$$

For $\lambda = 0$, define $v(0;\theta) := \lim_{\lambda\to 0^+} v(-\lambda;\theta)$ for $\theta > 1$ and $+\infty$ otherwise. Such a fixed-point equation has appeared in the literature before. See, e.g., Dobriban & Wager (2018); Hastie et al. (2022); Mei & Montanari (2022) in the context of ridge regression; and more generally, for other M-estimators, see, e.g., El Karoui (2013; 2018); Thrampoulidis et al. (2015; 2018); Sur et al. (2019); Miolane & Montanari (2021), among others. The fact that the fixed-point equation (18) has a unique nonnegative solution follows from Patil et al. (2022b, Lemma S.6.14).

We then define the nonnegative constants $\widetilde{v}(-\lambda;\vartheta,\theta)$, and $\widetilde{c}(-\lambda;\theta)$ via the following equations:

$$\widetilde{v}(-\lambda;\vartheta,\theta) = \frac{\vartheta \int r^2(1 + v(-\lambda;\theta)r)^{-2}\,\mathrm{d}H(r)}{v(-\lambda;\theta)^{-2} - \vartheta \int r^2(1 + v(-\lambda;\theta)r)^{-2}\,\mathrm{d}H(r)},\ \widetilde{c}(-\lambda;\theta) = \int r(1 + v(-\lambda;\theta))r)^{-2}\,\mathrm{d}G(r). \tag{19}$$

Lemma A.2 (Risk characterization of ridge ensembles, adapted from Patil et al. (2022a))**.** *Suppose Assumptions 2.1-2.2 hold for the dataset $\mathcal{D}_n$. Then, as $k, n, p \to \infty$ such that $p/n \to \phi \in (0,\infty)$ and $p/k \to \phi_s \in [\phi, \infty]$ (and $\phi_s \neq 1$ if $\lambda = 0$), there exist deterministic functions $\mathscr{R}^{\lambda}_{M}(\phi,\phi_s)$ for $M \in \mathbb{N}$, such that for $I_1, \ldots, I_M \overset{\text{SRS}}{\sim} \mathcal{I}_k$,*

$$\sup_{M\in\mathbb{N}} |R(\widetilde{f}_{M,\mathcal{I}_k}; \mathcal{D}_n, \{I_\ell\}_{\ell=1}^{M}) - \mathscr{R}^{\lambda}_{M}(\phi,\phi_s)| \xrightarrow{\text{p}} 0. \tag{20}$$

Furthermore, the function $\mathscr{R}^{\lambda}_{M}(\phi,\phi_s)$ decomposes as

$$\mathscr{R}^{\lambda}_{M}(\phi,\phi_s) = \sigma^2 + \mathscr{B}^{\lambda}_{M}(\phi,\phi_s) + \mathscr{V}^{\lambda}_{M}(\phi,\phi_s), \tag{21}$$

where the bias and variance terms are given by

$$\mathscr{B}^{\lambda}_{M}(\phi,\phi_s) = M^{-1}B_\lambda(\phi_s,\phi_s) + (1 - M^{-1})B_\lambda(\phi,\phi_s), \tag{22}$$

$$\mathscr{V}^{\lambda}_{M}(\phi,\phi_s) = M^{-1}V_\lambda(\phi_s,\phi_s) + (1 - M^{-1})V_\lambda(\phi,\phi_s), \tag{23}$$

and the functions $B_\lambda(\cdot,\cdot)$ and $V_\lambda(\cdot,\cdot)$ are defined as

$$B_\lambda(\vartheta,\theta) = \rho^2(1 + \widetilde{v}(-\lambda;\vartheta,\theta))\widetilde{c}(-\lambda;\theta), \qquad V_\lambda(\vartheta,\theta) = \sigma^2\widetilde{v}(-\lambda;\vartheta,\theta), \qquad \theta \in (0,\infty],\ \vartheta \leq \theta. \tag{24}$$

A.3. Proof of Theorem 2.3

Proof of Theorem 2.3. Define $\phi_s^*(\phi) := \operatorname{argmin}_{\phi_s\geq\phi} \mathscr{R}_{\lambda,\infty}(\phi,\phi_s)$ and $\lambda^*(\phi) := \operatorname{argmin}_{\lambda\geq 0} \mathscr{R}^{\lambda}_{1}(\phi,\phi)$. We will write ϕ_s^* and λ^* for simplicity and split the proof into different cases.

Part (1) Case of $\texttt{SNR} > 0$ ($\rho^2 > 0, \sigma^2 > 0$):

From Patil et al. (2022a, Proposition 5.7) we have that $\phi_s^* \in (\phi \vee 1, \infty)$. From Lemma F.11 (1), the function $\phi_s \mapsto v(0;\phi_s)$ is strictly decreasing over $\phi_s \in [1,\infty]$ with range

$$v(0;\phi_s \vee 1) = \begin{cases} v(0;\phi_s), & \phi \in (1,\infty) \\ \lim_{\phi_s\to 1^+} v(0;\phi_s) = +\infty, & \phi \in (0,1] \end{cases}, \qquad v(0;+\infty) := \lim_{\phi_s\to+\infty} v(0;\phi_s) = 0.$$

From Lemma F.12 (1), the function $\lambda \mapsto v(-\lambda;\phi)$ is strictly decreasing over $\lambda \in [0,\infty]$ with range

$$v(0;\phi \vee 1) = \begin{cases} v(0;\phi), & \phi \in (1,\infty) \\ \lim_{\lambda\to 0^+} v(-\lambda;\phi) = +\infty, & \phi \in (0,1] \end{cases}, \qquad v(-\infty;\phi) := \lim_{\lambda\to+\infty} v(-\lambda;\phi) = 0.$$

By the intermediate value theorem, there exists unique $\lambda_0 \in (0,\infty)$ such that $v(-\lambda_0;\phi) = v(0;\phi_s^*)$. Then we also have $\widetilde{c}(-\lambda_0;\phi) = \widetilde{c}(0;\phi_s^*)$ and $\widetilde{v}(-\lambda_0;\phi,\phi) = \widetilde{v}(0;\phi_s^*)$. Substituting this into the optimal ensemble risk, we have

$$\begin{aligned}
\min_{\phi_s\geq\phi} \mathscr{R}_\infty^0(\phi,\phi_s) &= \mathscr{R}_\infty^0(\phi,\phi_s^*) \\
&= (\sigma^2 + \rho^2\widetilde{c}(0;\phi_s^*))(1+\widetilde{v}(0;\phi,\phi_s^*)) \\
&= (\sigma^2 + \rho^2\widetilde{c}(-\lambda_0;\phi))(1+\widetilde{v}(-\lambda_0;\phi,\phi)) \\
&= \mathscr{R}_1^{\lambda_0}(\phi,\phi) \\
&\leq \min_{\lambda\geq 0} \mathscr{R}_1^{\lambda}(\phi,\phi).
\end{aligned}$$

On the other hand, there exists unique $\phi_0 \in [1,\infty)$ such that $v(-\lambda^*;\phi) = v(0;\phi_0)$, and thus, we have

$$\begin{aligned}
\min_{\lambda\geq 0} \mathscr{R}_1^{\lambda}(\phi,\phi) &= \mathscr{R}_1^{\lambda^*}(\phi,\phi) \\
&= (\sigma^2 + \rho^2\widetilde{c}(-\lambda^*;\phi))(1+\widetilde{v}(-\lambda^*;\phi,\phi)) \\
&= (\sigma^2 + \rho^2\widetilde{c}(0;\phi_0))(1+\widetilde{v}(0;\phi,\phi_0)) \\
&= \mathscr{R}_\infty^0(\phi,\phi_0) \\
&\leq \min_{\phi_s\geq\phi} \mathscr{R}_\infty^0(\phi,\phi_s).
\end{aligned}$$

Combining the above two inequalities, we have that $\min_{\phi_s\geq\phi} \mathscr{R}_\infty^0(\phi,\phi_s) = \min_{\lambda\geq 0} \mathscr{R}_1^{\lambda}(\phi,\phi)$.

Part (2) Case of $\mathtt{SNR} = 0$ ($\rho^2 = 0, \sigma^2 > 0$):

From Patil et al. (2022a, Proposition 5.7) we have that $\phi_s^* = +\infty$, which implies that $v(0;\phi_s^*) = 0$. Then, from Lemma F.11 (1) we have $v(0;+\infty) := \lim_{\phi_s\to+\infty} v(0;\phi_s) = 0$,

$$\begin{aligned}
\lim_{\phi_s\to+\infty} \widetilde{v}(0;\phi,+\infty) &= \lim_{\phi_s\to+\infty} \frac{\phi\int r^2(1+v(0;\phi_s)r)^{-2}\,\mathrm{d}H(r)}{v(0;\phi_s)^{-2} - \phi\int r^2(1+v(0;\phi_s)r)^{-2}\,\mathrm{d}H(r)} \\
&= \lim_{\phi_s\to+\infty} \frac{\phi\int (v(0;\phi_s)r)^2(1+v(0;\phi_s)r)^{-2}\,\mathrm{d}H(r)}{1 - \phi\int (v(0;\phi_s)r)^2(1+v(0;\phi_s)r)^{-2}\,\mathrm{d}H(r)} \\
&= \frac{\phi\int (v(0;+\infty)r)^2(1+v(0;+\infty)r)^{-2}\,\mathrm{d}H(r)}{1 - \phi\int (v(0;+\infty)r)^2(1+v(0;+\infty)r)^{-2}\,\mathrm{d}H(r)} \\
&= 0,
\end{aligned}$$

and thus,

$$\min_{\phi_s\geq\phi} \mathscr{R}_\infty^0(\phi,\phi_s) = \mathscr{R}_\infty^0(\phi,\infty) = \sigma^2(1+\widetilde{v}(0;\phi,+\infty)) = \sigma^2.$$

On the other hand,

$$\min_{\lambda\geq 0} \mathscr{R}_1^{\lambda}(\phi,\phi) = \mathscr{R}_1^{\lambda^*}(\phi,\phi) = \sigma^2\widetilde{v}(-\lambda^*;\phi,\phi) \geq \sigma^2$$

where the equality holds when $\lambda^* = +\infty$ because $\widetilde{v}(-\lambda^*;\phi,\phi) \geq 0$ from Lemma F.10 (4). Thus, the optimal parameters to the two optimization problems are given by $\phi_s^* = \lambda^* = +\infty$, with $v(0;\phi_s^*) = v(-\lambda^*;\phi) = 0$.

Part (3) Case of $\mathtt{SNR} = \infty$ ($\rho^2 > 0, \sigma^2 = 0$):

When $\phi \le 1$, from Patil et al. (2022a, Proposition 5.7) we have that any $\phi_s^* \in [\phi, 1]$ minimizes $\min_{\phi_s \ge \phi} \mathscr{R}_\infty^0(\phi, \phi_s)$ and the minimum is 0, which is also the smallest possible prediction risk. As $\mathscr{R}_1^\lambda(\phi, \phi) = 0$ for $\lambda = 0$, the conclusion still holds.

When $\phi \in (1, \infty)$, we know that $\phi_s^* \in (1, \infty)$ from Patil et al. (2022a, Proposition 5.7). Analogous to Part (1), we have that $\min_{\phi_s \ge \phi} \mathscr{R}_\infty^0(\phi, \phi_s) = \min_{\lambda \ge 0} \mathscr{R}_1^\lambda(\phi, \phi)$.

Part (4) Relationship between ϕ^* and λ^*:

Each pair of the optimal solution (ϕ^*, λ^*) satisfies that $v(0; \phi_s^*) = v(-\lambda^*; \phi) =: v^*$, where $v(0; \phi_s^*)$ and $v(-\lambda^*; \phi)$ are non-negative solutions to the following fixed-point equations:

$$\frac{1}{v(0; \phi_s^*)} = \phi_s^* \int \frac{r}{1 + v(0; \phi_s^*) r} \, \mathrm{d}H(r), \qquad \frac{1}{v(-\lambda^*; \phi)} = \lambda^* + \phi \int \frac{r}{1 + v(-\lambda^*; \phi) r} \, \mathrm{d}H(r)$$

From the previous parts, if $\mathtt{SNR} = 0$, then $\lambda^* = \phi_s^* = +\infty$ and $v^* = 0$. Otherwise, we have

$$\frac{1}{v^*} = \phi_s^* \int \frac{r}{1 + v^* r} \, \mathrm{d}H(r) = \lambda^* + \phi \int \frac{r}{1 + v^* r} \, \mathrm{d}H(r),$$

which yields that

$$\lambda^* = (\phi_s^* - \phi) \int \frac{r}{1 + v^* r} \, \mathrm{d}H(r).$$

Part (5) Individual and joint optimization:

Note that from Lemma F.10 (2) and Lemma F.11 (1), the function $\phi_s \mapsto v(-\lambda; \phi_s)$ is decreasing with the range $[0, \lambda^{-1}]$ for $\lambda \in [0, \infty]$. Then the function $(\lambda, \phi_s) \mapsto v(-\lambda; \phi_s)$ has the range $[0, +\infty]$, which is the same as $v(0; \phi_s)$. It follows that $\min_{\phi_s \ge \phi} \mathscr{R}_\infty^0(\phi, \phi_s) = \min_{\phi_s \ge \phi, \lambda \ge 0} \mathscr{R}_1^\lambda(\phi, \phi)$ by the analogous argument in Part (1)-(3).

When $\lambda^* = 0$, the curve reduces to a singleton, which is a trivial case. When $\lambda^* > 0$, for any $t \in [0, \lambda^*]$, let $\lambda = \lambda^* - t$ and $\phi_s = \phi + t(\phi_s^* - \phi)/\lambda^*$. Note that

$$\begin{aligned}
\frac{1}{v(\lambda; \phi_s)} &= \lambda + \phi_s \int \frac{r}{1 + v^* r} \, \mathrm{d}H(r) \\
&= \lambda^* - t + (\phi + t(\phi_s^* - \phi)/\lambda^*) \int \frac{r}{1 + v(0; \phi_s^*) r} \, \mathrm{d}H(r) \\
&= \lambda^* + \phi \int \frac{r}{1 + v(0; \phi_s^*) r} \, \mathrm{d}H(r) + \frac{t}{\lambda^*} (\phi_s^* - \phi - \lambda) \int \frac{r}{1 + v(0; \phi_s^*) r} \, \mathrm{d}H(r) \\
&= \frac{1}{v^*} + \frac{t}{\lambda^*} \left(\frac{1}{v^*} - \frac{1}{v^*} \right) \\
&= \frac{1}{v^*},
\end{aligned}$$

which implies that $v(\lambda; \phi_s) = v^*$. Then, we have

$$\widetilde{c}(-\lambda; \phi_s) = \int \frac{r}{(1 + v(-\lambda; \phi_s)) r)^2} \, \mathrm{d}G(r) = \widetilde{c}(-\lambda^*; \phi) = \widetilde{c}(0; \phi_s^*).$$

and

$$\widetilde{v}(-\lambda; \phi, \phi_s) = \frac{\phi \int r^2 (1 + v(-\lambda; \phi_s) r)^{-2} \, \mathrm{d}H(r)}{v(-\lambda; \phi_s)^{-2} - \phi \int r^2 (1 + v(-\lambda; \phi_s) r)^{-2} \, \mathrm{d}H(r)} = \widetilde{v}(-\lambda^*; \phi, \phi) = \widetilde{v}(0; \phi, \phi_s^*).$$

It then follows that $\mathscr{R}_\infty^\lambda(\phi, \phi_s) = \mathscr{R}_\infty^{\lambda^*}(\phi, \phi) = \mathscr{R}_\infty^0(\phi, \phi_s^*)$, which completes the proof for Theorem 2.3.

Part (6)* Extension of risk equivalence:

Here we extend the results in Theorem 2.3 to a more general equivalence of (λ, ϕ_s), as indicated following Theorem 2.3 towards the end of Section 2.2.

For any $\bar{\phi}_s \in [\phi, +\infty]$, let $\bar{\lambda} = (\bar{\phi}_s - \phi) \int r(1 + v(0;\phi_s)r)^{-1} \, \mathrm{d}H(r) \geq 0$. Then, we have

$$\frac{1}{v(0;\bar{\phi}_s)} = \bar{\phi}_s \int \frac{r}{1 + v(0;\bar{\phi}_s)r} \, \mathrm{d}H(r) = \bar{\lambda} + \phi \int \frac{r}{1 + v(0;\bar{\phi}_s)r} \, \mathrm{d}H(r),$$

It follows that $v(-\bar{\lambda};\phi) = v(0;\bar{\phi}_s)$, and consequently, $\mathscr{R}^0_\infty(\phi, \bar{\phi}_s) = \mathscr{R}^{\bar{\lambda}}_\infty(\phi, \phi)$.

□

B. Proofs of results in Section 3

B.1. Proof of Theorem 3.1

To prove Theorem 3.1, we first prove pointwise convergence (over k and λ) as stated in Lemma B.1, which is based on Proposition 3.6 proved in Appendix E.

Lemma B.1 (Consistency of GCV in full ensemble)**.** *Under Assumptions 2.1-2.2, as $k, n, p \to \infty$, $p/n \to \phi \in (0, \infty)$ and $p/k \to \phi_s \in [\phi, +\infty]$, for $\lambda \geq 0$, it holds that*

$$|\mathrm{gcv}_k^\lambda - R^\lambda_{k,\infty}| \xrightarrow{\text{a.s.}} 0. \tag{25}$$

Proof of Lemma B.1. We will first show that proof for $\lambda > 0$ and $\phi_s < \infty$, and then extend the results to these boundary cases.

Recall that from Proposition 3.6, we have

$$\mathscr{G}^\lambda_\infty(\phi, \phi_s) = \frac{\frac{2\phi(2\phi_s - \phi)}{\phi_s^2} \mathscr{T}_2^\lambda(\phi, \phi_s) + \frac{2(\phi_s - \phi)^2}{\phi_s^2} \mathscr{R}_2^\lambda(\phi, \phi_s) - \frac{\phi}{\phi_s} \mathscr{T}_1^\lambda(\phi, \phi_s) - \frac{\phi_s - \phi}{\phi_s} \mathscr{R}_1^\lambda(\phi, \phi_s)}{\mathscr{D}^\lambda_\infty(\phi, \phi_s)}.$$

We next simplify the expression of the numerator:

$$\begin{aligned}
&\frac{2\phi(2\phi_s - \phi)}{\phi_s^2} \mathscr{T}_2^\lambda(\phi, \phi_s) + \frac{2(\phi_s - \phi)^2}{\phi_s^2} \mathscr{R}_2^\lambda(\phi, \phi_s) - \frac{\phi}{\phi_s} \mathscr{T}_1^\lambda(\phi, \phi_s) - \frac{\phi_s - \phi}{\phi_s} \mathscr{R}_1^\lambda(\phi, \phi_s) \\
&= \frac{\phi(\phi_s - \phi)}{\phi_s^2} \mathscr{R}_1^\lambda(\phi, \phi_s) + \mathscr{D}^\lambda(\phi_s, \phi_s) \left(\frac{\phi}{\phi_s} \mathscr{R}_1^\lambda(\phi, \phi_s) + \left(\frac{2\phi(\phi_s - \phi)}{\phi_s^2} \frac{1}{\lambda v(-\lambda;\phi_s)} + \frac{\phi^2}{\phi_s^2} \right) \mathscr{R}^\lambda_\infty(\phi, \phi_s) \right) \\
&\quad + \frac{2(\phi_s - \phi)^2}{\phi_s^2} \mathscr{R}_2^\lambda(\phi, \phi_s) - \frac{\phi}{\phi_s} \mathscr{D}^\lambda(\phi_s, \phi_s) \mathscr{R}_1^\lambda(\phi, \phi_s) - \frac{\phi_s - \phi}{\phi_s} \mathscr{R}_1^\lambda(\phi, \phi_s) \\
&= \left(\frac{2\phi(\phi_s - \phi)}{\phi_s^2} \lambda v(-\lambda;\phi_s) + \frac{\phi^2}{\phi_s^2} \lambda^2 v(-\lambda;\phi_s)^2 \right) \mathscr{R}^\lambda_\infty(\phi, \phi_s) + \frac{2(\phi_s - \phi)^2}{\phi_s^2} (\mathscr{R}_2^\lambda(\phi, \phi_s) - \mathscr{R}_1^\lambda(\phi, \phi_s)) \\
&= \left(\frac{(\phi_s - \phi)^2}{\phi_s^2} + \frac{2\phi(\phi_s - \phi)}{\phi_s^2} \lambda v(-\lambda;\phi_s) + \frac{\phi^2}{\phi_s^2} \lambda^2 v(-\lambda;\phi_s)^2 \right) \mathscr{R}^\lambda_\infty(\phi, \phi_s) \\
&= \mathscr{D}^\lambda_\infty(\phi, \phi_s) \mathscr{R}^\lambda_\infty(\phi, \phi_s).
\end{aligned}$$

Then, it follows that $\mathscr{G}^\lambda_\infty(\phi, \phi_s) = \mathscr{R}^\lambda_\infty(\phi, \phi_s)$. From Lemma A.2 and Proposition 3.6, we have that $\mathrm{gcv}_k^\lambda \xrightarrow{\text{a.s.}} \mathscr{G}^\lambda_\infty(\phi, \phi_s)$ and $R^\lambda_{k,\infty} \xrightarrow{\text{a.s.}} \mathscr{R}^\lambda_\infty(\phi, \phi_s)$, which finishes the proof. □

We are now ready to prove Theorem 3.1.

Proof of Theorem 3.1. Let $R_{n,k} = \mathrm{gcv}^\lambda_{k,\infty} - R^\lambda_{k,\infty}$ for $n \in \mathbb{N}$ and $k \in \mathcal{K}_n$. From Lemma B.1 we have that $R_{n,k} \xrightarrow{\text{a.s.}} 0$ as $k, n, p \to \infty$, $p/n \to \phi \in (0, \infty)$ and $p/k \to \phi_s \in [\phi, \infty]$. Here we view k and p as k_n and p_n that are indexed by n. Then from Lemma G.5 (1) the conclusion follows. □

B.2. Proof of Corollary 3.2

Proof of Corollary 3.2. From Theorem 3.1, we have

$$\max_{k\in\mathcal{K}_n} |\mathrm{gcv}^{\lambda}_{k,\infty} - R^{\lambda}_{k,\infty}| \xrightarrow{\text{a.s.}} 0.$$

This implies that

$$\begin{aligned}
\min_{k\in\mathcal{K}_n} \mathrm{gcv}^0_k &= \min_{k\in\mathcal{K}_n} \mathscr{R}^0_\infty(p/n, p/k)\left(1 + \frac{\mathrm{gcv}^0_k - \mathscr{R}^0_\infty(p/n,p/k)}{\mathscr{R}^0_\infty(p/n,p/k)}\right) \\
&\leqslant \min_{k\in\mathcal{K}_n} \mathscr{R}^0_\infty(p/n, p/k)\left(1 \pm \max_{k\in\mathcal{K}_n}\left|\frac{\mathrm{gcv}^0_k - \mathscr{R}^0_\infty(p/n,p/k)}{\mathscr{R}^0_\infty(p/n,p/k)}\right|\right) \\
&\leqslant \min_{k\in\mathcal{K}_n} \mathscr{R}^0_\infty(p/n, p/k)\left(1 \pm \frac{1}{\sigma^2}\max_{k\in\mathcal{K}_n} |\mathrm{gcv}^0_k - \mathscr{R}^0_\infty(p/n,p/k)|\right) \\
&\xrightarrow{\text{a.s.}} \inf_{\phi_s\geq\phi} \mathscr{R}^0_\infty(\phi,\phi_s) \\
&= \inf_{\phi_s\geq\phi,\lambda\geq 0} \mathscr{R}^{\lambda}_\infty(\phi,\phi_s),
\end{aligned}$$

where the last equality is from Theorem 2.3. This finishes the proof. □

B.3. Proof of Proposition 3.3

Proof of Proposition 3.3. From the proof of Lemma 3.4, we have

$$\frac{1}{k}\operatorname{tr}(\boldsymbol{M}_m\boldsymbol{\Sigma}_m) \xrightarrow{\text{a.s.}} (1 - \lambda v(-\lambda;\phi_s)).$$

Then, as $k, n, p \to \infty$, $p/n \to \phi$ and $p/k \to \phi_s$, we have

$$\begin{aligned}
D^{\lambda}_{k,2} &= \left(1 - \frac{1}{|I_1\cup I_2|}\frac{1}{2}\sum_{m=1}^{2}\operatorname{tr}(\boldsymbol{M}_m\boldsymbol{\Sigma}_m)\right)^2 \\
&= \left(1 - \frac{k}{|I_1\cup I_2|}\frac{1}{k}\frac{1}{2}\sum_{m=1}^{2}\operatorname{tr}(\boldsymbol{M}_m\boldsymbol{\Sigma}_m)\right)^2 \\
&\xrightarrow{\text{a.s.}} \left(1 - \frac{\phi_s}{2\phi_s - \phi}(1 - \lambda v(-\lambda;\phi_s))\right)^2 =: \mathscr{D}^{\lambda}_2(\phi,\phi_s).
\end{aligned}$$

where the convergence of $k/|I_1 \cup I_2|$ is from Lemma G.2. It then follows that

$$\mathrm{gcv}^{\lambda}_{k,2} = \frac{T^{\lambda}_{k,2}}{D^{\lambda}_{k,2}} \xrightarrow{\text{a.s.}} \mathscr{G}^{\lambda}_2(\phi,\phi_s) := \frac{\mathscr{T}^{\lambda}_2}{\mathscr{D}^{\lambda}_2},$$

where $\mathscr{T}^{\lambda}_2$ defined in (32) has the following expression:

$$\begin{aligned}
\mathscr{T}^{\lambda}_2(\phi,\phi_s) &= \frac{1}{2}\frac{\phi_s - \phi}{2\phi_s - \phi}\mathscr{R}^{\lambda}_1(\phi,\phi_s) + \frac{1}{2}\mathscr{D}^{\lambda}(\phi_s,\phi_s) \\
&\left(\frac{\phi_s}{2\phi_s - \phi}\mathscr{R}^{\lambda}_1(\phi,\phi_s) + \left(\frac{2(\phi_s - \phi)}{2\phi_s - \phi}\ \frac{1}{\lambda v(-\lambda;\phi_s)} + \ \frac{\phi}{2\phi_s - \phi}\right)\mathscr{R}^{\lambda}_\infty(\phi,\phi_s)\right).
\end{aligned}$$

On the other hand, we have

$$\mathscr{R}^{\lambda}_2(\phi,\phi_s) = \frac{1}{2}\mathscr{R}^{\lambda}_1(\phi,\phi_s) + \frac{1}{2}\mathscr{R}^{\lambda}_\infty(\phi,\phi_s).$$

Note that $\mathscr{R}_1^\lambda(\phi,\phi_s) > \mathscr{R}_\infty^\lambda(\phi,\phi_s) > \sigma^2$ when $\phi < \phi_s < \infty$ because $\rho^2 > 0$. When $\lambda = 0$ and $\phi_s > 1$, we have $\lambda v(-\lambda;\phi_s) = 0$ and

$$\mathscr{G}_2^0(\phi,\phi_s) = \frac{\frac{\phi_s-\phi}{2\phi_s-\phi}\mathscr{R}_1^\lambda(\phi,\phi_s) + \frac{2(\phi_s-\phi)}{2\phi_s-\phi}\mathscr{R}_\infty^\lambda(\phi,\phi_s)}{2\left(1-\frac{\phi_s}{2\phi_s-\phi}\right)^2} = \frac{1}{2}\cdot\frac{2\phi_s-\phi}{\phi_s-\phi}\left(\mathscr{R}_1^\lambda(\phi,\phi_s) + 2\mathscr{R}_\infty^\lambda(\phi,\phi_s)\right).$$

It follows that

$$\mathscr{G}_2^0(\phi,\phi_s) - \mathscr{R}_2^\lambda(\phi,\phi_s) = \frac{1}{2(\phi_s-\phi)}\left(\phi_s\mathscr{R}_1^\lambda(\phi,\phi_s) + (3\phi_s-\phi)\mathscr{R}_\infty^\lambda(\phi,\phi_s)\right) =: c,$$

and $\mathrm{gcv}_{k,2}^\lambda - \mathscr{R}_2^\lambda(\phi,\phi_s) \xrightarrow{\mathrm{p}} c > 0$, which completes the proof. □

C. Proof of Lemma 3.4 (convergence of the GCV denominator functional)

Proof of Lemma 3.4. By the definition, the smooth matrix for $M = \infty$ is given by

$$\boldsymbol{S}_{\lambda,\infty}(\{\mathcal{D}_{I_\ell}\}_{\ell=1}^M) := \lim_{M\to\infty}\frac{1}{M}\sum_{\ell=1}^M \boldsymbol{S}_\lambda(\mathcal{D}_{I_\ell}).$$

For the denominator, note that for any fixed $n \in \mathbb{N}$, as $M \to \infty$, $I_{1:M} \xrightarrow{\text{a.s.}} [n]$. Then from Lemma C.1 (stated and proved below),

$$\frac{1}{|I_{1:M}|}\operatorname{tr}(\boldsymbol{S}_\lambda(\mathcal{D}_{I_\ell})) \xrightarrow{\text{a.s.}} \frac{1}{n}\operatorname{tr}\left(\boldsymbol{X}\boldsymbol{M}_\ell\frac{\boldsymbol{X}^\top\boldsymbol{L}_\ell}{k}\right) = \frac{1}{n}\operatorname{tr}\left(\boldsymbol{M}_\ell\boldsymbol{\Sigma}_\ell\right) \xrightarrow{\text{a.s.}} \frac{\phi}{\phi_s}(1-\lambda v(-\lambda;\phi_s)).$$

By continuous mapping theorem, we have

$$\left(1-\frac{1}{n}\operatorname{tr}(\boldsymbol{S}_{\lambda,\infty})\right)^2 \xrightarrow{\text{a.s.}} \mathscr{D}_\infty^\lambda(\phi,\phi_s) := \left(\frac{\phi_s-\phi}{\phi_s} + \frac{\phi}{\phi_s}\lambda v(-\lambda;\phi_s)\right)^2. \tag{26}$$

□

Lemma C.1 (Deterministic approximation of the denominator functional)**.** *Under Assumption 2.1, for all $m \in [M]$ and $I_m \in \mathcal{I}_k$, let $\widehat{\boldsymbol{\Sigma}}_m = \boldsymbol{X}^\top\boldsymbol{L}_m\boldsymbol{X}/k$, $\boldsymbol{L}_m \in \mathbb{R}^{n\times n}$ be a diagonal matrix with $(\boldsymbol{L}_m)_{ll} = 1$ if $l \in I_m$ and 0 otherwise, and $\boldsymbol{M}_m = (\boldsymbol{X}^\top\boldsymbol{L}_m\boldsymbol{X}/k + \lambda\boldsymbol{I}_p)^{-1}$. Then, it holds that for all $m \in [M]$ and $I_m \in \mathcal{I}_k$,*

$$\frac{1}{n}\operatorname{tr}\left(\boldsymbol{M}_m\widehat{\boldsymbol{\Sigma}}_m\right) \xrightarrow{\text{a.s.}} \frac{\phi}{\phi_s}(1-\lambda v(-\lambda;\phi_s)),$$

as $n,k,p \to \infty$, $p/n \to \phi \in (0,\infty)$, and $p/k \to \phi_s \in [\phi,\infty)$, where the nonnegative constant $\widetilde{v}(\lambda;\phi,\phi_s)$ is as defined in (19).

Proof of Lemma C.1. Note that $\boldsymbol{M}_m\widehat{\boldsymbol{\Sigma}}_m = \boldsymbol{I}_p - \lambda\boldsymbol{M}_m$. From Corollary F.5, we have that $\lambda\boldsymbol{M}_m \simeq (v(-\lambda;\phi_s)\boldsymbol{\Sigma} + \boldsymbol{I}_p)^{-1}$. Then by Lemma F.3 (4), it follows that

$$\begin{aligned}
\frac{1}{n}\operatorname{tr}\left(\boldsymbol{M}_m\boldsymbol{\Sigma}_m\right) &\xrightarrow{\text{a.s.}} \phi\lim_{p\to\infty}\frac{1}{p}\operatorname{tr}\left(\boldsymbol{I}_p - (v(-\lambda;\phi_s)\boldsymbol{\Sigma}+\boldsymbol{I}_p)^{-1}\right)\\
&= \phi\lim_{p\to\infty}\left(1-\int\frac{1}{1+v(-\lambda;\phi_s)r}\,\mathrm{d}H_p(r)\right)\\
&= \phi\int\frac{v(-\lambda;\phi_s)r}{1+v(-\lambda;\phi_s)r}\,\mathrm{d}H(r)\\
&= \frac{\phi}{\phi_s}(1-\lambda v(-\lambda;\phi_s))
\end{aligned}$$

where in the second last line, we used the fact that H_p and H have compact supports and Assumption 2.2 and the last equality is due to the definition of $v(-\lambda;\phi_s)$ in (18). □

D. Proof of Lemma 3.5 (convergence of the GCV numerator functional)

Proof of Lemma 3.5. For any $m \in [M]$, let I_m be a sample from $\mathcal{I}_k$, and $\boldsymbol{L}_{I_m} \in \mathbb{R}^{n\times n}$ be a diagonal matrix with $(\boldsymbol{L}_{I_m})_{ll} = 1$ if $l \in I_m$ and 0 otherwise. The ingredient estimator takes the form:

$$\begin{aligned}
\widetilde{\boldsymbol{\beta}}^{\lambda}_{k,M}(\{\mathcal{D}_{I_\ell}\}_{\ell=1}^{M}) &= \frac{1}{M}\sum_{m=1}^{M}\widehat{\boldsymbol{\beta}}^{\lambda}(\mathcal{D}_{I_m}) \\
&= \frac{1}{M}\sum_{m=1}^{M}(\boldsymbol{X}^\top \boldsymbol{L}_{I_m}\boldsymbol{X}/k + \lambda \boldsymbol{I}_p)^{-1}(\boldsymbol{X}^\top \boldsymbol{L}_{I_m}\boldsymbol{y}/k) \\
&= \frac{1}{M}\sum_{m=1}^{M}\left[\left(\frac{\boldsymbol{X}^\top \boldsymbol{L}_{I_m}\boldsymbol{X}}{k} + \lambda \boldsymbol{I}_p\right)^{-1}\frac{\boldsymbol{X}^\top \boldsymbol{L}_{I_m}}{k}\boldsymbol{\beta}_0 + \left(\frac{\boldsymbol{X}^\top \boldsymbol{L}_{I_m}\boldsymbol{X}}{k} + \lambda \boldsymbol{I}_p\right)^{-1}\frac{\boldsymbol{X}^\top \boldsymbol{L}_{I_m}}{k}\boldsymbol{\epsilon}\right].
\end{aligned}$$

We will write $\widetilde{\boldsymbol{\beta}}_{\lambda,M} = \widetilde{\boldsymbol{\beta}}^{\lambda}_{k,M}$ and $\boldsymbol{L}_m = \boldsymbol{L}_{I_m}$ for simplicity when they are clear from the context. The set operation will be propagated to such notations, e.g., $\boldsymbol{L}_{m\cup l} = \boldsymbol{L}_{I_m\cup I_l}$, $\boldsymbol{L}_{m\cap l} = \boldsymbol{L}_{I_m\cap I_l}$, $\boldsymbol{L}_{m\setminus l} = \boldsymbol{L}_{I_m\setminus I_l}$, etc. Let $\boldsymbol{M}_m = (\boldsymbol{X}^\top \boldsymbol{L}_{I_m}\boldsymbol{X}/k + \lambda \boldsymbol{I}_p)^{-1}$ for $m \in [M]$, we have

$$\widetilde{\boldsymbol{\beta}}_{\lambda,M} = \frac{1}{M}\sum_{m=1}^{M}(\boldsymbol{I}_p - \lambda \boldsymbol{M}_m)\boldsymbol{\beta}_0 + \frac{1}{M}\sum_{m=1}^{M}\boldsymbol{M}_m(\boldsymbol{X}^\top \boldsymbol{L}_m/k)\boldsymbol{\epsilon}. \tag{27}$$

The proof follows by combing the squared error decomposition in Lemma D.1, with the component convergence of test errors in Lemma D.2 and of train errors in Lemma D.3. To prove Lemma D.3, we further make of the component concentration results presented in Appendices D.4 and D.5, and component deterministic approximation results presented in Appendix D.5. □

D.1. Decomposition of the mean squared error (Lemma D.1)

Lemma D.1 (Decomposition of the mean squared error for the M-ensemble estimator)**.** *For a dataset $\mathcal{D}_n$, let $\boldsymbol{X} \in \mathbb{R}^{n\times p}$ and $\boldsymbol{y} \in \mathbb{R}^n$ be the design matrix and response vector. Let $\boldsymbol{L}_I \in \mathbb{R}^{n\times n}$ be a diagonal matrix with $(\boldsymbol{L}_I)_{ll} = 1$ if $l \in I$ and 0 otherwise. Then the mean squared error evaluated on $\mathcal{D}_n$ decomposes as*

$$\begin{aligned}
\|\boldsymbol{y} - \boldsymbol{X}\widetilde{\boldsymbol{\beta}}^{\lambda}_k\|_2^2 = &-\mathbb{E}\left[\mathrm{Err}_{\mathrm{train}}(\widehat{\boldsymbol{\beta}}^{\lambda}_k(\mathcal{D}_I)) + \mathrm{Err}_{\mathrm{test}}(\widehat{\boldsymbol{\beta}}^{\lambda}_k(\{\mathcal{D}_I\}))\} \mid \mathcal{D}_n\right] \\
&+ 2\mathbb{E}\left[\mathrm{Err}_{\mathrm{train}}(\widetilde{\boldsymbol{\beta}}^{\lambda}_{k,2}(\{\mathcal{D}_I, \mathcal{D}_J\})) + \mathrm{Err}_{\mathrm{test}}(\widetilde{\boldsymbol{\beta}}^{\lambda}_{k,2}(\{\mathcal{D}_I, \mathcal{D}_J\})) \mid \mathcal{D}_n\right]
\end{aligned} \tag{28}$$

where the training and test errors are defined by

$$\begin{aligned}
\mathrm{Err}_{\mathrm{train}}(\widehat{\boldsymbol{\beta}}^{\lambda}_k(\mathcal{D}_{I_\ell})) &= \|\boldsymbol{L}_{I_\ell}(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}^{\lambda}_k(\mathcal{D}_{I_\ell}))\|_2^2 \\
\mathrm{Err}_{\mathrm{test}}(\widehat{\boldsymbol{\beta}}^{\lambda}_k(\mathcal{D}_{I_\ell})) &= \|\boldsymbol{L}_{I_\ell^c}(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}^{\lambda}_k(\mathcal{D}_{I_\ell}))\|_2^2 \\
\mathrm{Err}_{\mathrm{train}}(\widetilde{\boldsymbol{\beta}}^{\lambda}_{k,2}(\{\mathcal{D}_{I_m}, \mathcal{D}_{I_\ell}\})) &= \|\boldsymbol{L}_{I_m\cup \boldsymbol{L}_\ell}(\boldsymbol{y} - \boldsymbol{X}\widetilde{\boldsymbol{\beta}}^{\lambda}_{k,2}(\{\mathcal{D}_{I_m}, \mathcal{D}_{I_\ell}\}))\|_2^2 \\
\mathrm{Err}_{\mathrm{test}}(\widetilde{\boldsymbol{\beta}}^{\lambda}_{k,2}(\{\mathcal{D}_{I_m}, \mathcal{D}_{I_\ell}\})) &= \|\boldsymbol{L}_{I_m^c\cap \boldsymbol{L}_\ell^c}(\boldsymbol{y} - \boldsymbol{X}\widetilde{\boldsymbol{\beta}}^{\lambda}_{k,2}(\{\mathcal{D}_{I_m}, \mathcal{D}_{I_\ell}\}))\|_2^2.
\end{aligned} \tag{29}$$

From Lemma D.1, the numerator of the GCV estimate for a M-ensemble estimator decomposes into a linear combination of the training and test error of all possible 1-ensemble and 2-ensemble estimators. Then the asymptotics of the numerator can be obtained, if we can show that the limits of all components exist and their linear combination remains invariable when M goes off to infinity.

Proof of Lemma D.1. We first decompose the training error into the linear combination of the mean squared errors (evaluated on $\mathcal{D}_n$) for 1-ensemble and 2-ensemble estimators:

$$\frac{1}{n}\|\boldsymbol{y} - \boldsymbol{X}\widetilde{\boldsymbol{\beta}}^{\lambda}_{k,M}\|_2^2$$

$$
\begin{aligned}
&= \frac{1}{n}\sum_{i=1}^{n}\left(\frac{1}{M}\sum_{\ell=1}^{M}(y_i - \boldsymbol{x}_i\widehat{\boldsymbol{\beta}}_k^{\lambda}(\mathcal{D}_{I_\ell})\right)^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}\frac{1}{M^2}\sum_{\ell=1}^{M}(y_i - \boldsymbol{x}_i\widehat{\boldsymbol{\beta}}_k^{\lambda}(\mathcal{D}_{I_\ell}))^2 + \frac{1}{n}\sum_{i=1}^{n}\frac{1}{M^2}\sum_{\substack{m,\ell\in[M]\\ i\neq j}}(y_i - \boldsymbol{x}_i\widehat{\boldsymbol{\beta}}_k^{\lambda}(\mathcal{D}_{I_m}))(y_i - \boldsymbol{x}_i\widehat{\boldsymbol{\beta}}_k^{\lambda}(\mathcal{D}_{I_\ell})) \\
&= \frac{1}{n}\frac{1}{M^2}\sum_{\ell=1}^{M}\|\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}_k^{\lambda}(\mathcal{D}_{I_\ell})\|_2^2 \\
&\quad + \frac{1}{n}\sum_{i=1}^{n}\frac{1}{M^2}\sum_{\substack{m,\ell\in[M]\\ i\neq j}}\frac{1}{2}[4(y_i - \boldsymbol{x}_i\widetilde{\boldsymbol{\beta}}_{\lambda,2}(\{\mathcal{D}_{I_m}, D_{I_\ell}\}))^2 - (y_i - \boldsymbol{x}_i\widehat{\boldsymbol{\beta}}_k^{\lambda}(\mathcal{D}_{I_m}))^2 - (y_i - \boldsymbol{x}_i\widehat{\boldsymbol{\beta}}_k^{\lambda}(\mathcal{D}_{I_\ell}))^2] \\
&= -\left(\frac{1}{M} - \frac{2}{M^2}\right)\sum_{\ell=1}^{M}\frac{1}{n}\|\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}_k^{\lambda}(\mathcal{D}_{I_\ell})\|_2^2 + \frac{2}{M^2}\sum_{\substack{m,\ell\in[M]\\ i\neq j}}\frac{1}{n}\|\boldsymbol{y} - \boldsymbol{X}\widetilde{\boldsymbol{\beta}}_{\lambda,2}(\{\mathcal{D}_{I_m}, \mathcal{D}_{I_\ell}\})\|_2^2.
\end{aligned}
$$

Next, we further decompose the MSE into training and test errors for 1-ensemble and 2-ensemble estimators:

$$
\begin{aligned}
\frac{1}{n}\|\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}_k^{\lambda}(\mathcal{D}_{I_\ell})\|_2^2 &= \frac{1}{n}\|\boldsymbol{L}_{I_\ell}(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}_k^{\lambda}(\mathcal{D}_{I_\ell}))\|_2^2 + \frac{1}{n}\|\boldsymbol{L}_{I_\ell^c}(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}_k^{\lambda}(\mathcal{D}_{I_\ell}))\|_2^2, \\
\frac{1}{n}\|\boldsymbol{y} - \boldsymbol{X}\widetilde{\boldsymbol{\beta}}_{\lambda,2}(\{\mathcal{D}_{I_m}, \mathcal{D}_{I_\ell}\})\|_2^2 &= \frac{1}{n}\|\boldsymbol{L}_{I_m\cup\boldsymbol{L}_\ell}(\boldsymbol{y} - \boldsymbol{X}\widetilde{\boldsymbol{\beta}}_{\lambda,2}(\{\mathcal{D}_{I_m}, \mathcal{D}_{I_\ell}\}))\|_2^2 + \frac{1}{n}\|\boldsymbol{L}_{I_m^c\cap\boldsymbol{L}_\ell^c}(\boldsymbol{y} - \boldsymbol{X}\widetilde{\boldsymbol{\beta}}_{\lambda,2}(\{\mathcal{D}_{I_m}, \mathcal{D}_{I_\ell}\}))\|_2^2
\end{aligned}
$$

The conclusion then readily follows. □

D.2. Convergence of test errors (Lemma D.2)

Lemma D.2 (Convergence of test errors)**.** *Under Assumptions 2.1-2.2, for the test error defined in (29) with* $I_1, I_2 \overset{\text{SRS}}{\sim} \mathcal{I}_k$*, we have that as* $k, n, p \to \infty$*,* $p/n \to \phi \in (0, \infty)$ *and* $p/k \to \phi_s \in [\phi, +\infty]$*,*

$$
\frac{\mathrm{Err}_{\text{test}}(\widehat{\boldsymbol{\beta}}_k^{\lambda}(\mathcal{D}_{I_\ell}))}{n-k} \xrightarrow{\text{a.s.}} \mathscr{R}_1^{\lambda}(\phi, \phi_s) \tag{30}
$$

$$
\frac{\mathrm{Err}_{\text{test}}(\widetilde{\boldsymbol{\beta}}_{k,2}^{\lambda}(\{\mathcal{D}_{I_m}, \mathcal{D}_{I_\ell}\}))}{|I_m^c \cap I_\ell^c|} \xrightarrow{\text{a.s.}} \mathscr{R}_2^{\lambda}(\phi, \phi_s), \tag{31}
$$

where the deterministic functions $\mathscr{R}_M$ *is defined in Lemma A.2.*

Proof of Lemma D.2. From the strong law of large numbers, we have

$$
\begin{aligned}
\frac{1}{k}\mathrm{Err}_{\text{test}}(\widehat{\boldsymbol{\beta}}_k^{\lambda}(\mathcal{D}_{I_\ell})) &\xrightarrow{\text{a.s.}} \mathbb{E}\left[(y_0 - \boldsymbol{x}_0^\top\widehat{\boldsymbol{\beta}}_k^{\lambda}(\mathcal{D}_{I_1}))^2 \,\middle|\, \mathcal{D}_n\right] \\
\frac{1}{|I_m^c \cap I_\ell^c|}\mathrm{Err}_{\text{test}}(\widehat{\boldsymbol{\beta}}_k^{\lambda}(\{\mathcal{D}_{I_m}, \mathcal{D}_{I_\ell}\})) &\xrightarrow{\text{a.s.}} \mathbb{E}\left[(y_0 - \boldsymbol{x}_0^\top\widehat{\boldsymbol{\beta}}_k^{\lambda}(\{\mathcal{D}_{I_m}, \mathcal{D}_{I_\ell}\}))^2 \,\middle|\, \mathcal{D}_n\right].
\end{aligned}
$$

From Lemma A.2 (Patil et al., 2022a, Theorem 4.1), the condition prediction risks converge in the sense that

$$
\begin{aligned}
\mathbb{E}\left[(y_0 - \boldsymbol{x}_0^\top\widehat{\boldsymbol{\beta}}_k^{\lambda}(\mathcal{D}_{I_1}))^2 \,\middle|\, \mathcal{D}_n\right] &\xrightarrow{\text{a.s.}} \mathscr{R}_1(\phi, \phi_s) \\
\mathbb{E}\left[(y_0 - \boldsymbol{x}_0^\top\widetilde{\boldsymbol{\beta}}_{k,2}^{\lambda}(\{\mathcal{D}_{I_m}, \mathcal{D}_{I_\ell}\}))^2 \,\middle|\, \mathcal{D}_n\right] &\xrightarrow{\text{a.s.}} \mathscr{R}_2(\phi, \phi_s),
\end{aligned}
$$

and the conclusions follow. □

D.3. Convergence of train errors (Lemma D.3)

Lemma D.3 (Convergence of train errors)**.** *Under Assumptions 2.1-2.2, for the train error defined in (29) with* $I_1, I_2 \overset{\text{SRS}}{\sim} \mathcal{I}_k$*, we have that as* $k, n, p \to \infty$*,* $p/n \to \phi \in (0, \infty)$ *and* $p/k \to \phi_s \in [\phi, +\infty)$*,*

$$\begin{aligned}
\frac{1}{k}\mathrm{Err}_{\text{train}}(\widehat{\boldsymbol{\beta}}_k^\lambda(\mathcal{D}_{I_\ell})) &\xrightarrow{\text{a.s.}} \mathscr{T}_1^\lambda(\phi, \phi_s) := \mathscr{D}^\lambda(\phi_s, \phi_s)\mathscr{R}_1^\lambda(\phi, \phi_s) \\
\frac{1}{|I_m \cup I_\ell|}\mathrm{Err}_{\text{train}}(\widetilde{\boldsymbol{\beta}}_{k,1}^\lambda(\{\mathcal{D}_{I_m}, \mathcal{D}_{I_\ell}\})) &\xrightarrow{\text{a.s.}} \mathscr{T}_2^\lambda(\phi, \phi_s) := \frac{1}{2}\frac{\phi_s - \phi}{2\phi_s - \phi}\mathscr{R}_1^\lambda(\phi, \phi_s) + \frac{1}{2}\mathscr{D}^\lambda(\phi_s, \phi_s) \\
&\left(\frac{\phi_s}{2\phi_s - \phi}\mathscr{R}_1^\lambda(\phi, \phi_s) + \left(\frac{2(\phi_s - \phi)}{2\phi_s - \phi}\,\frac{1}{\lambda v(-\lambda; \phi_s)} + \frac{\phi}{2\phi_s - \phi}\right)(2\mathscr{R}_2^\lambda(\phi, \phi_s) - \mathscr{R}_1^\lambda(\phi, \phi_s))\right),
\end{aligned} \tag{32}$$

where the deterministic function $\mathscr{R}_M$ *is defined in Lemma A.2.*

Proof of Lemma D.3. From (27), we have

$$\boldsymbol{\beta}_0 - \widetilde{\boldsymbol{\beta}}_{\lambda,M} = \frac{1}{M}\sum_{m=1}^M \lambda \boldsymbol{M}_m \boldsymbol{\beta}_0 - \frac{1}{M}\sum_{m=1}^M \boldsymbol{M}_m(\boldsymbol{X}^\top \boldsymbol{L}_m/k)\boldsymbol{\epsilon}. \tag{33}$$

Part (1) Case of $M = 1$:

From (33), the train error can be decomposed as follows:

$$\begin{aligned}
\frac{1}{k}\mathrm{Err}_{\text{train}}(\widehat{\boldsymbol{\beta}}_k^\lambda(\mathcal{D}_{I_\ell})) &= \|\boldsymbol{L}_\ell\boldsymbol{\epsilon} + \boldsymbol{L}_\ell\boldsymbol{X}(\boldsymbol{\beta}_0 - \widehat{\boldsymbol{\beta}}_k^\lambda(\mathcal{D}_{I_\ell}))\|_2^2/k \\
&= \|(\boldsymbol{L}_\ell - \boldsymbol{L}_\ell\boldsymbol{X}\boldsymbol{M}_\ell\boldsymbol{X}^\top\boldsymbol{L}_\ell/k)\boldsymbol{\epsilon} + \lambda\boldsymbol{L}_\ell\boldsymbol{X}\boldsymbol{M}_\ell\boldsymbol{\beta}_0\|_2^2/k \\
&= T_C + T_B + T_V,
\end{aligned}$$

where the constant term T_C, bias term T_B, and the variance term T_V are given by

$$T_C = \frac{2\lambda}{k}\boldsymbol{\epsilon}^\top\boldsymbol{L}_\ell\left(\boldsymbol{I}_n - \boldsymbol{X}\boldsymbol{M}_\ell\frac{\boldsymbol{X}^\top\boldsymbol{L}_\ell}{k}\right)^\top\boldsymbol{L}_\ell\boldsymbol{X}\boldsymbol{M}_\ell\boldsymbol{\beta}_0, \tag{34}$$

$$T_B = \lambda^2\boldsymbol{\beta}_0^\top\boldsymbol{M}_\ell\widehat{\boldsymbol{\Sigma}}_\ell\boldsymbol{M}_\ell\boldsymbol{\beta}_0, \tag{35}$$

$$T_V = \frac{1}{k}\boldsymbol{\epsilon}^\top\boldsymbol{L}_\ell\left(\boldsymbol{I}_n - \boldsymbol{X}\boldsymbol{M}_\ell\frac{\boldsymbol{X}^\top\boldsymbol{L}_\ell}{k}\right)^\top\boldsymbol{L}_\ell\left(\boldsymbol{I}_n - \boldsymbol{X}\boldsymbol{M}_\ell\frac{\boldsymbol{X}^\top\boldsymbol{L}_\ell}{k}\right)\boldsymbol{L}_\ell\boldsymbol{\epsilon}. \tag{36}$$

Next, we analyze the three terms separately. From Lemmas D.4 and D.5 with $n = k$, we have that $T_C \xrightarrow{\text{a.s.}} 0$, and

$$T_V \xrightarrow{\text{a.s.}} \sigma^2\left(1 - \frac{2}{k}\operatorname{tr}(\boldsymbol{M}_m\widehat{\boldsymbol{\Sigma}}_m) + \frac{1}{k}\operatorname{tr}(\boldsymbol{M}_m\widehat{\boldsymbol{\Sigma}}_m\boldsymbol{M}_m\widehat{\boldsymbol{\Sigma}}_m)\right) := T_{VT}.$$

Thus, it remains to obtain the deterministic equivalent for the bias term T_B and the trace term T_{VT}.

From Lemma D.6 and Lemma D.7, we have that for all $I_1 \in \mathcal{I}_k$,

$$\begin{aligned}
T_B &\xrightarrow{\text{a.s.}} \rho^2\mathscr{D}^\lambda(\phi_s, \phi_s)(1 + \widetilde{v}(-\lambda; \phi_s, \phi_s))\widetilde{c}(-\lambda; \phi_s) \\
T_{VT} &\xrightarrow{\text{a.s.}} \sigma^2\mathscr{D}^\lambda(\phi_s, \phi_s)(1 + \widetilde{v}(-\lambda; \phi_s, \phi_s)).
\end{aligned}$$

Then, we have

$$\frac{1}{k}\mathrm{Err}_{\text{train}}(\widehat{\boldsymbol{\beta}}_k^\lambda(\mathcal{D}_{I_\ell})) \xrightarrow{\text{a.s.}} \mathscr{R}_{\lambda,1}(\phi, \phi_s)\mathscr{D}^\lambda(\phi_s, \phi_s)$$

where $\mathscr{R}_{\lambda,1}$ and $\mathscr{D}_\lambda$ are defined in (21) and (26), respectively.

Part (2) Case of $M = 2$:

From (33), the train error can be analogously decomposed as follows:

$$\frac{1}{|I_m \cup I_l|}\mathrm{Err}_{\mathrm{train}}(\widetilde{\boldsymbol{\beta}}^{\lambda}_{k,2}(\{\mathcal{D}_{I_1}, \mathcal{D}_{I_2}\})) = \frac{1}{|I_m \cup I_l|}\|\boldsymbol{L}_{m\cup l}(\boldsymbol{\epsilon} + \boldsymbol{X}(\boldsymbol{\beta}_0 - \widehat{\boldsymbol{\beta}}_{\lambda,2}))\|_2^2 = T'_C + T'_B + T'_V,$$

where the constant term T_C, bias term T_B, and the variance term T_V are given by

$$T'_C = \frac{\lambda}{2|I_m \cup I_l|}\sum_{i,j\in\{m,l\}} \boldsymbol{\epsilon}^\top \left(\boldsymbol{I}_n - \boldsymbol{X}\boldsymbol{M}_i\frac{\boldsymbol{X}^\top \boldsymbol{L}_i}{k}\right)^\top \boldsymbol{L}_{m\cup l}\boldsymbol{X}\boldsymbol{M}_j\boldsymbol{\beta}_0, \tag{37}$$

$$\begin{aligned} T'_B &= \frac{\lambda^2}{4|I_m \cup I_l|}\sum_{i,j\in\{m,l\}} \boldsymbol{\beta}_0^\top \boldsymbol{M}_i\widehat{\boldsymbol{\Sigma}}_{m\cup l}\boldsymbol{M}_j\boldsymbol{\beta}_0 \\ &= \frac{\lambda^2 k}{4|I_m \cup I_l|}\sum_{i\in\{m,l\}} \boldsymbol{\beta}_0^\top \boldsymbol{M}_i\widehat{\boldsymbol{\Sigma}}_i\boldsymbol{M}_i\boldsymbol{\beta}_0 + \frac{\lambda^2}{4|I_m \cup I_l|}\sum_{i\in\{m,l\}} |I_{m+l-i} \setminus I_i|\boldsymbol{\beta}_0^\top \boldsymbol{M}_i\widehat{\boldsymbol{\Sigma}}_{(m+l-i)\setminus i}\boldsymbol{M}_i\boldsymbol{\beta}_0 \\ &\quad + \frac{\lambda^2}{4}\sum_{i\in\{m,l\}} \boldsymbol{\beta}_0^\top \boldsymbol{M}_i\widehat{\boldsymbol{\Sigma}}_{m\cup l}\boldsymbol{M}_{m+l-i}\boldsymbol{\beta}_0, \end{aligned} \tag{38}$$

$$T'_V = \frac{1}{4|I_m \cup I_l|}\sum_{i,j\in\{m,l\}} \boldsymbol{\epsilon}^\top \left(\boldsymbol{I}_n - \boldsymbol{X}\boldsymbol{M}_i\frac{\boldsymbol{X}^\top \boldsymbol{L}_i}{k}\right)^\top \boldsymbol{L}_{m\cup l}\left(\boldsymbol{I}_n - \boldsymbol{X}\boldsymbol{M}_j\frac{\boldsymbol{X}^\top \boldsymbol{L}_j}{k}\right)\boldsymbol{\epsilon}. \tag{39}$$

Next, we analyze the three terms separately. From Lemmas D.4 and D.5, we have that $T_C \xrightarrow{\text{a.s.}} 0$, and

$$\begin{aligned} T'_V &\xrightarrow{\text{a.s.}} \frac{\sigma^2}{4}\sum_{i\in\{m,l\}}\left(1 - \frac{2}{|I_m \cup I_l|}\operatorname{tr}(\boldsymbol{M}_i\widehat{\boldsymbol{\Sigma}}_i) + \frac{1}{k}\operatorname{tr}(\boldsymbol{M}_i\widehat{\boldsymbol{\Sigma}}_i\boldsymbol{M}_i\widehat{\boldsymbol{\Sigma}}_{m\cup l})\right) \\ &\quad + \frac{\sigma^2}{2}\left(1 - \frac{1}{|I_m \cup I_l|}\sum_{j\in\{m,l\}}\operatorname{tr}(\boldsymbol{M}_j\widehat{\boldsymbol{\Sigma}}_j) + \frac{1}{n}\operatorname{tr}(\boldsymbol{M}_l\widehat{\boldsymbol{\Sigma}}_{m\cap l}\boldsymbol{M}_l\widehat{\boldsymbol{\Sigma}}_{m\cup l})\right) \\ &= \frac{\phi_s}{2\phi_s - \phi}\frac{T_{VT}}{2} + \frac{\sigma^2}{4}\frac{\phi_s - \phi}{2\phi_s - \phi}\left(2 + \frac{1}{k}\sum_{i\in\{m,l\}}\operatorname{tr}(\boldsymbol{M}_i\widehat{\boldsymbol{\Sigma}}_i\boldsymbol{M}_i\widehat{\boldsymbol{\Sigma}}_{(m+l-i)\setminus i})\right) \\ &\quad + \frac{\sigma^2}{2}\left(1 - \frac{1}{|I_m \cup I_l|}\sum_{j\in\{m,l\}}\operatorname{tr}(\boldsymbol{M}_j\widehat{\boldsymbol{\Sigma}}_j) + \frac{1}{n}\operatorname{tr}(\boldsymbol{M}_l\widehat{\boldsymbol{\Sigma}}_{m\cap l}\boldsymbol{M}_m\widehat{\boldsymbol{\Sigma}}_{m\cup l})\right) := T'_{VT}. \end{aligned}$$

Thus, it remains to obtain the deterministic equivalent for the bias term T'_B and the trace term T'_{VT}.

From Lemma D.6 and Lemma D.7, and the convergence of the cardinality from Lemma G.2, we have that for all $I_m, I_l \overset{\text{SRS}}{\sim} \mathcal{I}_k$,

$$T'_B \xrightarrow{\text{a.s.}} \frac{\rho^2}{2}\widetilde{t}(\phi, \phi_s)\widetilde{c}(-\lambda; \phi_s), \quad \text{and} \quad T'_V \xrightarrow{\text{a.s.}} \frac{\sigma^2}{2}\widetilde{t}(\phi, \phi_s),$$

where

$$\begin{aligned} \widetilde{t}(\phi, \phi_s) &= \frac{\phi_s}{2\phi_s - \phi}\mathscr{D}^{\lambda}(\phi_s, \phi_s)(1 + \widetilde{v}(-\lambda; \phi_s, \phi_s)) + \frac{\phi_s - \phi}{2\phi_s - \phi}(1 + \widetilde{v}(-\lambda; \phi_s, \phi_s)) \\ &\quad + \mathscr{D}^{\lambda}(\phi_s, \phi_s)\left(\frac{2(\phi_s - \phi)}{2\phi_s - \phi}\frac{1}{\lambda v(-\lambda; \phi_s)} + \frac{\phi}{2\phi_s - \phi}\right)(1 + \widetilde{v}(-\lambda; \phi, \phi_s)). \end{aligned}$$

Then, we have

$$\frac{1}{|I_m \cup I_l|}\mathrm{Err}_{\mathrm{train}}(\widetilde{\boldsymbol{\beta}}^{\lambda}_{k,2}(\{\mathcal{D}_{I_1}, \mathcal{D}_{I_2}\}))$$

$$\xrightarrow{\text{a.s.}} \frac{\rho^2}{2}\widetilde{t}(\phi,\phi_s)\widetilde{c}(-\lambda;\phi_s) + \frac{\sigma^2}{2}\widetilde{t}(\phi,\phi_s)$$
$$= \frac{1}{2}\mathscr{D}^{\lambda}(\phi_s,\phi_s)\left(\frac{\phi_s}{2\phi_s-\phi}\mathscr{R}_{\lambda,1}(\phi,\phi_s) + \left(\frac{2(\phi_s-\phi)}{2\phi_s-\phi}\frac{1}{\lambda v(-\lambda;\phi_s)} + \frac{\phi}{2\phi_s-\phi}\right)(2\mathscr{R}_{\lambda,2}(\phi,\phi_s) - \mathscr{R}_{\lambda,2}(\phi,\phi_s))\right)$$
$$+ \frac{1}{2}\frac{\phi_s-\phi}{2\phi_s-\phi}\mathscr{R}_1(\phi,\phi_s),$$

which finishes the proof. □

D.4. Component concentrations

In this subsection, we will show that the cross-term T_C converges to zero and the variance term T_V converges to its corresponding trace expectation.

D.4.1. Convergence of the cross term

Lemma D.4 (Convergence of the cross term)**.** *Under Assumptions 2.1-2.2, for T_C and T_C' as defined in (34) and (37)), we have $T_C \xrightarrow{\text{a.s.}} 0$ and $T_C' \xrightarrow{\text{a.s.}} 0$ as $k,p\to\infty$ and $p/k\to\phi_s$.*

Proof of Lemma D.4. We first prove the result for T_C'. Note that

$$T_C' = -\frac{\lambda}{M^2}\cdot\frac{1}{|I_1\cup I_2|}\left\langle \sum_{m=1}^{2}\left(\boldsymbol{I}_n - \boldsymbol{X}\frac{\boldsymbol{M}_m\boldsymbol{X}^\top\boldsymbol{L}_m}{k}\right)^\top \boldsymbol{L}_{m\cup l}\boldsymbol{X}\sum_{m=1}^{2}\boldsymbol{M}_m\boldsymbol{\beta}_0, \boldsymbol{\epsilon}\right\rangle.$$

We next bound the squared norm:

$$\frac{1}{|I_m\cup I_l|}\left\|\frac{1}{2}\sum_{m=1}^{2}\left(\boldsymbol{I}_n - \frac{\boldsymbol{X}\boldsymbol{M}_m\boldsymbol{X}^\top\boldsymbol{L}_m}{k}\right)^\top \boldsymbol{L}_{m\cup l}\boldsymbol{X}\sum_{m=1}^{2}\boldsymbol{M}_m\boldsymbol{\beta}_0\right\|_2^2$$
$$\le \sum_{j=1}^{2}\sum_{l=1}^{2}\left[\frac{|I_m\cup I_l|}{4k^2}\left\|(\boldsymbol{M}_j\boldsymbol{X}^\top\boldsymbol{L}_j)^\top\widehat{\boldsymbol{\Sigma}}_{m\cup l}\boldsymbol{M}_l\boldsymbol{\beta}_0\right\|_2^2 + \frac{1}{4|I_1\cup I_2|}\|\boldsymbol{L}_j\boldsymbol{X}\boldsymbol{M}_l\boldsymbol{\beta}_0\|_2^2\right]$$
$$\le \frac{\|\boldsymbol{\beta}_0\|_2^2}{4}\cdot\sum_{j=1}^{2}\sum_{l=1}^{2}\left[\frac{|I_m\cup I_l|}{k^2}\left\|\boldsymbol{M}_l\widehat{\boldsymbol{\Sigma}}_{m\cup l}\boldsymbol{M}_j\boldsymbol{X}^\top\boldsymbol{L}_j\boldsymbol{X}\boldsymbol{M}_j\widehat{\boldsymbol{\Sigma}}_{m\cup l}\boldsymbol{M}_l\right\|_{\text{op}} + \frac{k}{|I_1\cup I_2|}\left\|\widehat{\boldsymbol{\Sigma}}_j\right\|_{\text{op}}\|\boldsymbol{M}_l\|_{\text{op}}\right]$$
$$\le \frac{\|\boldsymbol{\beta}_0\|_2^2}{4}\cdot\sum_{j=1}^{2}\sum_{l=1}^{2}\left[\frac{|I_m\cup I_l|}{k}\|\boldsymbol{M}_l\|_{\text{op}}^2\left\|\widehat{\boldsymbol{\Sigma}}_l\right\|_{\text{op}}^2\left\|\boldsymbol{M}_j(\boldsymbol{X}^\top\boldsymbol{L}_j\boldsymbol{X}/k)\boldsymbol{M}_j\right\|_{\text{op}} + \frac{k}{|I_1\cup I_2|}\|\boldsymbol{M}_l\|_{\text{op}}\left\|\widehat{\boldsymbol{\Sigma}}_l\right\|_{\text{op}}\right]$$
$$= \frac{\|\boldsymbol{\beta}_0\|_2^2}{4}\cdot\sum_{j=1}^{2}\sum_{l=1}^{2}\left[\frac{|I_m\cup I_l|}{k}\|\boldsymbol{M}_l\|_{\text{op}}^2\left\|\widehat{\boldsymbol{\Sigma}}_l\right\|_{\text{op}}^2\|\boldsymbol{M}_j\|_{\text{op}}\|\boldsymbol{I}_p-\lambda\boldsymbol{M}_j\|_{\text{op}} + \frac{k}{|I_1\cup I_2|}\|\boldsymbol{M}_l\|_{\text{op}}\left\|\widehat{\boldsymbol{\Sigma}}_l\right\|_{\text{op}}\right]$$
$$\le \frac{\|\boldsymbol{\beta}_0\|_2^2}{\lambda}\left(\frac{|I_m\cup I_l|}{k\lambda^2} + \frac{k}{|I_1\cup I_2|}\right)\left\|\widehat{\boldsymbol{\Sigma}}_l\right\|_{\text{op}}^2,$$

where the last inequality is due to the fact that $\|\boldsymbol{M}_j\|_{\text{op}}\le 1/\lambda$ and $\|\boldsymbol{I}_p-\lambda\boldsymbol{M}_j\|_{\text{op}}\le 1$. By Assumption 2.2, $\|\boldsymbol{\beta}_0\|_2^2$ is uniformly bounded in p. From Bai & Silverstein (2010), we have $\limsup\left\|\widehat{\boldsymbol{\Sigma}}\right\|_{\text{op}} \le \limsup\max_{1\le i\le p}s_i^2 \le r_{\max}(1+\sqrt{\phi_s})^2$ almost surely as $k,p\to\infty$ and $p/k\to\phi_s\in(0,\infty)$. From Lemma G.2, we have that $k/|I_1\cup I_2| \xrightarrow{\text{a.s.}} k/(2k-k^2/n)$, which is $\phi_s/(2\phi_s-\phi)$ almost surely. Then we have that the square norm is almost surely upper bounded by some constant. Applying Lemma G.3, we thus have that $T_C' \xrightarrow{\text{a.s.}} 0$.

Note that when $I_1=I_2$, T_C' reduces to T_C, and thus the conclusion for T_C also holds. □

D.4.2. Convergence of the variance term

Lemma D.5 (Convergence of the variance term)**.** *Under Assumptions 2.1-2.2, let $M\in\mathbb{N}$ and $\widehat{\boldsymbol{\Sigma}} = \boldsymbol{X}^\top\boldsymbol{X}/n$. For all $m\in[M]$ and $I_m\in\mathcal{I}_k$, let $\widehat{\boldsymbol{\Sigma}}_m = \boldsymbol{X}^\top\boldsymbol{L}_m\boldsymbol{X}/k$, $\boldsymbol{L}_m\in\mathbb{R}^{n\times n}$ be a diagonal matrix with $(\boldsymbol{L}_m)_{ll}=1$ if $l\in I_m$ and 0*

otherwise, and $\boldsymbol{M}_m = (\boldsymbol{X}^\top \boldsymbol{L}_m \boldsymbol{X}/k + \lambda \boldsymbol{I}_p)^{-1}$. *Then, for all* $m, l \in [M]$ *and* $m \neq l$, *it holds that*

$$\frac{1}{k}\boldsymbol{\epsilon}^\top \left(\boldsymbol{I}_n - \boldsymbol{X}\boldsymbol{M}_m \frac{\boldsymbol{X}^\top \boldsymbol{L}_m}{k}\right)^\top \boldsymbol{L}_m \left(\boldsymbol{I}_n - \boldsymbol{X}\boldsymbol{M}_m \frac{\boldsymbol{X}^\top \boldsymbol{L}_m}{k}\right)\boldsymbol{\epsilon} - \sigma^2\left(1 - \frac{2}{k}\operatorname{tr}(\boldsymbol{M}_m \widehat{\boldsymbol{\Sigma}}_m) + \frac{1}{k}\operatorname{tr}(\boldsymbol{M}_m \widehat{\boldsymbol{\Sigma}}_m \boldsymbol{M}_m \widehat{\boldsymbol{\Sigma}}_m)\right) \xrightarrow{\text{a.s.}} 0, \tag{40}$$

$$\frac{1}{|I_m \cup I_l|}\boldsymbol{\epsilon}^\top \left(\boldsymbol{I}_n - \boldsymbol{X}\boldsymbol{M}_i \frac{\boldsymbol{X}^\top \boldsymbol{L}_i}{k}\right)^\top \boldsymbol{L}_{m\cup l} \left(\boldsymbol{I}_n - \boldsymbol{X}\boldsymbol{M}_i \frac{\boldsymbol{X}^\top \boldsymbol{L}_i}{k}\right)\boldsymbol{\epsilon} - \sigma^2\left(1 - \frac{2}{|I_m \cup I_l|}\operatorname{tr}(\boldsymbol{M}_i \widehat{\boldsymbol{\Sigma}}_i) + \frac{1}{k}\operatorname{tr}(\boldsymbol{M}_i \widehat{\boldsymbol{\Sigma}}_i \boldsymbol{M}_i \widehat{\boldsymbol{\Sigma}}_{m\cup l})\right) \xrightarrow{\text{a.s.}} 0, \tag{41}$$

$$\frac{1}{|I_m \cup I_l|}\boldsymbol{\epsilon}^\top \left(\boldsymbol{I}_n - \boldsymbol{X}\boldsymbol{M}_m \frac{\boldsymbol{X}^\top \boldsymbol{L}_m}{k}\right)^\top \boldsymbol{L}_{m\cup l} \left(\boldsymbol{I}_n - \boldsymbol{X}\boldsymbol{M}_l \frac{\boldsymbol{X}^\top \boldsymbol{L}_l}{k}\right)\boldsymbol{\epsilon} - \sigma^2\left(1 - \frac{1}{|I_m \cup I_l|}\sum_{\ell\in\{i,j\}} \operatorname{tr}(\boldsymbol{M}_\ell \widehat{\boldsymbol{\Sigma}}_\ell) + \frac{1}{n}\operatorname{tr}(\boldsymbol{M}_i \widehat{\boldsymbol{\Sigma}}_{i\cap j} \boldsymbol{M}_j \widehat{\boldsymbol{\Sigma}}_{m\cup l})\right) \xrightarrow{\text{a.s.}} 0, \tag{42}$$

where $i, j \in \{m, l\}$, $i \neq j$, *and* $\widehat{\boldsymbol{\Sigma}}_{m\cup l} = \boldsymbol{X}^\top \boldsymbol{L}_{m\cup l}\boldsymbol{X}/|I_m \cup I_l|$, *as* $n, k, p \to \infty$, $p/n \to \phi \in (0, \infty)$, *and* $p/k \to \phi_s \in [\phi, \infty)$.

Proof of Lemma D.5. We first prove the last convergence result. Note that

$$\begin{aligned}
&\left\|\left(\boldsymbol{I}_n - \boldsymbol{X}\boldsymbol{M}_m \frac{\boldsymbol{X}^\top \boldsymbol{L}_m}{k}\right)^\top \boldsymbol{L}_{m\cup l} \left(\boldsymbol{I}_n - \boldsymbol{X}\boldsymbol{M}_l \frac{\boldsymbol{X}^\top \boldsymbol{L}_l}{k}\right)\right\|_{\text{op}} \\
&= \left\|\boldsymbol{L}_{m\cup l} - \frac{1}{k}\boldsymbol{L}_m \boldsymbol{X}\boldsymbol{M}_m \boldsymbol{X}^\top \boldsymbol{L}_{m\cup l} - \frac{1}{k}\boldsymbol{L}_{m\cup l}\boldsymbol{X}\boldsymbol{M}_l \boldsymbol{X}^\top \boldsymbol{L}_l + \frac{1}{k^2}\boldsymbol{L}_m \boldsymbol{X}\boldsymbol{M}_m \boldsymbol{X}^\top \boldsymbol{L}_{m\cup l}\boldsymbol{X}\boldsymbol{M}_l \boldsymbol{X}^\top \boldsymbol{L}_l\right\|_{\text{op}} \\
&\leq 1 + \sqrt{\frac{|I_m \cup I_l|}{k}} \sum_{j\in\{m,l\}} \left\|\widehat{\boldsymbol{\Sigma}}_j\right\|_{\text{op}}^{\frac{1}{2}} \|\boldsymbol{M}_j\|_{\text{op}} \left\|\widehat{\boldsymbol{\Sigma}}_{m\cup l}\right\|_{\text{op}}^{\frac{1}{2}} + \frac{|I_m \cup I_l|}{k} \left\|\widehat{\boldsymbol{\Sigma}}_m\right\|_{\text{op}}^{\frac{1}{2}} \|\boldsymbol{M}_m\|_{\text{op}} \left\|\widehat{\boldsymbol{\Sigma}}_{m\cup l}\right\|_{\text{op}} \|\boldsymbol{M}_l\|_{\text{op}} \left\|\widehat{\boldsymbol{\Sigma}}_l\right\|_{\text{op}}^{\frac{1}{2}} \\
&\leq 1 + \frac{2}{\lambda}\sqrt{\frac{|I_m \cup I_l|}{k}} \left\|\widehat{\boldsymbol{\Sigma}}_m\right\|_{\text{op}}^{\frac{1}{2}} \left\|\widehat{\boldsymbol{\Sigma}}_l\right\|_{\text{op}}^{\frac{1}{2}} + \frac{1}{\lambda^2}\frac{|I_m \cup I_l|}{k} \left\|\widehat{\boldsymbol{\Sigma}}_m\right\|_{\text{op}}^{\frac{1}{2}} \left\|\widehat{\boldsymbol{\Sigma}}_{m\cup l}\right\|_{\text{op}} \left\|\widehat{\boldsymbol{\Sigma}}_l\right\|_{\text{op}}^{\frac{1}{2}}.
\end{aligned}$$

Now, we have $\limsup \|\widehat{\boldsymbol{\Sigma}}\|_{\text{op}} \leq \limsup \max_{1\leq i\leq p} s_i^2 \leq r_{\max}(1 + \sqrt{\phi})^2$ almost surely as $n, p \to \infty$ and $p/n \to \phi \in (0, \infty)$ from Bai & Silverstein (2010). Similarly, $\limsup \|\widehat{\boldsymbol{\Sigma}}_m\|_{\text{op}} \leq r_{\max}(1 + \sqrt{\phi_s})^2$ almost surely. From Lemma G.2, $|I_m \cup I_l|/k \xrightarrow{\text{a.s.}} (2\phi_s - \phi)/\phi_s$. Then the above quantity is asymptotically upper bounded by some constant as $n, k, p \to \infty$, $p/n \to \phi \in (0, \infty)$ and $p/k \to \phi_s \in [\phi, \infty)$. From Lemma G.4, it follows that

$$\frac{1}{|I_m \cup I_l|}\boldsymbol{\epsilon}^\top \left(\boldsymbol{I}_n - \boldsymbol{X}\boldsymbol{M}_i \frac{\boldsymbol{X}^\top \boldsymbol{L}_i}{k}\right)^\top \boldsymbol{L}_{m\cup l} \left(\boldsymbol{I}_n - \boldsymbol{X}\boldsymbol{M}_j \frac{\boldsymbol{X}^\top \boldsymbol{L}_j}{k}\right)\boldsymbol{\epsilon} - \frac{\sigma^2}{|I_m \cup I_l|}\operatorname{tr}\left(\left(\boldsymbol{I}_n - \boldsymbol{X}\boldsymbol{M}_i \frac{\boldsymbol{X}^\top \boldsymbol{L}_i}{k}\right)^\top \boldsymbol{L}_{m\cup l} \left(\boldsymbol{I}_n - \boldsymbol{X}\boldsymbol{M}_j \frac{\boldsymbol{X}^\top \boldsymbol{L}_j}{k}\right)\right) \xrightarrow{\text{a.s.}} 0.$$

Expanding the trace term above, we have

$$\begin{aligned}
&\frac{\sigma^2}{|I_m \cup I_l|}\operatorname{tr}\left(\left(\boldsymbol{I}_n - \boldsymbol{X}\boldsymbol{M}_i \frac{\boldsymbol{X}^\top \boldsymbol{L}_i}{k}\right)^\top \boldsymbol{L}_{m\cup l} \left(\boldsymbol{I}_n - \boldsymbol{X}\boldsymbol{M}_j \frac{\boldsymbol{X}^\top \boldsymbol{L}_j}{k}\right)\right) \\
&= \sigma^2\left(1 - \frac{1}{|I_m \cup I_l|}\sum_{\ell\in\{i,j\}} \operatorname{tr}(\boldsymbol{M}_\ell \widehat{\boldsymbol{\Sigma}}_\ell) + \frac{|I_i \cap I_j|}{k^2}\operatorname{tr}(\boldsymbol{M}_i \widehat{\boldsymbol{\Sigma}}_{i\cup j} \boldsymbol{M}_j \widehat{\boldsymbol{\Sigma}}_{m\cap l})\right).
\end{aligned} \tag{43}$$

Since $I_m, I_l \overset{\text{SRS}}{\sim} \mathcal{I}_k$, from Lemma G.2 we have that $|I_m \cap I_l|/k \xrightarrow{\text{a.s.}} k/n$. Then, we have

$$\frac{1}{|I_m \cup I_l|}\boldsymbol{\epsilon}^\top \left(\boldsymbol{I}_n - \boldsymbol{X}\boldsymbol{M}_i\frac{\boldsymbol{X}^\top \boldsymbol{L}_i}{k}\right)^\top \boldsymbol{L}_{m\cup l}\left(\boldsymbol{I}_n - \boldsymbol{X}\boldsymbol{M}_j\frac{\boldsymbol{X}^\top \boldsymbol{L}_j}{k}\right)\boldsymbol{\epsilon} - \sigma^2\left(1 - \frac{1}{|I_m \cup I_l|}\sum_{\ell\in\{i,j\}}\operatorname{tr}(\boldsymbol{M}_\ell \widehat{\boldsymbol{\Sigma}}_\ell) + \frac{1}{n}\operatorname{tr}(\boldsymbol{M}_i\widehat{\boldsymbol{\Sigma}}_{i\cap j}\boldsymbol{M}_j\widehat{\boldsymbol{\Sigma}}_{m\cup l})\right) \xrightarrow{\text{a.s.}} 0,$$

and thus (42) follows.

Setting $i = j$ in (43) yields (41).

Finally, setting $i = j = l = m$ in (43) finishes the proof for (40). □

D.5. Component deterministic approximations

D.5.1. Deterministic approximation of the bias functional

Lemma D.6 (Deterministic approximation of the bias functional)**.** *Under Assumptions 2.1-2.2, for all $m \in [M]$ and $I_m \in \mathcal{I}_k$, let $\widehat{\boldsymbol{\Sigma}}_m = \boldsymbol{X}^\top \boldsymbol{L}_m \boldsymbol{X}/k$, $\boldsymbol{L}_m \in \mathbb{R}^{n\times n}$ be a diagonal matrix with $(\boldsymbol{L}_m)_{ll} = 1$ if $l \in I_m$ and 0 otherwise, and $\boldsymbol{M}_m = (\boldsymbol{X}^\top \boldsymbol{L}_m \boldsymbol{X}/k + \lambda \boldsymbol{I}_p)^{-1}$. Then, it holds that*

1. *For all $m \in [M]$,*

$$\lambda^2 \boldsymbol{\beta}_0^\top \boldsymbol{M}_m \widehat{\boldsymbol{\Sigma}}_m \boldsymbol{M}_m \boldsymbol{\beta}_0 \xrightarrow{\text{a.s.}} \rho^2\lambda^2 v(-\lambda;\phi_s)^2(1+\widetilde{v}(-\lambda;\phi_s,\phi_s))\widetilde{c}(-\lambda;\phi_s).$$

2. *For all $m, l \in [M]$, $m \neq l$ and $I_m, I_l \overset{\text{SRSWR}}{\sim} \mathcal{I}_k$,*

$$\lambda^2 \boldsymbol{\beta}_0^\top \boldsymbol{M}_l \widehat{\boldsymbol{\Sigma}}_{m\setminus l} \boldsymbol{M}_l \boldsymbol{\beta}_0 \xrightarrow{\text{a.s.}} \rho^2(1+\widetilde{v}(-\lambda;\phi_s,\phi_s))\widetilde{c}(-\lambda;\phi_s). \tag{44}$$

3. *For all $m, l \in [M]$, $m \neq l$ and $I_m, I_l \overset{\text{SRSWR}}{\sim} \mathcal{I}_k$,*

$$\lambda^2 \boldsymbol{\beta}_0^\top \boldsymbol{M}_m \widehat{\boldsymbol{\Sigma}}_{m\cup l} \boldsymbol{M}_l \boldsymbol{\beta}_0 \xrightarrow{\text{a.s.}} \rho^2\lambda^2\left(\frac{2(\phi_s-\phi)}{2\phi_s-\phi}\frac{1}{\lambda v(-\lambda;\phi_s)} + \frac{\phi}{2\phi_s-\phi}\right) v(-\lambda;\phi_s)^2(1+\widetilde{v}(-\lambda;\phi,\phi_s))\widetilde{c}(-\lambda;\phi_s), \tag{45}$$

where $\widehat{\boldsymbol{\Sigma}}_{m\cup l} = |I_m \cup I_l|^{-1}\boldsymbol{X}^\top \boldsymbol{L}_{m\cup l}\boldsymbol{X}$, as $n, k, p \to \infty$, $p/n \to \phi \in (0,\infty)$, and $p/k \to \phi_s \in [\phi, \infty)$, where $\phi_0 = \phi_s^2/\phi$, T_B is as defined in (35), and the nonnegative constants $\widetilde{v}(-\lambda;\phi,\phi_s)$ and $\widetilde{c}(-\lambda;\phi_s)$ are as defined in (19).

Proof of Lemma D.6. We split the proof into different parts.

Part (1) From Lemma F.6 (2) (with $\boldsymbol{A} = \boldsymbol{I}_p$), we have that

$$\lambda^2 \boldsymbol{M}_m \widehat{\boldsymbol{\Sigma}}_m \boldsymbol{M}_m \simeq v(-\lambda;\phi_s)^2(1+\widetilde{v}(-\lambda;\phi_s,\phi_s))\cdot(v(-\lambda;\phi_s)\boldsymbol{\Sigma}+\boldsymbol{I}_p)^{-1}\boldsymbol{\Sigma}(v(-\lambda;\phi_s)\boldsymbol{\Sigma}+\boldsymbol{I}_p)^{-1}. \tag{46}$$

By the definition of deterministic equivalent, we have

$$\begin{aligned}\lambda^2 \boldsymbol{\beta}_0^\top \boldsymbol{M}_m \widehat{\boldsymbol{\Sigma}}_m \boldsymbol{M}_m \boldsymbol{\beta}_0 &\xrightarrow{\text{a.s.}} \lim_{p\to\infty} v(-\lambda;\phi_s)^2(1+\widetilde{v}(-\lambda;\phi_s,\phi_s))\sum_{i=1}^p \frac{r_i}{(1+r_i v(-\lambda;\phi_s))^2}(\boldsymbol{\beta}_0^\top w_i)^2 \\ &= \lim_{p\to\infty}\|\boldsymbol{\beta}_0\|_2^2 v(-\lambda;\phi_s)^2(1+\widetilde{v}(-\lambda;\phi_s,\phi_s))\int \frac{r}{(1+v(-\lambda;\phi_s)r)^2}\,\mathrm{d}G_p(r) \\ &= \rho^2 v(-\lambda;\phi_s)^2(1+\widetilde{v}(-\lambda;\phi_s,\phi_s))\int \frac{r}{(1+v(-\lambda;\phi_s)r)^2}\,\mathrm{d}G(r),\end{aligned} \tag{47}$$

where the last equality holds since G_p and G have compact supports and invoking Assumption 2.2.

Part (2) From Lemma F.8 (1), we have

$$\boldsymbol{M}_l \widehat{\boldsymbol{\Sigma}}_{m\setminus l} \boldsymbol{M}_l \simeq \boldsymbol{M}_l \boldsymbol{\Sigma} \boldsymbol{M}_l.$$

Then, from Patil et al. (2022a, Lemma S.2.4), the conclusion follows.

Part (3) For the cross term, it suffices to derive the deterministic equivalent of $\boldsymbol{\beta}_0^\top \boldsymbol{M}_1 \widehat{\boldsymbol{\Sigma}}_{1\cup 2} \boldsymbol{M}_2 \boldsymbol{\beta}_0$. We begin with analyzing the deterministic equivalent of $\boldsymbol{M}_1 \widehat{\boldsymbol{\Sigma}}_{1\cup 2} \boldsymbol{M}_2$. Let $i_0 = \operatorname{tr}(\boldsymbol{L}_1 \boldsymbol{L}_2)$ be the number of shared samples between $\mathcal{D}_{I_1}$ and $\mathcal{D}_{I_2}$, we use the decomposition

$$\boldsymbol{M}_j^{-1} = \frac{i_0}{k}(\widehat{\boldsymbol{\Sigma}}_0 + \lambda \boldsymbol{I}_p) + \frac{k - i_0}{k}(\widehat{\boldsymbol{\Sigma}}_j^{\text{ind}} + \lambda \boldsymbol{I}_p), \qquad j = 1, 2,$$

where $\widehat{\boldsymbol{\Sigma}}_0 = \boldsymbol{X}^\top \boldsymbol{L}_1 \boldsymbol{L}_2 \boldsymbol{X}/i_0$ and $\widehat{\boldsymbol{\Sigma}}_j^{\text{ind}} = \boldsymbol{X}^\top (\boldsymbol{L}_j - \boldsymbol{L}_1 \boldsymbol{L}_2) \boldsymbol{X}/(k - i_0)$ are the common and individual covariance estimators of the two datasets. Let $\boldsymbol{N}_0 = (\widehat{\boldsymbol{\Sigma}}_0 + \lambda \boldsymbol{I}_p)^{-1}$ and $\boldsymbol{N}_j = (\widehat{\boldsymbol{\Sigma}}_j^{\text{ind}} + \lambda \boldsymbol{I}_p)^{-1}$ for $j = 1, 2$. Then

$$\boldsymbol{M}_j = \left(\frac{i_0}{k} \boldsymbol{N}_0^{-1} + \frac{k - i_0}{k} \boldsymbol{N}_j^{-1} \right)^{-1}, \qquad j = 1, 2, \tag{48}$$

where the equalities hold because $\boldsymbol{N}_0$ is invertible when $\lambda > 0$. Note that

$$\begin{aligned} \widehat{\boldsymbol{\Sigma}}_{1\cup 2} &= \frac{k}{2k - i_0} \widehat{\boldsymbol{\Sigma}}_1 + \frac{k}{2k - i_0} \widehat{\boldsymbol{\Sigma}}_2 - \frac{i_0}{2k - i_0} \widehat{\boldsymbol{\Sigma}}_0, \\ &= \frac{k}{2k - i_0} \sum_{j=1}^{2} (\boldsymbol{M}_j^{-1} - \lambda \boldsymbol{I}_p) - \frac{i_0}{2k - i_0} \widehat{\boldsymbol{\Sigma}}_0. \end{aligned}$$

We have that

$$\lambda^2 \boldsymbol{M}_1 \widehat{\boldsymbol{\Sigma}}_{1\cup 2} \boldsymbol{M}_2 = \frac{k}{2k - i_0} \lambda^2 \sum_{j=1}^{2} \boldsymbol{M}_j - \frac{2k}{2k - i_0} \lambda^3 \boldsymbol{M}_1 \boldsymbol{M}_2 - \frac{i_0}{2k - i_0} \lambda^2 \boldsymbol{M}_1 \widehat{\boldsymbol{\Sigma}}_0 \boldsymbol{M}_2. \tag{49}$$

Next, we derive the deterministic equivalents for the three terms in (49). From Corollary F.5, the first term admits

$$\lambda \boldsymbol{M}_j \simeq (v(-\lambda; \phi_s) \boldsymbol{\Sigma} + \boldsymbol{I}_p)^{-1}. \tag{50}$$

Note that

$$\begin{aligned} \lambda^2 \boldsymbol{M}_1 \boldsymbol{M}_2 &\simeq (v(-\lambda; \phi_s) \boldsymbol{\Sigma} + \boldsymbol{I}_p)^{-1} \left(\widetilde{v}(-\lambda; \phi, \phi_s, \boldsymbol{I}_p) \boldsymbol{\Sigma} + \boldsymbol{I}_p \right) (v(-\lambda; \phi_s) \boldsymbol{\Sigma} + \boldsymbol{I}_p)^{-1} &(51) \\ &\simeq (v(-\lambda; \phi_s) \boldsymbol{\Sigma} + \boldsymbol{I}_p)^{-2} \left(\widetilde{v}(-\lambda; \phi, \phi_s, \boldsymbol{I}_p) \boldsymbol{\Sigma} + \boldsymbol{I}_p \right), &(52) \end{aligned}$$

where

$$\widetilde{v}(-\lambda; \phi, \phi_s, \boldsymbol{I}_p) = \frac{\phi \int \frac{r}{(1 + v(-\lambda; \phi_s) r)^2} \, \mathrm{d}H(r)}{v(-\lambda; \phi_s)^{-2} - \phi \int \frac{r^2}{(1 + v(-\lambda; \phi_s) r)^2} \, \mathrm{d}H(r)}.$$

For the third term,

$$\boldsymbol{M}_1 \widehat{\boldsymbol{\Sigma}}_0 \boldsymbol{M}_2 \simeq \widetilde{v}_v(-\lambda; \phi, \phi_s) (v(-\lambda; \phi_s) \boldsymbol{\Sigma} + \boldsymbol{I}_p)^{-2} \boldsymbol{\Sigma}, \tag{53}$$

where

$$\widetilde{v}_v(-\lambda; \phi, \phi_s) := \frac{1}{v(-\lambda; \phi_s)^{-2} - \phi \int \frac{r^2}{(1 + v(-\lambda; \phi_s) r)^2} \, \mathrm{d}H(r)}.$$

Combining (50)-(53), we get

$$\lambda^2 \boldsymbol{M}_1 \widehat{\boldsymbol{\Sigma}}_{1\cup 2} \boldsymbol{M}_2 \simeq \left(\frac{2\phi_s}{2\phi_s - \phi}(v(-\lambda;\phi_s) - \widetilde{v}(-\lambda;\phi,\phi_s,\boldsymbol{I}_p)) - \frac{\phi}{2\phi_s-\phi}\lambda\widetilde{v}_v(-\lambda;\phi,\phi_s) \right) \lambda(v(-\lambda;\phi_s)\boldsymbol{\Sigma} + \boldsymbol{I}_p)^{-2}\boldsymbol{\Sigma}$$
$$= \lambda^2 v(-\lambda;\phi_s)^2(1+\widetilde{v}(-\lambda;\phi,\phi_s)) \left(\frac{2(\phi_s-\phi)}{2\phi_s-\phi}\frac{1}{\lambda v(-\lambda;\phi_s)} + \frac{\phi}{2\phi_s-\phi} \right) (v(-\lambda;\phi_s)\boldsymbol{\Sigma} + \boldsymbol{I}_p)^{-2}\boldsymbol{\Sigma}. \tag{54}$$

The last conclusion then follows analogously as in (47). □

D.5.2. Deterministic approximation of the variance functional

Lemma D.7 (Deterministic approximation of the variance functional)**.** *Under Assumptions 2.1-2.2, for all $m \in [M]$ and $I_m \in \mathcal{I}_k$, let $\widehat{\boldsymbol{\Sigma}}_m = \boldsymbol{X}^\top \boldsymbol{L}_m \boldsymbol{X}/k$, $\boldsymbol{L}_m \in \mathbb{R}^{n\times n}$ be a diagonal matrix with $(\boldsymbol{L}_m)_{ll} = 1$ if $l \in I_m$ and 0 otherwise, and $\boldsymbol{M}_m = (\boldsymbol{X}^\top \boldsymbol{L}_m \boldsymbol{X}/k + \lambda \boldsymbol{I}_p)^{-1}$. Then, it holds that*

1. *For all $m \in [M]$ and $I_m \in \mathcal{I}_k$,*

$$1 - \frac{2}{k}\operatorname{tr}(\boldsymbol{M}_m\widehat{\boldsymbol{\Sigma}}_m) + \frac{1}{k}\operatorname{tr}(\boldsymbol{M}_m\widehat{\boldsymbol{\Sigma}}_m\boldsymbol{M}_m\widehat{\boldsymbol{\Sigma}}_m) \xrightarrow{\text{a.s.}} \lambda^2 v(-\lambda;\phi_s)^2(1+\widetilde{v}(-\lambda;\phi_s,\phi_s)). \tag{55}$$

2. *For all $m,l \in [M]$, $m \neq l$ and $I_m, I_l \overset{\text{SRSWR}}{\sim} \mathcal{I}_k$,*

$$\frac{1}{k}\operatorname{tr}(\boldsymbol{M}_m\widehat{\boldsymbol{\Sigma}}_m\boldsymbol{M}_m\widehat{\boldsymbol{\Sigma}}_{l\setminus m}) \xrightarrow{\text{a.s.}} \widetilde{v}(-\lambda;\phi_s,\phi_s). \tag{56}$$

3. *For all $m,l \in [M]$, $m \neq l$ and $I_m, I_l \overset{\text{SRSWR}}{\sim} \mathcal{I}_k$,*

$$1 - \frac{1}{|I_m \cup I_l|}\sum_{j\in\{m,l\}}\operatorname{tr}(\boldsymbol{M}_j\widehat{\boldsymbol{\Sigma}}_j) + \frac{1}{n}\operatorname{tr}(\boldsymbol{M}_l\widehat{\boldsymbol{\Sigma}}_{m\cap l}\boldsymbol{M}_m\widehat{\boldsymbol{\Sigma}}_{m\cup l}) \xrightarrow{\text{a.s.}}$$
$$\mathscr{D}^\lambda(\phi_s,\phi_s)\left(\frac{2(\phi_s-\phi)}{2\phi_s-\phi}\frac{1}{\lambda v(-\lambda;\phi_s)} + \frac{\phi}{2\phi_s-\phi}\right)(1+\widetilde{v}(-\lambda;\phi,\phi_s)), \tag{57}$$

where $\widehat{\boldsymbol{\Sigma}}_{l\setminus m} = |I_l \setminus I_m|^{-1}\boldsymbol{X}^\top \boldsymbol{L}_{l\setminus m}\boldsymbol{X}$ and $\widehat{\boldsymbol{\Sigma}}_{m\cup l} = |I_m \cup I_l|^{-1}\boldsymbol{X}^\top \boldsymbol{L}_{m\cup l}\boldsymbol{X}$, as $n,k,p \to \infty$, $p/n \to \phi \in (0,\infty)$, and $p/k \to \phi_s \in [\phi,\infty)$, where the nonnegative constant $\widetilde{v}(\lambda;\phi,\phi_s)$ is as defined in (19).

Proof of Lemma D.7. We split the proof into different parts.

Part (1) Note that

$$\operatorname{tr}(\boldsymbol{M}_m\widehat{\boldsymbol{\Sigma}}_m\boldsymbol{M}_m\widehat{\boldsymbol{\Sigma}}_m) = \operatorname{tr}(\boldsymbol{M}_m\widehat{\boldsymbol{\Sigma}}_m) - \lambda\operatorname{tr}(\boldsymbol{M}_n^2\widehat{\boldsymbol{\Sigma}}_m).$$

We now have

$$1 - \frac{2}{k}\operatorname{tr}(\boldsymbol{M}_m\widehat{\boldsymbol{\Sigma}}_m) + \frac{1}{k}\operatorname{tr}(\boldsymbol{M}_m\widehat{\boldsymbol{\Sigma}}_m\boldsymbol{M}_m\widehat{\boldsymbol{\Sigma}}_m) = 1 - \frac{1}{k}\operatorname{tr}(\boldsymbol{M}_m\widehat{\boldsymbol{\Sigma}}_m) - \frac{\lambda}{k}\operatorname{tr}(\boldsymbol{M}_m^2\widehat{\boldsymbol{\Sigma}}_m)$$
$$= 1 - \frac{p}{k} + \frac{\lambda}{k}\operatorname{tr}(\boldsymbol{M}_m) - \frac{\lambda}{k}\operatorname{tr}(\boldsymbol{M}_m^2\widehat{\boldsymbol{\Sigma}}_m).$$

From Corollary F.5 we have that $\lambda\boldsymbol{M}_m \simeq (v(-\lambda;\phi_s)\boldsymbol{\Sigma} + \boldsymbol{I}_p)^{-1}$. From Lemma F.6 (2) (with $\boldsymbol{A} = \boldsymbol{I}$), we have that for $j \in [M]$,

$$\boldsymbol{M}_m\widehat{\boldsymbol{\Sigma}}_m\boldsymbol{M}_m \simeq \widetilde{v}_v(-\lambda;\phi_s)(v(-\lambda;\phi_s)\boldsymbol{\Sigma} + \boldsymbol{I}_p)^{-2}\boldsymbol{\Sigma}. \tag{58}$$

By the trace rule Lemma F.3 (4) , we have

$$\begin{aligned}\frac{\lambda}{k}\operatorname{tr}(\boldsymbol{M}_m)-\frac{\lambda}{k}\operatorname{tr}(\boldsymbol{M}_m^2\widehat{\boldsymbol{\Sigma}}_m) &\xrightarrow{\text{a.s.}} \lim_{p\to\infty}\frac{p}{k}\cdot\frac{1}{p}\left(\operatorname{tr}((v(-\lambda;\phi_s)\boldsymbol{\Sigma}+\boldsymbol{I}_p)^{-1})-\operatorname{tr}(\lambda\widetilde{v}_v(-\lambda;\phi_s)(v(-\lambda;\phi_s)\boldsymbol{\Sigma}+\boldsymbol{I}_p)^{-2}\boldsymbol{\Sigma})\right)\\ &=\phi_s\lim_{p\to\infty}\frac{1}{p}\sum_{i=1}^{p}\frac{1+v(-\lambda;\phi_s)r_i-\lambda\widetilde{v}_v(-\lambda;\phi_s)r_i}{(v(-\lambda;\phi_s)r_i+1)^2}\\ &=\phi_s\lim_{p\to\infty}\int\frac{1+v(-\lambda;\phi_s)r-\lambda\widetilde{v}_v(-\lambda;\phi_s)r}{(1+v(-\lambda;\phi_s)r)^2}\,\mathrm{d}H_p(r)\\ &=\phi_s\int\frac{1+v(-\lambda;\phi_s)r-\lambda\widetilde{v}_v(-\lambda;\phi_s)r}{(1+v(-\lambda;\phi_s)r)^2}\,\mathrm{d}H(r),\qquad j=1,2,\end{aligned}\tag{59}$$

where in the last line we used the fact that H_p and H have compact supports and Assumption 2.2. Then, we have

$$\begin{aligned}&1-\frac{2}{k}\operatorname{tr}(\boldsymbol{M}_m\widehat{\boldsymbol{\Sigma}}_m)+\frac{1}{k}\operatorname{tr}(\boldsymbol{M}_m\widehat{\boldsymbol{\Sigma}}_m\boldsymbol{M}_m\widehat{\boldsymbol{\Sigma}}_m)\\ &\xrightarrow{\text{a.s.}}1-\phi_s+\phi_s\int\frac{1+v(-\lambda;\phi_s)r-\lambda\widetilde{v}_v(-\lambda;\phi_s)r}{(1+v(-\lambda;\phi_s)r)^2}\,\mathrm{d}H(r)\\ &=1-\phi_s\int\frac{v(-\lambda;\phi_s)r}{1+v(-\lambda;\phi_s)r}\,\mathrm{d}H(r)-\phi_s\int\frac{\lambda\widetilde{v}_v(-\lambda;\phi_s)r}{(1+v(-\lambda;\phi_s)r)^2}\,\mathrm{d}H(r)\\ &=\lambda v(-\lambda;\phi_s)-\phi_s\int\frac{\lambda\widetilde{v}_v(-\lambda;\phi_s)r}{(1+v(-\lambda;\phi_s)r)^2}\,\mathrm{d}H(r)\\ &=\lambda\widetilde{v}_v(-\lambda;\phi_s)\left(v(-\lambda;\phi_s)^{-1}-\phi_s\int\frac{v(-\lambda;\phi_s)r^2}{(1+v(-\lambda;\phi_s)r)^2}\,\mathrm{d}H(r)-\phi_s\int\frac{r}{(1+v(-\lambda;\phi_s)r)^2}\,\mathrm{d}H(r)\right)\\ &=\lambda\widetilde{v}_v(-\lambda;\phi_s)\left(v(-\lambda;\phi_s)^{-1}-\phi_s\int\frac{r}{1+v(-\lambda;\phi_s)r}\,\mathrm{d}H(r)\right)\\ &=\lambda^2\widetilde{v}_v(-\lambda;\phi_s)\\ &=\lambda^2v(-\lambda;\phi_s)^2(1+\widetilde{v}(-\lambda;\phi_s,\phi_s)),\end{aligned}$$

and thus, (55) follows.

Part (2) Since $\boldsymbol{M}_m\widehat{\boldsymbol{\Sigma}}_m\boldsymbol{M}_m$ and $\widehat{\boldsymbol{\Sigma}}_{l\setminus m}$ are independent, from Lemma F.8 (1), we have

$$\boldsymbol{M}_m\widehat{\boldsymbol{\Sigma}}_m\boldsymbol{M}_m\widehat{\boldsymbol{\Sigma}}_{l\setminus m}\simeq\boldsymbol{M}_m\widehat{\boldsymbol{\Sigma}}_m\boldsymbol{M}_m\boldsymbol{\Sigma}.$$

Then, by the definition of deterministic equivalents, it follows that

$$\frac{1}{k}\operatorname{tr}(\boldsymbol{M}_m\widehat{\boldsymbol{\Sigma}}_m\boldsymbol{M}_m\widehat{\boldsymbol{\Sigma}}_{l\setminus m})=\frac{1}{k}\operatorname{tr}(\boldsymbol{M}_m\widehat{\boldsymbol{\Sigma}}_m\boldsymbol{M}_m\boldsymbol{\Sigma})\xrightarrow{\text{a.s.}}\widetilde{v}(-\lambda;\phi_s,\phi_s),$$

where the convergence is due to Patil et al. (2022a, Lemma S.2.5).

Part (3) Let $i_0=|I_m\cap I_l|$. The first two terms in (59) satisfy that

$$1-\frac{1}{|I_m\cup I_l|}\sum_{j\in\{m,l\}}\operatorname{tr}(\boldsymbol{M}_j\widehat{\boldsymbol{\Sigma}}_j)=1-\frac{p}{2k-i_0}\sum_{j\in\{m,l\}}\frac{1}{p}\operatorname{tr}(\boldsymbol{M}_j\widehat{\boldsymbol{\Sigma}}_j)\xrightarrow{\text{a.s.}}1-\frac{2\phi_s^2}{2\phi_s-\phi}\int\frac{v(-\lambda;\phi_s)r}{1+v(-\lambda;\phi_s)r}\,\mathrm{d}H(r),$$

where the convergence is from Part 1. The last term can be further decomposed because

$$\begin{aligned}\operatorname{tr}(\boldsymbol{M}_l\widehat{\boldsymbol{\Sigma}}_{m\cap l}\boldsymbol{M}_m\widehat{\boldsymbol{\Sigma}}_{m\cup l})&=\frac{k}{2k-i_0}\sum_{j\in\{m,l\}}\operatorname{tr}(\boldsymbol{M}_l\widehat{\boldsymbol{\Sigma}}_{m\cap l}\boldsymbol{M}_m\widehat{\boldsymbol{\Sigma}}_j)-\frac{i_0}{2k-i_0}\operatorname{tr}(\boldsymbol{M}_l\widehat{\boldsymbol{\Sigma}}_{m\cap l}\boldsymbol{M}_m\widehat{\boldsymbol{\Sigma}}_{m\cap l})\\ &=\frac{k}{2k-i_0}\sum_{j\in\{m,l\}}[\operatorname{tr}(\boldsymbol{M}_j\widehat{\boldsymbol{\Sigma}}_{m\cap l})-\lambda\operatorname{tr}(\boldsymbol{M}_m\widehat{\boldsymbol{\Sigma}}_{m\cap l}\boldsymbol{M}_l)]-\frac{i_0}{2k-i_0}\operatorname{tr}(\boldsymbol{M}_l\widehat{\boldsymbol{\Sigma}}_{m\cap l}\boldsymbol{M}_m\widehat{\boldsymbol{\Sigma}}_{m\cap l}).\end{aligned}$$

From Lemma F.8 (4), (3), and (5), we have

$$\begin{aligned}
\boldsymbol{M}_j\widehat{\boldsymbol{\Sigma}}_{m\cap l} &\simeq \boldsymbol{I}_p - (v(-\lambda;\phi_s)\boldsymbol{\Sigma}+\boldsymbol{I}_p)^{-1}\\
\boldsymbol{M}_m\widehat{\boldsymbol{\Sigma}}_{m\cap l}\boldsymbol{M}_l &\simeq \widetilde{v}_v(-\lambda;\phi,\phi_s)(v(-\lambda;\phi_s)\boldsymbol{\Sigma}+\boldsymbol{I}_p)^{-2}\boldsymbol{\Sigma}\\
\boldsymbol{M}_l\widehat{\boldsymbol{\Sigma}}_{m\cap l}\boldsymbol{M}_m\widehat{\boldsymbol{\Sigma}}_{m\cap l} &\simeq \left(\frac{\phi_s}{\phi}v(-\lambda;\phi_s) - \frac{\phi_s-\phi}{\phi}\lambda\widetilde{v}_v(-\lambda;\phi,\phi_s)\right)(v(-\lambda;\phi_s)\boldsymbol{\Sigma}+\boldsymbol{I}_p)^{-1}\boldsymbol{\Sigma}\\
&\quad - \lambda\widetilde{v}_v(-\lambda;\phi,\phi_s)(v(-\lambda;\phi_s)\boldsymbol{\Sigma}+\boldsymbol{I}_p)^{-2}\boldsymbol{\Sigma}.
\end{aligned}$$

Combining the above terms and by Assumption 2.2, we have

$$\begin{aligned}
&1 - \frac{1}{|I_m\cup I_l|}\sum_{j\in\{m,l\}}\operatorname{tr}(\boldsymbol{M}_j\widehat{\boldsymbol{\Sigma}}_j) + \frac{1}{n}\operatorname{tr}(\boldsymbol{M}_l\widehat{\boldsymbol{\Sigma}}_{m\cap l}\boldsymbol{M}_m\widehat{\boldsymbol{\Sigma}}_{m\cup l})\\
&\xrightarrow{\text{a.s.}} 1 - \frac{2\phi_s^2}{2\phi_s-\phi}\int\frac{v(-\lambda;\phi_s)r}{1+v(-\lambda;\phi_s)r}\,\mathrm{d}H(r)\\
&\quad + \phi\frac{2\phi_s}{2\phi_s-\phi}\left(\int\frac{v(-\lambda;\phi_s)r}{1+v(-\lambda;\phi_s)r}\,\mathrm{d}H(r) - \lambda\widetilde{v}_v(-\lambda;\phi,\phi_s)\int\frac{r}{(1+v(-\lambda;\phi_s)r)^2}\,\mathrm{d}H(r)\right)\\
&\quad - \phi\frac{\phi}{2\phi_s-\phi}\left(\frac{\phi_s}{\phi}\int\frac{r}{1+v(-\lambda;\phi_s)r}\,\mathrm{d}H(r) - \frac{\phi_s-\phi}{\phi}\lambda\widetilde{v}_v(-\lambda;\phi,\phi_s)\int\frac{r}{1+v(-\lambda;\phi_s)r}\,\mathrm{d}H(r)\right.\\
&\quad\quad \left.-\lambda\widetilde{v}_v(-\lambda;\phi,\phi_s)\int\frac{r}{(1+v(-\lambda;\phi_s)r)^2}\,\mathrm{d}H(r)\right)\\
&= 1 - \phi_s\int\frac{v(-\lambda;\phi_s)r}{1+v(-\lambda;\phi_s)r}\,\mathrm{d}H(r) + \frac{\phi(\phi_s-\phi)}{2\phi_s-\phi}\lambda\widetilde{v}_v(-\lambda;\phi,\phi_s)\int\frac{r}{1+v(-\lambda;\phi_s)r}\,\mathrm{d}H(r)\\
&\quad - \phi\lambda\widetilde{v}_v(-\lambda;\phi,\phi_s)\int\frac{r}{(1+v(-\lambda;\phi_s)r)^2}\,\mathrm{d}H(r)\\
&= \lambda v(-\lambda;\phi_s) + \frac{\phi(\phi_s-\phi)}{2\phi_s-\phi}\int\frac{\lambda\widetilde{v}_v(-\lambda;\phi,\phi_s)r}{1+v(-\lambda;\phi_s)r}\,\mathrm{d}H(r) - \phi\int\frac{\lambda\widetilde{v}_v(-\lambda;\phi,\phi_s)r}{(1+v(-\lambda;\phi_s)r)^2}\,\mathrm{d}H(r)\\
&= \frac{\lambda^2\widetilde{v}_v(-\lambda;\phi,\phi_s)}{2\phi_s-\phi}\left(\frac{2\phi_s-\phi}{\lambda v(-\lambda;\phi_s)} - \frac{\phi(2\phi_s-\phi)}{\lambda}\int\frac{v(-\lambda;\phi_s)r^2}{(1+v(-\lambda;\phi_s)r)^2}\,\mathrm{d}H(r)\right.\\
&\quad \left.+\frac{\phi(\phi_s-\phi)}{\lambda}\int\frac{r}{1+v(-\lambda;\phi_s)r}\,\mathrm{d}H(r) - \frac{\phi}{\lambda}\int\frac{r}{(1+v(-\lambda;\phi_s)r)^2}\,\mathrm{d}H(r)\right)\\
&= \frac{\lambda^2\widetilde{v}_v(-\lambda;\phi,\phi_s)}{2\phi_s-\phi}\left(\frac{2\phi_s-\phi}{\lambda v(-\lambda;\phi_s)} - \frac{\phi(2\phi_s-\phi)}{\lambda}\int\frac{r}{1+v(-\lambda;\phi_s)r}\,\mathrm{d}H(r)\right.\\
&\quad \left.+\frac{\phi(\phi_s-\phi)}{\lambda}\int\frac{r}{1+v(-\lambda;\phi_s)r}\,\mathrm{d}H(r)\right)\\
&= \lambda^2\widetilde{v}_v(-\lambda;\phi,\phi_s)^2\left(\frac{2(\phi_s-\phi)}{2\phi_s-\phi}\frac{1}{\lambda v(-\lambda;\phi_s)} + \frac{\phi}{2\phi_s-\phi}\right)\\
&= \mathscr{D}^\lambda(\phi_s,\phi_s)\left(\frac{2(\phi_s-\phi)}{2\phi_s-\phi}\frac{1}{\lambda v(-\lambda;\phi_s)} + \frac{\phi}{2\phi_s-\phi}\right)(1+\widetilde{v}(-\lambda;\phi,\phi_s)).
\end{aligned}$$

□

E. Proof of Proposition 3.6

Proof of Proposition 3.6. From Lemma D.1, we have

$$\begin{aligned}
\frac{1}{n}\|\boldsymbol{y}-\boldsymbol{X}\widetilde{\boldsymbol{\beta}}_k^\lambda\|_2^2 = &-\frac{1}{n}\mathbb{E}_{I\overset{\text{SRS}}{\sim}\mathcal{I}_k}\left[\operatorname{Err}_{\text{train}}(\widehat{\boldsymbol{\beta}}_k^\lambda(\mathcal{D}_I)) + \operatorname{Err}_{\text{test}}(\widehat{\boldsymbol{\beta}}_k^\lambda(\{\mathcal{D}_I\}))\right]\\
&+\frac{2}{n}\mathbb{E}_{(I_m,I_\ell)\overset{\text{SRS}}{\sim}\mathcal{I}_k}\left[\operatorname{Err}_{\text{train}}(\widetilde{\boldsymbol{\beta}}_k^\lambda(\{\mathcal{D}_{I_m},\mathcal{D}_{I_\ell}\})) + \operatorname{Err}_{\text{test}}(\widetilde{\boldsymbol{\beta}}_k^\lambda(\{\mathcal{D}_{I_m},\mathcal{D}_{I_\ell}\}))\right].
\end{aligned}$$

Since by Lemma D.2, Lemma D.3 and Lemma G.6, each expectation converges, we have that

$$\frac{1}{n}\|\boldsymbol{y}-\boldsymbol{X}\widetilde{\boldsymbol{\beta}}_k^\lambda\|_2^2 \xrightarrow{\text{a.s.}} \frac{2\phi(2\phi_s-\phi)}{\phi_s^2}\mathscr{T}_2^\lambda(\phi,\phi_s)+\frac{2(\phi_s-\phi)^2}{\phi_s^2}\mathscr{R}_2^\lambda(\phi,\phi_s)-\frac{\phi}{\phi_s}\mathscr{T}_1^\lambda(\phi,\phi_s)-\frac{\phi_s-\phi}{\phi_s}\mathscr{R}_1^\lambda(\phi,\phi_s),$$

where the convergence of the averages is from Lemma G.5 and the convergence of coefficients is from Lemma G.2. Since the denominator converges from Lemma 3.4, we further have

$$\mathrm{gcv}_{k,\infty}^\lambda \xrightarrow{\text{a.s.}} \mathscr{G}_\infty^\lambda(\phi,\phi_s)=\frac{\frac{2\phi(2\phi_s-\phi)}{\phi_s^2}\mathscr{T}_2^\lambda(\phi,\phi_s)+\frac{2(\phi_s-\phi)^2}{\phi_s^2}\mathscr{R}_2^\lambda(\phi,\phi_s)-\frac{\phi}{\phi_s}\mathscr{T}_1^\lambda(\phi,\phi_s)-\frac{\phi_s-\phi}{\phi_s}\mathscr{R}_1^\lambda(\phi,\phi_s)}{\mathscr{D}_\infty^\lambda(\phi,\phi_s)},$$

for $\lambda>0$ and $\phi_s\in[\phi,+\infty)$.

For the boundary case when $\lambda>0$ but $\phi_s=+\infty$, we require Proposition E.1; for the boundary case when $\lambda=0$, we require Proposition E.2. Applying Proposition E.1 and Proposition E.2 finishes the proof. □

E.1. Boundary case: diverging subsample aspect ratio for the ridge predictor

Proposition E.1 (Risk approximation when $\phi_s\to+\infty$)**.** *Under Assumptions 2.1-2.2, it holds for all $\lambda>0$,*

$$\mathrm{gcv}_k^\lambda \xrightarrow{\text{a.s.}} \mathscr{G}_\infty^\lambda(\phi,\infty),$$

as $k,n,p\to\infty$, $p/n\to\phi\in(0,\infty)$ and $p/k\to\infty$, where $\mathscr{G}_\infty^\lambda(\cdot,\cdot)$ is defined in Proposition 3.6.

Proof of Proposition E.1. Recall that

$$\frac{1}{n}\|\boldsymbol{y}-\boldsymbol{X}\widetilde{\boldsymbol{\beta}}_k^\lambda\|_2^2=\lim_{M\to\infty}\frac{1}{n}\|\boldsymbol{y}-\boldsymbol{X}\widetilde{\boldsymbol{\beta}}_{k,M}^\lambda\|_2^2=(\boldsymbol{\beta}_0-\widetilde{\boldsymbol{\beta}}_k^\lambda)^\top\widehat{\boldsymbol{\Sigma}}(\boldsymbol{\beta}_0-\widetilde{\boldsymbol{\beta}}_k^\lambda)+\frac{1}{n}\boldsymbol{\epsilon}^\top\boldsymbol{\epsilon}+\frac{2}{n}(\boldsymbol{\beta}_0-\widetilde{\boldsymbol{\beta}}_k^\lambda)^\top\widehat{\boldsymbol{\Sigma}}\boldsymbol{\epsilon}$$

From Lemma G.3 and Lemma G.4, we have that $\boldsymbol{\beta}_0^\top\boldsymbol{X}^\top\boldsymbol{\epsilon}/n\xrightarrow{\text{a.s.}}0$ and $\boldsymbol{\epsilon}^\top\boldsymbol{\epsilon}/n\xrightarrow{\text{a.s.}}\sigma^2$ as $n\to\infty$. For the other term, note that for any $(I_1,\dots,I_M)\overset{\text{SRS}}{\sim}\mathcal{I}_k$,

$$\begin{aligned}
\|\widetilde{\boldsymbol{\beta}}_k^\lambda\|_2 &\le \lim_{M\to\infty}\mathbb{E}_{(I_1,\dots,I_M)\overset{\text{SRS}}{\sim}\mathcal{I}_k}\left[\frac{1}{M}\sum_{m=1}^M\|(\boldsymbol{X}^\top\boldsymbol{L}_m\boldsymbol{X}/k+\lambda\boldsymbol{I}_p)^{-1}(\boldsymbol{X}^\top\boldsymbol{L}_m\boldsymbol{y}/k)\|_2\right]\\
&\le \lim_{M\to\infty}\mathbb{E}_{(I_1,\dots,I_M)\overset{\text{SRS}}{\sim}\mathcal{I}_k}\left[\frac{1}{M}\sum_{m=1}^M\|(\boldsymbol{X}^\top\boldsymbol{L}_m\boldsymbol{X}/k+\lambda\boldsymbol{I}_p)^{-1}\boldsymbol{X}^\top\boldsymbol{L}_m/\sqrt{k}\|\cdot\|\boldsymbol{L}_m\boldsymbol{y}/\sqrt{k}\|_2\right]\\
&\le C\sqrt{\rho^2+\sigma^2}\cdot\max_{I_m\in\mathcal{I}_k}\|(\boldsymbol{X}^\top\boldsymbol{L}_m\boldsymbol{X}/k+\lambda\boldsymbol{I}_p)^{-1}\boldsymbol{X}^\top\boldsymbol{L}_m/\sqrt{k}\|_{\mathrm{op}},
\end{aligned}$$

where the last inequality holds eventually almost surely since Assumptions 2.1-2.2 imply that the entries of $\boldsymbol{y}$ have bounded 4-th moment, and thus from the strong law of large numbers, $\|\boldsymbol{L}_m\boldsymbol{y}/\sqrt{k}\|_2$ is eventually almost surely bounded above by $C\sqrt{\mathbb{E}[y_1^2]}=C\sqrt{\rho^2+\sigma^2}$ for some constant C. Observe that operator norm of the matrix $(\boldsymbol{X}^\top\boldsymbol{L}_m\boldsymbol{X}/k+\lambda\boldsymbol{I}_p)^{-1}\boldsymbol{X}\boldsymbol{L}_m/\sqrt{k}$ is upper bounded $\max_i s_i/(s_i^2+\lambda)\le 1/s_{\min}$ where s_i's are the singular values of $\boldsymbol{X}$ and $s_{\min}$ is the smallest nonzero singular value. As $k,p\to\infty$ such that $p/k\to\infty$, $s_{\min}\to\infty$ almost surely (e.g., from results of Bloemendal et al. (2016)) and therefore, $\|\widetilde{\boldsymbol{\beta}}_k^\lambda\|_2\to 0$ almost surely. Because $\|\widehat{\boldsymbol{\Sigma}}\|_{\mathrm{op}}$ is upper bounded almost surely, we further have $\widetilde{\boldsymbol{\beta}}_k^\top\boldsymbol{X}^\top\boldsymbol{\epsilon}/n\xrightarrow{\text{a.s.}}0$. Consequently we have $(\boldsymbol{\beta}_0-\widetilde{\boldsymbol{\beta}}_k)^\top\boldsymbol{X}^\top\boldsymbol{\epsilon}/n\xrightarrow{\text{a.s.}}0$ and

$$\frac{1}{n}\|\boldsymbol{y}-\boldsymbol{X}\widetilde{\boldsymbol{\beta}}_k^\lambda\|_2^2\xrightarrow{\text{a.s.}}\boldsymbol{\beta}_0^\top\widehat{\boldsymbol{\Sigma}}\boldsymbol{\beta}_0+\sigma^2.$$

Finally, from Lemma F.8 (1) $\boldsymbol{\beta}_0^\top\widehat{\boldsymbol{\Sigma}}\boldsymbol{\beta}_0\xrightarrow{\text{a.s.}}\boldsymbol{\beta}_0^\top\boldsymbol{\Sigma}\boldsymbol{\beta}_0$ and from Assumption 2.2, we have

$$\frac{1}{n}\|\boldsymbol{y}-\boldsymbol{X}\widetilde{\boldsymbol{\beta}}_k^\lambda\|_2^2\xrightarrow{\text{a.s.}}\sigma^2+\rho^2\int r\,\mathrm{d}G(r).$$

Since $\boldsymbol{S}_k^\lambda=\boldsymbol{X}\widetilde{\boldsymbol{\beta}}_k^\lambda$, we have that $\mathrm{tr}(\boldsymbol{S}_k^\lambda)/n\xrightarrow{\text{a.s.}}0$. So the denominator converges to 1 almost surely.

From Lemma F.10, we have $\mathscr{G}_\infty^\lambda(\phi,\infty):=\lim_{\phi_s\to+\infty}\mathscr{G}_\infty^\lambda(\phi,\phi_s)=\sigma^2+\rho^2\int r\,\mathrm{d}G(r)$, which is also the limit of the GCV estimate. Thus, $\mathscr{G}_\infty^\lambda(\phi,\infty)$ is well defined and $\mathscr{G}_\infty^\lambda(\phi,\phi_s)$ is right continuous at $\phi_s=+\infty$. □

E.2. Boundary case: the ridgeless predictor

Proposition E.2 (Risk approximation when $\lambda = 0$)**.** *Under Assumptions 2.1-2.2, suppose that the conclusion of Proposition 3.6 holds for $\lambda > 0$. Then it holds that,*

$$\mathrm{gcv}_k^0 \xrightarrow{\text{a.s.}} \mathscr{G}_\infty^0(\phi, \phi_s) := \lim_{\lambda \to 0^+} \mathscr{G}_\infty^\lambda(\phi, \phi_s),$$

as $k, n, p \to \infty$, $p/n \to \phi \in (0, \infty)$ and $p/k \to [\phi, +\infty]$, where $\mathscr{G}_\infty^\lambda(\cdot, \cdot)$ is defined in Proposition 3.6.

Proof of Proposition E.2. We analyze the numerator and the denominator separately.

Part (1) For the denominator, note that

$$P_{n,\lambda} := (1 - \mathrm{tr}(\boldsymbol{S}_k^\lambda)/n)^2 = \lim_{M \to \infty} (1 - \mathrm{tr}(\boldsymbol{S}_{k,M}^\lambda)/n)^2,$$

where $\boldsymbol{S}_k^\lambda = \lim_{M \to \infty} \boldsymbol{S}_{k,M}^\lambda$ is the smoothing matrix. Since $\boldsymbol{S}_k^\lambda \succeq \boldsymbol{0}_{n \times n}$ and

$$\|\boldsymbol{S}_k^\lambda\|_{\mathrm{op}} \le \max_{I_m \in \mathcal{I}_k} \|\boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{L}_m \boldsymbol{X}/k + \lambda \boldsymbol{I}_p)^{-1} \boldsymbol{X}^\top \boldsymbol{L}_m / \sqrt{k}\|_{\mathrm{op}}, \tag{60}$$

which is also upper bounded almost surely from the proof in Proposition E.1 (when $\lambda = 0$, the inverse in the above display is replaced by pseudo-inverse). Thus, we have $P_{n,\lambda}$ is almost surely upper bounded $\lambda \in \Lambda := [0, \lambda_{\max}]$ for any $\lambda_{\max} \in (0, \infty)$ fixed.

Next we inspect the boundedness of the derivative of $P_{n,\lambda}$:

$$\frac{\partial}{\partial \lambda} P_{n,\lambda} = \frac{\partial}{\partial \lambda} \lim_{M \to \infty} \lim_{M \to \infty} (1 - \mathrm{tr}(\boldsymbol{S}_{k,M}^\lambda)/n)^2 =: \frac{\partial}{\partial \lambda} \lim_{M \to \infty} Q_{M,\lambda}.$$

We claim that

$$\frac{\partial}{\partial \lambda} \lim_{M \to \infty} Q_{M,\lambda} = \lim_{M \to \infty} \frac{\partial}{\partial \lambda} Q_{M,\lambda}.$$

To see this, we need to show that $Q_{M,\lambda}$ is equicontinuous in λ over Λ. First we know that $Q_{M,\lambda}$ is differentiable in λ. From (60), we have that $Q_{M,\lambda}$ is uniformly upper bounded over $\lambda \in \Lambda$ almost surely. Note that

$$\frac{\partial}{\partial \lambda} Q_{M,\lambda} = 2(1 - \mathrm{tr}(\boldsymbol{S}_{k,M}^\lambda)) \,\mathrm{tr}\left(\frac{\partial}{\partial \lambda} \boldsymbol{S}_{k,M}^\lambda\right),$$

where

$$\frac{\partial}{\partial \lambda} \boldsymbol{S}_{k,M}^\lambda = \frac{1}{M} \sum_{m=1}^M \boldsymbol{X} \left(\frac{\boldsymbol{X}^\top \boldsymbol{L}_m \boldsymbol{X}}{k} + \lambda \boldsymbol{I}\right)^{-2} \frac{\boldsymbol{X}^\top \boldsymbol{L}_m}{k}.$$

By the similar arguments as in Proposition E.1, we have that $\left\|\partial \boldsymbol{S}_{k,M}^\lambda / \partial \lambda\right\|_{\mathrm{op}}$, and $\|\boldsymbol{S}_{k,M}^\lambda\|_2^2$ are uniformly upper bounded almost surely over Λ, the equicontinuity conclusion follows. Then by Moore-Osgood theorem, we have

$$\frac{\partial}{\partial \lambda} P_{n,\lambda} = \lim_{M \to \infty} 2(1 - \mathrm{tr}(\boldsymbol{S}_{k,M}^\lambda)) \,\mathrm{tr}\left(\frac{\partial}{\partial \lambda} \boldsymbol{S}_{k,M}^\lambda\right)$$

is uniformly upper bounded almost surely over $[0, +\infty]$ independent of λ and M. Therefore, we conclude that $|\partial P_{n,\lambda}/\partial \lambda|$ is upper bounded almost surely over $\lambda \in \Lambda$.

On the other hand, we know that $P_{n,\lambda} \xrightarrow{\text{a.s.}} \mathscr{D}_\infty^\lambda(\phi, \phi_s)$ for $\lambda > 0$. Define $\mathscr{D}^0(\phi, \phi_s) := \lim_{\lambda \to 0^+} \mathscr{D}_\infty^\lambda(\phi, \phi_s)$. When $\lambda = 0$ and $\phi_s > 1$, we know that $\mathscr{D}^0(\phi, \phi_s)$ is well-defined because $v(-\lambda; \phi_s)$ is finite and continuous from Lemma F.12. When $\lambda = 0$ and $\phi_s \in (0, 1]$, from the definition of fixed-point solution (18), we have

$$1 = v(-\lambda; \phi_s)\lambda + \phi_s \int \frac{v(-\lambda; \phi_s) r}{1 + v(-\lambda; \phi_s) r} \,\mathrm{d}H(r).$$

In this case, $v(0;\phi_s) = +\infty$ from Lemma F.12. Let $\lambda \to 0^+$, we have

$$1 = \lim_{\lambda\to 0^+} v(-\lambda;\phi_s)\lambda + \phi_s \lim_{\lambda\to 0^+} \int \frac{v(-\lambda;\phi_s)r}{1+v(-\lambda;\phi_s)r}\,\mathrm{d}H(r) = \lim_{\lambda\to 0^+} v(-\lambda;\phi_s)\lambda + \phi_s.$$

Then we have $\lim_{\lambda\to 0^+} v(-\lambda;\phi_s)\lambda = 1-\phi_s$ and $\mathscr{D}^\lambda_\infty(\phi,\phi_s) = (1-\phi_s)^2$. Thus, $\mathscr{D}^0(\phi,\phi_s)$ is always well-defined.

From Lemma F.12, there exists $M' > 0$ such that the magnitudes of $v(-\lambda;\phi_s)$ and its derivative with respect to λ are continuous and bounded by M' for all $\lambda \in [0,+\infty]$. It follows that $|\mathscr{D}^\lambda_\infty(\phi,\phi_s)|$ and $|\partial\mathscr{D}^\lambda_\infty(\phi,\phi_s)/\partial\lambda|$ are uniformly upper bounded almost surely. From Moore-Osgood theorem and the continuity property from Lemma F.12, we have

$$\lim_{n\to\infty}\lim_{\lambda\to 0^+} P_{n,\lambda} = \lim_{\lambda\to 0^+}\lim_{n\to\infty} P_{n,\lambda} = \lim_{\lambda\to 0^+} \mathscr{D}^\lambda_\infty(\phi,\phi_s) = \mathscr{D}^0(\phi,\phi_s).$$

Part (2) For the numerator, note that

$$P'_{n,\lambda} := \frac{1}{n}\|\boldsymbol{y} - \boldsymbol{X}\widetilde{\boldsymbol{\beta}}^\lambda_k\|_2^2 = \frac{1}{n}\|(\boldsymbol{I}_n - \boldsymbol{S}^\lambda_k)\boldsymbol{y}\|_2^2.$$

Assumptions 2.1-2.2 imply that the entries of $\boldsymbol{y}$ have bounded 4-th moment, and thus from the strong law of large numbers, $\|\boldsymbol{y}/\sqrt{n}\|_2$ is eventually almost surely bounded above by $C\sqrt{\mathbb{E}[y_1^2]} = C\sqrt{\rho^2+\sigma^2}$ for some constant C. On the other hand, $\boldsymbol{S}^\lambda_k \succeq \boldsymbol{0}_{n\times n}$ and $\|\boldsymbol{S}^\lambda_k\|_{\mathrm{op}}$ is also upper bounded almost surely from Part (1). Thus, we have $P'_{n,\lambda}$ is almost surely upper bounded $\lambda \in \Lambda$.

Next we inspect the boundedness of the derivative of $P'_{n,\lambda}$:

$$\begin{aligned}
\frac{\partial}{\partial\lambda}P'_{n,\lambda} &= \frac{2}{n}(\boldsymbol{y} - \boldsymbol{X}\widetilde{\boldsymbol{\beta}}^\lambda_k)^\top \frac{\partial}{\partial\lambda}\boldsymbol{S}^\lambda_k\boldsymbol{y} \\
&= \frac{\partial}{\partial\lambda}\lim_{M\to\infty}\frac{2}{n}(\boldsymbol{y} - \boldsymbol{X}\widetilde{\boldsymbol{\beta}}^\lambda_{k,M})^\top \boldsymbol{S}^\lambda_{k,M}\boldsymbol{y} \\
&= \frac{\partial}{\partial\lambda}\lim_{M\to\infty}\frac{2}{n}\boldsymbol{y}^\top(\boldsymbol{I}_n - \boldsymbol{S}^\lambda_{k,M})\boldsymbol{S}^\lambda_{k,M}\boldsymbol{y} =: \frac{\partial}{\partial\lambda}\lim_{M\to\infty} Q'_{M,\lambda}.
\end{aligned}$$

We claim that

$$\frac{\partial}{\partial\lambda}\lim_{M\to\infty} Q'_{M,\lambda} = \lim_{M\to\infty}\frac{\partial}{\partial\lambda}Q'_{M,\lambda}.$$

To see this, we need to show that $Q'_{M,\lambda}$ is equicontinuous in λ over Λ. First we know that $Q'_{M,\lambda}$ is differentiable in λ. From (60), we have that $Q'_{M,\lambda}$ is uniformly upper bounded over $\lambda \in \Lambda$ almost surely. Similarly, we have

$$\frac{\partial}{\partial\lambda}Q'_{M,\lambda} = \frac{2}{n}\boldsymbol{y}^\top(\boldsymbol{I}_n - 2\boldsymbol{S}^\lambda_{k,M})\frac{\partial}{\partial\lambda}\boldsymbol{S}^\lambda_{k,M}\boldsymbol{y},$$

and

$$\left|\frac{\partial}{\partial\lambda}Q'_{M,\lambda}\right| \le \left\|\boldsymbol{I}_n - 2\boldsymbol{S}^\lambda_{k,M}\right\|_{\mathrm{op}}\left\|\frac{\partial}{\partial\lambda}\boldsymbol{S}^\lambda_{k,M}\right\|_{\mathrm{op}}\frac{1}{n}\|\boldsymbol{y}\|_2^2.$$

and the equicontinuity conclusion follows analogously as in Part (1). Therefore, we conclude that $|\partial P'_{n,\lambda}/\partial\lambda|$ is upper bounded almost surely over $\lambda \in [0,+\infty]$.

On the other hand, we know that $P'_{n,\lambda} \xrightarrow{\text{a.s.}} \mathscr{G}^\lambda_\infty(\phi,\phi_s)$ for $\lambda > 0$. Define $\mathscr{D}^0(\phi,\phi_s)\mathscr{G}^0_\infty(\phi,\phi_s) := \lim_{\lambda\to 0^+}(\mathscr{D}^\lambda_\infty(\phi,\phi_s)\mathscr{G}^\lambda_\infty(\phi,\phi_s)) = \mathscr{D}^0(\phi,\phi_s)\mathscr{R}^0_M(\phi,\phi_s)$, which is well defined from Part (1) and Lemma A.2. From Lemma F.12, there exists $M' > 0$ such that the magnitudes of $v(-\lambda;\phi_s)$, $\widetilde{v}(\lambda;\phi_s,\phi)$ and $\widetilde{c}(\lambda;\phi_s)$, and their derivatives with respect to λ are continuous and bounded by M' for all $\lambda \in [0,+\infty]$. It follows that $|\mathscr{G}^\lambda_\infty(\phi,\phi_s)|$ is upper bounded almost surely. Analogously, we have that $|\partial(\mathscr{D}^\lambda_\infty(\phi,\phi_s)\mathscr{G}^\lambda_\infty(\phi,\phi_s))/\partial\lambda|$ is also upper bounded almost surely on $\lambda \in \Lambda$. From Moore-Osgood theorem and the continuity property from Lemma F.12, we have

$$\lim_{n\to\infty}\lim_{\lambda\to 0^+} P'_{n,\lambda} = \lim_{\lambda\to 0^+}\lim_{n\to\infty} P'_{n,\lambda} = \lim_{\lambda\to 0^+}\mathscr{G}^\lambda_\infty(\phi,\phi_s) = \mathscr{G}^0_\infty(\phi,\phi_s).$$

□

F. Auxiliary results on asymptotic equivalents

F.1. Preliminary background

We use the notion of asymptotic equivalence of sequences of random matrices in various proofs. In this section, we provide a basic review of the related definitions and corresponding calculus rules. See Dobriban & Wager (2018); Dobriban & Sheng (2021); Patil et al. (2022b;a) for more details.

Definition F.1 (Asymptotic equivalence)**.** Consider sequences $\{\boldsymbol{A}_p\}_{p\geq 1}$ and $\{\boldsymbol{B}_p\}_{p\geq 1}$ of (random or deterministic) matrices of growing dimensions. We say that $\boldsymbol{A}_p$ and $\boldsymbol{B}_p$ are equivalent and write $\boldsymbol{A}_p \simeq \boldsymbol{B}_p$ if $\lim_{p\to\infty} |\operatorname{tr}[\boldsymbol{C}_p(\boldsymbol{A}_p - \boldsymbol{B}_p)]| = 0$ almost surely for any sequence of random matrices $\boldsymbol{C}_p$ independent to $\boldsymbol{A}_p$ and $\boldsymbol{B}_p$, with bounded trace norm such that $\limsup_{p\to\infty} \|\boldsymbol{C}_p\|_{\operatorname{tr}} < \infty$ almost surely.

The notion of asymptotic equivalence of two sequences of random matrices from Definition F.1 can be further extended to incorporate conditioning on another sequence of random matrices.

Definition F.2 (Conditional asymptotic equivalence)**.** Consider sequences $\{\boldsymbol{A}_p\}_{p\geq 1}$, $\{\boldsymbol{B}_p\}_{p\geq 1}$ and $\{\boldsymbol{D}_p\}_{p\geq 1}$ of (random or deterministic) matrices of growing dimensions. We say that $\boldsymbol{A}_p$ and $\boldsymbol{B}_p$ are equivalent given $\boldsymbol{D}_p$ and write $\boldsymbol{A}_p \simeq \boldsymbol{B}_p \mid \boldsymbol{D}_p$ if $\lim_{p\to\infty} |\operatorname{tr}[\boldsymbol{C}_p(\boldsymbol{A}_p - \boldsymbol{B}_p)]| = 0$ almost surely conditional on $\{\boldsymbol{D}_p\}_{p\geq 1}$, i.e.,

$$\mathbb{P}\left(\lim_{p\to\infty} |\operatorname{tr}[\boldsymbol{C}_p(\boldsymbol{A}_p - \boldsymbol{B}_p)]| = 0 \,\middle|\, \{\boldsymbol{D}_p\}_{p\geq 1}\right) = 1,$$

for any sequence of random matrices $\boldsymbol{C}_p$ independent to $\boldsymbol{A}_p$ and $\boldsymbol{B}_p$ conditional on $\boldsymbol{D}_p$, with bounded trace norm such that $\limsup \|\boldsymbol{C}_p\|_{\operatorname{tr}} < \infty$ as $p \to \infty$.

Below we summarize the calculus rules for conditional asymptotic equivalence Definition F.2 adapted from Patil et al. (2022a, Lemma S.7.4 and S.7.6).

Lemma F.3 (Calculus of deterministic equivalents)**.** *Let $\boldsymbol{A}_p$, $\boldsymbol{B}_p$, $\boldsymbol{C}_p$ and $\boldsymbol{D}_p$ be sequences of random matrices. The calculus of deterministic equivalents ($\simeq_D$ and $\simeq_R$) satisfies the following properties:*

(1) Equivalence: The relation $\simeq$ is an equivalence relation.

(2) Sum: If $\boldsymbol{A}_p \simeq \boldsymbol{B}_p \mid \boldsymbol{E}_p$ and $\boldsymbol{C}_p \simeq \boldsymbol{D}_p \mid \boldsymbol{E}_p$, then $\boldsymbol{A}_p + \boldsymbol{C}_p \simeq \boldsymbol{B}_p + \boldsymbol{D}_p \mid \boldsymbol{E}_p$.

(3) Product: If $\boldsymbol{A}_p$ has bounded operator norms such that $\limsup_{p\to\infty} \|\boldsymbol{A}_p\|_{\operatorname{op}} < \infty$, $\boldsymbol{A}_p$ is conditional independent to $\boldsymbol{B}_p$ and $\boldsymbol{C}_p$ given $\boldsymbol{E}_p$ for $p \geq 1$, and $\boldsymbol{B}_p \simeq \boldsymbol{C}_p \mid \boldsymbol{E}_p$, then $\boldsymbol{A}_p\boldsymbol{B}_p \simeq \boldsymbol{A}_p\boldsymbol{C}_p \mid \boldsymbol{E}_p$.

(4) Trace: If $\boldsymbol{A}_p \simeq \boldsymbol{B}_p \mid \boldsymbol{E}_p$, then $\operatorname{tr}[\boldsymbol{A}_p]/p - \operatorname{tr}[\boldsymbol{B}_p]/p \to 0$ almost surely when conditioning on $\boldsymbol{E}_p$.

(5) Differentiation: Suppose $f(z, \boldsymbol{A}_p) \simeq g(z, \boldsymbol{B}_p) \mid \boldsymbol{E}_p$ where the entries of f and g are analytic functions in $z \in S$ and S is an open connected subset of $\mathbb{C}$. Suppose for any sequence $\boldsymbol{C}_p$ of deterministic matrices with bounded trace norm we have $|\operatorname{tr}[\boldsymbol{C}_p(f(z, \boldsymbol{A}_p) - g(z, \boldsymbol{B}_p))]| \leq M$ for every p and $z \in S$. Then we have $f'(z, \boldsymbol{A}_p) \simeq g'(z, \boldsymbol{B}_p) \mid \boldsymbol{E}_p$ for every $z \in S$, where the derivatives are taken entrywise with respect to z.

(6) Unconditioning: If $\boldsymbol{A}_p \simeq \boldsymbol{B}_p \mid \boldsymbol{E}_p$, then $\boldsymbol{A}_p \simeq \boldsymbol{B}_p$.

(7) Substitution: Let $v : \mathbb{R}^{p\times p} \to \mathbb{R}$ and $f(v(\boldsymbol{C}), \boldsymbol{C}) : \mathbb{R}^{p\times p} \to \mathbb{R}^{p\times p}$ be a matrix function for matrix $\boldsymbol{C} \in \mathbb{R}^{p\times p}$ and $p \in \mathbb{N}$, that is continuous in the first augment with respect to operator norm. If $v(\boldsymbol{C}) \overset{\text{a.s.}}{=} v(\boldsymbol{D})$ such that $\boldsymbol{C}$ is independent to $\boldsymbol{D}$, then $f(v(\boldsymbol{C}), \boldsymbol{C}) \simeq f(v(\boldsymbol{D}), \boldsymbol{C}) \mid \boldsymbol{C}$.

F.2. Standard ridge resolvents and various extensions

In this section, we collect various asymptotic equivalents. Appendix F.2.1 introduces the basic concepts and definitions. The extended equivalents developed in the work of Patil et al. (2022a) are summarized in Appendix F.2.2. Based on results in Appendices F.2.1 and F.2.2, we prove some useful deterministic equivalent relations in Appendix F.2.3, which are subsequently used in the proof of Lemma D.3 (Lemmas D.6 and D.7).

F.2.1. STANDARD RIDGE RESOLVENTS

The following lemma provides a deterministic equivalent for the standard ridge resolvent and implies Corollary F.5. It is adapted from Theorem 1 of Rubio & Mestre (2011). See also Theorem 3 of Dobriban & Sheng (2021).

Lemma F.4 (Deterministic equivalent for standard ridge resolvent)**.** *Suppose $\boldsymbol{x}_i \in \mathbb{R}^p$, $1 \leq i \leq n$, are i.i.d. random vectors such that each $\boldsymbol{x}_i = \boldsymbol{z}_i \boldsymbol{\Sigma}^{1/2}$, where $\boldsymbol{z}_i$ is a random vector consisting of i.i.d. entries z_{ij}, $1 \leq j \leq p$, satisfying $\mathbb{E}[z_{ij}] = 0$, $\mathbb{E}[z_{ij}^2] = 1$, and $\mathbb{E}[|z_{ij}|^{8+\alpha}] \leq M_\alpha$ for some constants $\alpha > 0$ and $M_\alpha < \infty$, and $\boldsymbol{\Sigma} \in \mathbb{R}^{p \times p}$ is a positive semidefinite matrix satisfying $0 \preceq \boldsymbol{\Sigma} \preceq r_{\max} I_p$ for some constant $r_{\max} < \infty$ (independent of p). Let $\boldsymbol{X} \in \mathbb{R}^{n \times p}$ the concatenated matrix with $\boldsymbol{x}_i^\top$, $1 \leq i \leq n$, as rows, and let $\widehat{\boldsymbol{\Sigma}} \in \mathbb{R}^{p \times p}$ denote the random matrix $\boldsymbol{X}^\top \boldsymbol{X}/n$. Let $\gamma := p/n$. Then, for $z \in \mathbb{C}^+$, as $n, p \to \infty$ such that $0 < \liminf \gamma \leq \limsup \gamma < \infty$, we have*

$$(\widehat{\boldsymbol{\Sigma}} - z\boldsymbol{I}_p)^{-1} \simeq (c(e(z;\gamma))\boldsymbol{\Sigma} - z\boldsymbol{I}_p)^{-1}, \tag{61}$$

where the scalar $c(e(z;\gamma))$ is defined in terms of another scalar $e(z;\gamma)$ by the equation

$$c(e(z;\gamma)) = \frac{1}{1 + \gamma e(z;\gamma)}, \tag{62}$$

and $e(z;\gamma)$ is the unique solution in $\mathbb{C}^+$ to the fixed-point equation

$$e(z;\gamma) = \operatorname{tr}[\boldsymbol{\Sigma}(c(e(z;\gamma))\boldsymbol{\Sigma} - zI_p)^{-1}]/p. \tag{63}$$

Note that both the scalars $c(e(z;\gamma))$ and $e(z;\gamma)$ also implicitly depend on $\boldsymbol{\Sigma}$. For notation brevity, we do not always explicitly indicate this dependence. However, we will be explicit in such dependence for certain extensions to follow. Additionally, observe that one can eliminate $e(z;\gamma)$ in the statement of Lemma F.4 by combining (62) and (63) so that for $z \in \mathbb{C}^+$, one has

$$(\widehat{\boldsymbol{\Sigma}} - z\boldsymbol{I}_p)^{-1} \simeq (c(z;\gamma)\boldsymbol{\Sigma} - z\boldsymbol{I}_p)^{-1},$$

where $c(z)$ is the unique solution in $\mathbb{C}^-$ to the fixed-point equation

$$\frac{1}{c(z;\gamma)} = 1 + \gamma \operatorname{tr}[\boldsymbol{\Sigma}(c(z;\gamma)\boldsymbol{\Sigma} - z\boldsymbol{I}_p)^{-1}]/p.$$

The following corollary is a simple consequence of Lemma F.4, which supplies a deterministic equivalent for the (regularization) scaled ridge resolvent. We will work with a real regularization parameter λ from here on.

Corollary F.5 (Deterministic equivalent for scaled ridge resolvent)**.** *Assume the setting of Lemma F.4. For $\lambda > 0$, we have*

$$\lambda(\widehat{\boldsymbol{\Sigma}} + \lambda\boldsymbol{I}_p)^{-1} \simeq (v(-\lambda;\gamma)\boldsymbol{\Sigma} + \boldsymbol{I}_p)^{-1},$$

where $v(-\lambda;\gamma) > 0$ is the unique solution to the fixed-point equation

$$\frac{1}{v(-\lambda;\gamma)} = \lambda + \gamma \int \frac{r}{1 + v(-\lambda;\gamma)r} \, \mathrm{d}H_n(r). \tag{64}$$

Here H_n is the empirical distribution (supported on $\mathbb{R}_{\geq 0}$) of the eigenvalues of $\boldsymbol{\Sigma}$.

As a side note, the parameter $v(-\lambda;\gamma)$ in Corollary F.5 is also the companion Stieltjes transform of the spectral distribution of the sample covariance matrix $\widehat{\boldsymbol{\Sigma}}$, which is also the Stieltjes transform of the spectral distribution of the gram matrix $\boldsymbol{X}\boldsymbol{X}^\top/n$.

The following lemma uses Corollary F.5 along with calculus of deterministic equivalents (from Lemma F.3), and provides deterministic equivalents for resolvents needed to obtain asymptotic bias and variance of standard ridge regression. It is adapted from Lemma S.6.10 of Patil et al. (2022b).

Lemma F.6 (Deterministic equivalents for ridge resolvents associated with generalized bias and variance)**.** *Suppose $\boldsymbol{x}_i \in \mathbb{R}^p$, $1 \leq i \leq n$, are i.i.d. random vectors with each $\boldsymbol{x}_i = \boldsymbol{z}_i \boldsymbol{\Sigma}^{1/2}$, where $\boldsymbol{z}_i \in \mathbb{R}^p$ is a random vector that contains i.i.d. random variables z_{ij}, $1 \leq j \leq p$, each with $\mathbb{E}[z_{ij}] = 0$, $\mathbb{E}[z_{ij}^2] = 1$, and $\mathbb{E}[|z_{ij}|^{8+\alpha}] \leq M_\alpha$ for some constants $\alpha > 0$ and $M_\alpha < \infty$, and $\boldsymbol{\Sigma} \in \mathbb{R}^{p \times p}$ is a positive semidefinite matrix with $r_{\min}\boldsymbol{I}_p \preceq \boldsymbol{\Sigma} \preceq r_{\max}\boldsymbol{I}_p$ for some constants $r_{\min} > 0$ and $r_{\max} < \infty$ (independent of p). Let $\boldsymbol{X} \in \mathbb{R}^{n \times p}$ be the concatenated random matrix with $\boldsymbol{x}_i$, $1 \leq i \leq n$, as its rows, and define $\widehat{\boldsymbol{\Sigma}} := \boldsymbol{X}^\top \boldsymbol{X}/n \in \mathbb{R}^{p \times p}$. Let $\gamma := p/n$. Then, for $\lambda > 0$, as $n, p \to \infty$ with $0 < \liminf \gamma \leq \limsup \gamma < \infty$, the following statements hold:*

(1) Bias of ridge regression:

$$\lambda^2(\widehat{\boldsymbol{\Sigma}}+\lambda\boldsymbol{I}_p)^{-1}\boldsymbol{A}(\widehat{\boldsymbol{\Sigma}}+\lambda\boldsymbol{I}_p)^{-1} \simeq (v(-\lambda;\gamma,\boldsymbol{\Sigma})\boldsymbol{\Sigma}+\boldsymbol{I}_p)^{-1}(\widetilde{v}_b(-\lambda;\gamma,\boldsymbol{\Sigma},\boldsymbol{A})\boldsymbol{\Sigma}+\boldsymbol{A})(v(-\lambda;\gamma,\boldsymbol{\Sigma})\boldsymbol{\Sigma}+\boldsymbol{I}_p)^{-1}. \quad (65)$$

(2) Variance of ridge regression:

$$(\widehat{\boldsymbol{\Sigma}}+\lambda\boldsymbol{I}_p)^{-2}\widehat{\boldsymbol{\Sigma}}\boldsymbol{A} \simeq \widetilde{v}_v(-\lambda;\gamma,\boldsymbol{\Sigma})(v(-\lambda;\gamma,\boldsymbol{\Sigma})\boldsymbol{\Sigma}+\boldsymbol{I}_p)^{-2}\boldsymbol{\Sigma}\boldsymbol{A}. \quad (66)$$

Here $v(-\lambda;\gamma,\boldsymbol{\Sigma})>0$ is the unique solution to the fixed-point equation

$$\frac{1}{v(-\lambda;\gamma,\boldsymbol{\Sigma})} = \lambda + \int \frac{\gamma r}{1+v(-\lambda;\gamma,\boldsymbol{\Sigma})r}\,\mathrm{d}H_n(r;\boldsymbol{\Sigma}), \quad (67)$$

and $\widetilde{v}_b(-\lambda;\gamma,\boldsymbol{\Sigma})$ and $\widetilde{v}_v(-\lambda;\gamma,\boldsymbol{\Sigma})$ are defined through $v(-\lambda;\gamma,\boldsymbol{\Sigma})$ by the following equations:

$$\widetilde{v}_b(-\lambda;\gamma,\boldsymbol{\Sigma},\boldsymbol{A}) = \frac{\gamma\,\mathrm{tr}[\boldsymbol{A}\boldsymbol{\Sigma}(v(-\lambda;\gamma,\boldsymbol{\Sigma})\boldsymbol{\Sigma}+\boldsymbol{I}_p)^{-2}]/p}{v(-\lambda;\gamma,\boldsymbol{\Sigma})^{-2}-\int \gamma r^2(1+v(-\lambda;\gamma,\boldsymbol{\Sigma})r)^{-2}\,\mathrm{d}H_n(r;\boldsymbol{\Sigma})}, \quad (68)$$

$$\widetilde{v}_v(-\lambda;\gamma,\boldsymbol{\Sigma})^{-1} = v(-\lambda;\gamma,\boldsymbol{\Sigma})^{-2}-\int \gamma r^2(1+v(-\lambda;\gamma,\boldsymbol{\Sigma})r)^{-2}\,\mathrm{d}H_n(r;\boldsymbol{\Sigma}), \quad (69)$$

where $H_n(\cdot;\boldsymbol{\Sigma})$ is the empirical distribution (supported on $[r_{\min},r_{\max}]$) of the eigenvalues of $\boldsymbol{\Sigma}$.

Though Lemma F.6 states the dependency explicitly, we will simply write $H_n(r)$, $v(-\lambda;\gamma)$, $\widetilde{v}_b(-\lambda;\gamma,\boldsymbol{A})$, and $\widetilde{v}_v(-\lambda;\gamma)$ to denote $H_n(r;\boldsymbol{\Sigma})$, $v(-\lambda;\gamma,\boldsymbol{\Sigma})$, $\widetilde{v}_b(-\lambda;\gamma,\boldsymbol{\Sigma},\boldsymbol{A})$, and $\widetilde{v}_v(-\lambda;\gamma,\boldsymbol{\Sigma})$, respectively, for simplifying notations when it is clear from the context. When $\boldsymbol{A}=\boldsymbol{\Sigma}$, we simply write $\widetilde{v}_b(-\lambda;\gamma)=\widetilde{v}_b(-\lambda;\gamma,\boldsymbol{A})$. The moment assumption of order $8+\alpha$ for some $\alpha>0$ in the above lemma can be relaxed to only requiring the existence of moments of order $4+\alpha$ by a truncation argument as in the proof of Theorem 6 of Hastie et al. (2022) (in Appendix A.4 therein). We omit the details and refer the readers to Hastie et al. (2022).

F.2.2. Extended ridge resolvents

The lemma below extends the deterministic equivalents of the ridge resolvents in Lemma F.6 to provide deterministic equivalents for Tikhonov resolvents, where the regularization matrix $\lambda\boldsymbol{I}_p$ is replaced with $\lambda(\boldsymbol{I}_p+\boldsymbol{C})$ and $\boldsymbol{C}\in\mathbb{R}^{p\times p}$ is an arbitrary positive semidefinite random matrix.

Lemma F.7 (Tikhonov resolvents, adapted from Patil et al. (2022a))**.** *Suppose the conditions in Lemma F.6 holds. Let $\boldsymbol{C}\in\mathbb{R}^{p\times p}$ be any symmetric and positive semidefinite random matrix with uniformly bounded operator norm in p that is independent to $\boldsymbol{X}$ for all $n,p\in\mathbb{N}$, and let $\boldsymbol{N}=(\widehat{\boldsymbol{\Sigma}}+\lambda\boldsymbol{I}_p)^{-1}$. Then the following statements hold:*

(1) Tikhonov resolvent:

$$\lambda(\boldsymbol{N}^{-1}+\lambda\boldsymbol{C})^{-1} \simeq \widetilde{\boldsymbol{\Sigma}}_{\boldsymbol{C}}^{-1}. \quad (70)$$

(2) Bias of Tikhonov regression:

$$\lambda^2(\boldsymbol{N}^{-1}+\lambda\boldsymbol{C})^{-1}\boldsymbol{\Sigma}(\boldsymbol{N}^{-1}+\lambda\boldsymbol{C})^{-1} \simeq \widetilde{\boldsymbol{\Sigma}}_{\boldsymbol{C}}^{-1}(\widetilde{v}_b(-\lambda;\gamma,\boldsymbol{\Sigma}_{\boldsymbol{C}})\boldsymbol{\Sigma}+\boldsymbol{\Sigma})\widetilde{\boldsymbol{\Sigma}}_{\boldsymbol{C}}^{-1}. \quad (71)$$

(3) Variance of Tikhonov regression:

$$(\boldsymbol{N}^{-1}+\lambda\boldsymbol{C})^{-1}\widehat{\boldsymbol{\Sigma}}(\boldsymbol{N}^{-1}+\lambda\boldsymbol{C})^{-1}\boldsymbol{\Sigma} \simeq \widetilde{v}_v(-\lambda;\gamma,\boldsymbol{\Sigma}_{\boldsymbol{C}})\widetilde{\boldsymbol{\Sigma}}_{\boldsymbol{C}}^{-1}\boldsymbol{\Sigma}\widetilde{\boldsymbol{\Sigma}}_{\boldsymbol{C}}^{-1}\boldsymbol{\Sigma}, \quad (72)$$

where $\boldsymbol{\Sigma}_{\boldsymbol{C}}=(\boldsymbol{I}_p+\boldsymbol{C})^{-\frac{1}{2}}\boldsymbol{\Sigma}(\boldsymbol{I}_p+\boldsymbol{C})^{-\frac{1}{2}}$, $\widetilde{\boldsymbol{\Sigma}}_{\boldsymbol{C}}=v(-\lambda;\gamma,\boldsymbol{\Sigma}_{\boldsymbol{C}})\boldsymbol{\Sigma}+\boldsymbol{I}_p+\boldsymbol{C}$. Here, $v(-\lambda;\gamma,\boldsymbol{\Sigma}_{\boldsymbol{C}})$, $\widetilde{v}_b(-\lambda;\gamma,\boldsymbol{\Sigma}_{\boldsymbol{C}})$, and $\widetilde{v}_v(-\lambda;\gamma,\boldsymbol{\Sigma}_{\boldsymbol{C}})$ defined by (67)-(69) simplify to

$$\frac{1}{v(-\lambda;\gamma,\boldsymbol{\Sigma}_{\boldsymbol{C}})} = \lambda+\gamma\,\mathrm{tr}[(v(-\lambda;\gamma,\boldsymbol{\Sigma}_{\boldsymbol{C}})\boldsymbol{\Sigma}+\boldsymbol{I}_p+\boldsymbol{C})^{-1}\boldsymbol{\Sigma}]/p, \quad (73)$$

$$\frac{1}{\widetilde{v}_v(-\lambda;\gamma,\boldsymbol{\Sigma}_{\boldsymbol{C}})} = \frac{1}{v(-\lambda;\gamma,\boldsymbol{\Sigma}_{\boldsymbol{C}})^2}-\gamma\,\mathrm{tr}[(v(-\lambda;\gamma,\boldsymbol{\Sigma}_{\boldsymbol{C}})\boldsymbol{\Sigma}+\boldsymbol{I}_p+\boldsymbol{C})^{-2}\boldsymbol{\Sigma}^2]/p, \quad (74)$$

$$\widetilde{v}_b(-\lambda;\gamma,\boldsymbol{\Sigma}_{\boldsymbol{C}}) = \gamma\,\mathrm{tr}[(v(-\lambda;\gamma,\boldsymbol{\Sigma}_{\boldsymbol{C}})\boldsymbol{\Sigma}+\boldsymbol{I}_p+\boldsymbol{C})^{-2}\boldsymbol{\Sigma}^2]/p\cdot\widetilde{v}_v(-\lambda;\gamma,\boldsymbol{\Sigma}_{\boldsymbol{C}}). \quad (75)$$

If $\gamma\to\phi\in(0,\infty)$, then γ in (1)-(3) can be replaced by ϕ, with the empirical distribution H_n of eigenvalues replaced by the limiting distribution H.

F.2.3. RESOLVENTS FOR TRAINING ERROR

The following lemma concerns the deterministic equivalents of quantities that arise in the proof for Lemma D.3.

Lemma F.8 (Resolvents for in-sample error)**.** *Suppose the conditions in Lemma F.6 holds. Let $\boldsymbol{C} \in \mathbb{R}^{p\times p}$ be any symmetric and positive semidefinite random matrix with uniformly bounded operator norm in p that is independent to $\boldsymbol{X}$ for all $n, p \in \mathbb{N}$. Let $I_1, I_2 \overset{\text{SRS}}{\sim} \mathcal{I}_k$ and $\widehat{\boldsymbol{\Sigma}}_j$ be the sample covariance matrix computed using k observations of $\boldsymbol{X}$ indexed by I_j $(j = 0, 1)$. For $j = 1, 2$, let $\boldsymbol{M}_j = (\widehat{\boldsymbol{\Sigma}}_j + \lambda \boldsymbol{I}_p)^{-1}$ be the resolvent for $\widehat{\boldsymbol{\Sigma}}_j$. Then, as $k, n, p \to \infty$ such that $p/n \to \phi \in (0, \infty)$ and $p/k \to \phi_s \in [\phi, \infty)$, the following statements hold:*

(1) Independent product with sample covariance:

$$\boldsymbol{C}\widehat{\boldsymbol{\Sigma}}_j \simeq \boldsymbol{C}\boldsymbol{\Sigma}.$$

(2) Bias term 1:

$$\lambda^2 \boldsymbol{M}_1 \boldsymbol{C} \boldsymbol{M}_2 \simeq (v(-\lambda; \phi_s)\boldsymbol{\Sigma} + \boldsymbol{I}_p)^{-1} \left(\widetilde{v}(-\lambda; \phi, \phi_s, \boldsymbol{C})\boldsymbol{\Sigma} + \boldsymbol{C}\right) (v(-\lambda; \phi_s)\boldsymbol{\Sigma} + \boldsymbol{I}_p)^{-1}. \tag{76}$$

(3) Bias term 2:

$$\boldsymbol{M}_1 \widehat{\boldsymbol{\Sigma}}_{1\cap 2} \boldsymbol{M}_2 \boldsymbol{C} \simeq \widetilde{v}_v(-\lambda; \phi, \phi_s)(v(-\lambda; \phi_s)\boldsymbol{\Sigma} + \boldsymbol{I}_p)^{-2}\boldsymbol{\Sigma}\boldsymbol{C}, \tag{77}$$

(4) Variance term 1:

$$\boldsymbol{M}_1 \widehat{\boldsymbol{\Sigma}}_{1\cap 2} \simeq \boldsymbol{I}_p - (v(-\lambda; \phi_s)\boldsymbol{\Sigma} + \boldsymbol{I}_p)^{-1}, \tag{78}$$

(5) Variance term 2:

$$\begin{aligned}\boldsymbol{M}_1 \widehat{\boldsymbol{\Sigma}}_{1\cap 2} \boldsymbol{M}_2 \widehat{\boldsymbol{\Sigma}}_{1\cap 2} \simeq \frac{\phi_s}{\phi}\left(v(-\lambda; \phi_s) - \frac{\phi_s - \phi}{\phi_s}\lambda \widetilde{v}_v(-\lambda; \phi, \phi_s)\right)(v(-\lambda; \phi_s)\boldsymbol{\Sigma} + \boldsymbol{I}_p)^{-1}\boldsymbol{\Sigma} \\ - \lambda \widetilde{v}_v(-\lambda; \phi, \phi_s)(v(-\lambda; \phi_s)\boldsymbol{\Sigma} + \boldsymbol{I}_p)^{-2}\boldsymbol{\Sigma},\end{aligned} \tag{79}$$

where

$$\begin{aligned}\widetilde{v}(-\lambda; \phi, \phi_s, \boldsymbol{C}) &= \frac{\lim\limits_{k,n,p} \phi \operatorname{tr}[\boldsymbol{C}\boldsymbol{\Sigma}(v(-\lambda; \phi_s)\boldsymbol{\Sigma} + \boldsymbol{I}_p)^{-2}]/p}{v(-\lambda; \phi_s)^{-2} - \phi \int \dfrac{r^2}{(1 + v(-\lambda; \phi_s)r)^2}\,\mathrm{d}H(r)}, \\ \widetilde{v}_v(-\lambda; \phi, \phi_s) &:= \frac{1}{v(-\lambda; \phi_s)^{-2} - \phi \int \dfrac{r^2}{(1 + v(-\lambda; \phi_s)r)^2}\,\mathrm{d}H(r)}.\end{aligned}$$

Proof of Lemma F.8. We split the proof into different parts.

Part (1) Note that $\operatorname{tr}(\widehat{\boldsymbol{\Sigma}}_j) = \sum_{i\in I_j} \|\boldsymbol{x}_i\|_2^2/k$ and $\boldsymbol{x}_i = \boldsymbol{z}_i^\top \boldsymbol{\Sigma} \boldsymbol{z}_i$. By Lemma G.4, we have that $\operatorname{tr}(\widehat{\boldsymbol{\Sigma}}_j)/p - \operatorname{tr}(\boldsymbol{\Sigma})/p \xrightarrow{\text{a.s.}} 0$. Since $\|\boldsymbol{C}\|_{\text{op}}$ is uniformly upper bounded and

$$\left|\frac{1}{p}\operatorname{tr}(\boldsymbol{C}\widehat{\boldsymbol{\Sigma}}_j) - \frac{1}{p}\operatorname{tr}(\boldsymbol{C}\boldsymbol{\Sigma})\right| \le \frac{1}{p}|\operatorname{tr}(\boldsymbol{C}(\widehat{\boldsymbol{\Sigma}}_j - \boldsymbol{\Sigma}))| \le \frac{1}{p}\|\boldsymbol{C}\|_{\text{op}}|\operatorname{tr}(\widehat{\boldsymbol{\Sigma}}_j - \boldsymbol{\Sigma})|,$$

it follows that $\frac{1}{p}\operatorname{tr}(\boldsymbol{C}\widehat{\boldsymbol{\Sigma}}_j) - \frac{1}{p}\operatorname{tr}(\boldsymbol{C}\boldsymbol{\Sigma}) \xrightarrow{\text{a.s.}} 0$, which implies that $\boldsymbol{C}\widehat{\boldsymbol{\Sigma}}_j \simeq \boldsymbol{C}\boldsymbol{\Sigma}$

Part (2) This is a direct consequence of Patil et al. (2022a, Part (c) of the proof for Lemma S.24).

Part (3) This is a direct consequence of Patil et al. (2022a, Part (c) of the proof for Lemma S.25).

Part (4) Let $i_0 = |I_1 \cap I_2|$. Conditioning on $\widehat{\boldsymbol{\Sigma}}_{1\cap 2}$ and i_0, from Definition F.2 and Lemma F.7 (1) we have

$$\lambda \boldsymbol{M}_1 \simeq \boldsymbol{M}^{\mathrm{det}}_{\boldsymbol{M}_{1\cap 2}, i_0} := \frac{k}{k-i_0}\left(v_1 \boldsymbol{\Sigma} + \boldsymbol{I}_p + \boldsymbol{C}_1\right)^{-1} \Big| \, i_0,$$

where $v_1 = v(-\lambda; \gamma_1, \boldsymbol{\Sigma}_{\boldsymbol{C}_1})$, $\boldsymbol{\Sigma}_{\boldsymbol{C}_1} = (\boldsymbol{I}_p + \boldsymbol{C}_1)^{-\frac{1}{2}} \boldsymbol{\Sigma} (\boldsymbol{I}_p + \boldsymbol{C}_1)^{-\frac{1}{2}}$, $\boldsymbol{C}_1 = i_0(\lambda(k-i_0))^{-1}\boldsymbol{M}_{1\cap 2}^{-1}$, and $\gamma_1 = p/(k-i_0)$. Here the subscripts of v_1 and $\boldsymbol{C}_1$ are related to the aspect ratio γ_1. Because

$$\limsup \left\| \widehat{\boldsymbol{\Sigma}}_{1\cap 2} \right\|_{\mathrm{op}} \leq r_{\max}(1 + \sqrt{\phi_s^2/\phi})^2,$$

almost surely as $k, n, p \to \infty$ such that $p/n \to \phi$ and $p/k \to \phi_s$, by Lemma F.3 (3), we have

$$\boldsymbol{M}_1 \widehat{\boldsymbol{\Sigma}}_{1\cap 2} \simeq \lambda^{-1} \boldsymbol{M}^{\mathrm{det}}_{\boldsymbol{M}_{1\cap 2}, i_0} \widehat{\boldsymbol{\Sigma}}_{1\cap 2} \mid i_0.$$

That is,

$$\boldsymbol{M}_1 \widehat{\boldsymbol{\Sigma}}_{1\cap 2} \simeq \frac{k}{i_0} (\boldsymbol{M}_{1\cap 2}^{-1} + \lambda \boldsymbol{C}_0)^{-1} \widehat{\boldsymbol{\Sigma}}_{1\cap 2} \mid i_0,$$

where $\boldsymbol{C}_0 = (k-i_0)/i_0 \cdot (v_1 \boldsymbol{\Sigma} + \boldsymbol{I}_p)$. Define $\boldsymbol{\Sigma}_{\boldsymbol{C}_0} = (\boldsymbol{I} + \boldsymbol{C}_0)^{-\frac{1}{2}} \boldsymbol{\Sigma} (\boldsymbol{I} + \boldsymbol{C}_0)^{-\frac{1}{2}}$. Conditioning on i_0, by Lemma F.7 (1), we have

$$\begin{aligned}
\operatorname{tr}[\boldsymbol{\Sigma}_{\boldsymbol{C}_1}(v_1 \boldsymbol{\Sigma}_{\boldsymbol{C}_1} + \boldsymbol{I}_p)^{-1}] &= \operatorname{tr}[\boldsymbol{\Sigma}(v_1 \boldsymbol{\Sigma} + \boldsymbol{I}_p + \boldsymbol{C}_1)^{-1}] \\
&= \frac{\lambda(k-i_0)}{i_0} \operatorname{tr}\left[\boldsymbol{\Sigma}\left(\boldsymbol{M}_{1\cap 2}^{-1} + \frac{\lambda(k-i_0)}{i_0}(v_1 \boldsymbol{\Sigma} + \boldsymbol{I}_p)\right)^{-1}\right] \\
&\overset{\text{a.s.}}{=} \frac{k-i_0}{i_0} \operatorname{tr}\left[\boldsymbol{\Sigma}\left(v_0 \boldsymbol{\Sigma} + \boldsymbol{I}_p + \frac{k-i_0}{i_0}(v_1 \boldsymbol{\Sigma} + \boldsymbol{I}_p)\right)^{-1}\right] \\
&= \operatorname{tr}\left[\boldsymbol{\Sigma}\left(\left(\frac{i_0}{k-i_0} v_0 + v_1\right)\boldsymbol{\Sigma} + \frac{k}{k-i_0}\boldsymbol{I}_p\right)^{-1}\right],
\end{aligned}$$

where $v_0 = v(-\lambda; \gamma_0, \boldsymbol{\Sigma}_{\boldsymbol{C}_0})$and $\gamma_0 = p/i_0$. Note that the fixed-point solution v_0 depends on v_1. The fixed-point equations reduce to

$$\begin{aligned}
\frac{1}{v_0} &= \lambda + \gamma_0 \operatorname{tr}[\boldsymbol{\Sigma}_{\boldsymbol{C}_0}(v_0 \boldsymbol{\Sigma}_{\boldsymbol{C}_0} + \boldsymbol{I}_p)^{-1}]/p = \lambda + \frac{p}{k} \operatorname{tr}\left[\boldsymbol{\Sigma}\left(\left(\frac{i_0}{k} v_0 + \frac{k-i_0}{k} v_1\right)\boldsymbol{\Sigma} + \boldsymbol{I}_p\right)^{-1}\right]/p \\
\frac{1}{v_1} &= \lambda + \gamma_1 \operatorname{tr}[\boldsymbol{\Sigma}_{\boldsymbol{C}_1}(v_1 \boldsymbol{\Sigma}_{\boldsymbol{C}_1} + \boldsymbol{I}_p)^{-1}]/p = \lambda + \frac{p}{k} \operatorname{tr}\left[\boldsymbol{\Sigma}\left(\left(\frac{i_0}{k} v_0 + \frac{k-i_0}{k} v_1\right)\boldsymbol{\Sigma} + \boldsymbol{I}_p\right)^{-1}\right]/p
\end{aligned}$$

almost surely. Note that the solution (v_0, v_1) to the above equations is a pair of positive numbers and does not depend on samples. If (v_0, v_1) is a solution to the above system, then (v_1, v_0) is also a solution. Thus, any solution to the above equations must be unique. On the other hand, since $v_0 = v_1 = v(-\lambda; p/k)$ satisfies the above equations, it is the unique solution. By Lemma F.3 (7), we can replace $v(-\lambda; \gamma_1, \boldsymbol{\Sigma}_{\boldsymbol{C}_1})$ by the solution $v_0 = v_1 = v(-\lambda; p/k)$ of the above system, which does not depend on samples. Thus,

$$\boldsymbol{M}_1 \widehat{\boldsymbol{\Sigma}}_{1\cap 2} \simeq= \frac{k}{i_0} (\boldsymbol{M}_{1\cap 2}^{-1} + \lambda \boldsymbol{C}^*)^{-1} \widehat{\boldsymbol{\Sigma}}_{1\cap 2} \mid i_0, \tag{80}$$

where $\boldsymbol{C}^* = (k-i_0)/i_0 \cdot (v(-\lambda; p/k)\boldsymbol{\Sigma} + \boldsymbol{I}_p)$. Again from Lemma F.7 (1) we have

$$\begin{aligned}
(\boldsymbol{M}_{1\cap 2}^{-1} + \lambda \boldsymbol{C}^*)^{-1} \widehat{\boldsymbol{\Sigma}}_{1\cap 2} &= \boldsymbol{I}_p - \lambda (\boldsymbol{M}_{1\cap 2}^{-1} + \lambda \boldsymbol{C}^*)^{-1} (\boldsymbol{I}_p + \boldsymbol{C}^*) \\
&\simeq \boldsymbol{I}_p - (v(-\lambda; p/k)\boldsymbol{\Sigma} + \boldsymbol{I}_p + \boldsymbol{C}^*)^{-1} (\boldsymbol{I}_p + \boldsymbol{C}^*) \mid i_0 \\
&= \frac{i_0}{k} (\boldsymbol{I}_p - (v(-\lambda; p/k)\boldsymbol{\Sigma} + \boldsymbol{I}_p)^{-1}).
\end{aligned}$$

Finally, from Lemma F.3 (6), we have

$$\boldsymbol{M}_1 \widehat{\boldsymbol{\Sigma}}_{1\cap 2} \simeq \boldsymbol{I}_p - (v(-\lambda; p/k)\boldsymbol{\Sigma} + \boldsymbol{I}_p)^{-1} \simeq \boldsymbol{I}_p - (v(-\lambda; \phi_s)\boldsymbol{\Sigma} + \boldsymbol{I}_p)^{-1}.$$

Part (5) From Patil et al. (2022a, Part (c) of the proof for Lemma S.2.5), we have that

$$\boldsymbol{M}_1\widehat{\boldsymbol{\Sigma}}_{1\cap 2}\boldsymbol{M}_2\widehat{\boldsymbol{\Sigma}}_{1\cap 2} \simeq \frac{k^2}{i_0^2}(\boldsymbol{M}_{1\cap 2}^{-1}+\lambda\boldsymbol{C}^*)^{-1}\widehat{\boldsymbol{\Sigma}}_{1\cap 2}(\boldsymbol{M}_{1\cap 2}^{-1}+\lambda\boldsymbol{C}^*)^{-1}\widehat{\boldsymbol{\Sigma}}_{1\cap 2},$$

where $\boldsymbol{M}_{1\cap 2}=(\widehat{\boldsymbol{\Sigma}}_{1\cap 2}+\lambda\boldsymbol{I}_p)^{-1}$ and $\boldsymbol{C}^*=(k-i_0)/i_0(v(-\lambda;\phi_s)\boldsymbol{\Sigma}+\boldsymbol{I}_p)$. Since

$$(\boldsymbol{M}_{1\cap 2}^{-1}+\lambda\boldsymbol{C}^*)^{-1}\widehat{\boldsymbol{\Sigma}}_{1\cap 2}=\boldsymbol{I}_p-\lambda(\boldsymbol{M}_{1\cap 2}^{-1}+\lambda\boldsymbol{C}^*)^{-1}(\boldsymbol{I}_p+\boldsymbol{C}^*),$$

we have

$$\begin{aligned}
&\boldsymbol{M}_1\widehat{\boldsymbol{\Sigma}}_{1\cap 2}\boldsymbol{M}_2\widehat{\boldsymbol{\Sigma}}_{1\cap 2}\\
&\simeq \frac{k^2}{i_0^2}(\boldsymbol{M}_{1\cap 2}^{-1}+\lambda\boldsymbol{C}^*)^{-1}\widehat{\boldsymbol{\Sigma}}_{1\cap 2}-\lambda\frac{k^2}{i_0^2}(\boldsymbol{M}_{1\cap 2}^{-1}+\lambda\boldsymbol{C}^*)^{-1}\widehat{\boldsymbol{\Sigma}}_{1\cap 2}(\boldsymbol{M}_{1\cap 2}^{-1}+\lambda\boldsymbol{C}^*)^{-1}(\boldsymbol{I}_p+\boldsymbol{C}^*)\\
&= \frac{k^2}{i_0^2}(\boldsymbol{I}_p-\lambda(\boldsymbol{M}_{1\cap 2}^{-1}+\lambda\boldsymbol{C}^*)^{-1}(\boldsymbol{I}_p+\boldsymbol{C}^*))-\lambda\frac{k^2}{i_0^2}(\boldsymbol{M}_{1\cap 2}^{-1}+\lambda\boldsymbol{C}^*)^{-1}\widehat{\boldsymbol{\Sigma}}_{1\cap 2}(\boldsymbol{M}_{1\cap 2}^{-1}+\lambda\boldsymbol{C}^*)^{-1}(\boldsymbol{I}_p+\boldsymbol{C}^*) \qquad (81)
\end{aligned}$$

From Lemma F.7 (1) and (3), we have that

$$\begin{aligned}
\lambda(\boldsymbol{M}_{1\cap 2}^{-1}+\lambda\boldsymbol{C}^*)^{-1} &\simeq \frac{\phi}{\phi_s}(v(-\lambda;\phi_s)\boldsymbol{\Sigma}+\boldsymbol{I}_p)^{-1}\\
(\boldsymbol{M}_{1\cap 2}^{-1}+\lambda\boldsymbol{C}^*)^{-1}\widehat{\boldsymbol{\Sigma}}_{1\cap 2}(\boldsymbol{M}_{1\cap 2}^{-1}+\lambda\boldsymbol{C}^*)^{-1} &\simeq \frac{\phi^2}{\phi_s^2}\widetilde{v}_v(-\lambda;\phi,\phi_s)(v(-\lambda;\phi_s)\boldsymbol{\Sigma}+\boldsymbol{I}_p)^{-2}\boldsymbol{\Sigma}.
\end{aligned}$$

Combing the above two equivalents, the expression in (81) can be further simplified as:

$$\begin{aligned}
\boldsymbol{M}_1\widehat{\boldsymbol{\Sigma}}_{1\cap 2}\boldsymbol{M}_2\widehat{\boldsymbol{\Sigma}}_{1\cap 2} \simeq \frac{\phi_s}{\phi}\left(v(-\lambda;\phi_s)-\frac{\phi_s-\phi}{\phi_s}\lambda\widetilde{v}_v(-\lambda;\phi,\phi_s)\right)(v(-\lambda;\phi_s)\boldsymbol{\Sigma}+\boldsymbol{I}_p)^{-1}\boldsymbol{\Sigma}\\
-\lambda\widetilde{v}_v(-\lambda;\phi,\phi_s)(v(-\lambda;\phi_s)\boldsymbol{\Sigma}+\boldsymbol{I}_p)^{-2}\boldsymbol{\Sigma}.
\end{aligned}$$

□

F.3. Analytic properties of associated fixed-point equations

In this section, we gather results on the properties of the fixed-point solution $v(-\lambda;\phi)$ defined in (64).

The following lemma provides the existence and uniqueness of the solution $v(-\lambda;\phi)$. The properties of the derivatives in Lemma F.9 are related to the properties of $\widetilde{v}_v(-\lambda;\phi)$ defined in Lemma F.10, which equals $-f'(x)$, where the function f is defined in (82).

Lemma F.9 (Properties of the solution to the fixed-point equation, adapted from Patil et al. (2022a))**.** *Let $\lambda,\phi,a>0$ and $b<\infty$ be real numbers. Let P be a probability measure supported on $[a,b]$. Define the function f such that*

$$f(x)=\frac{1}{x}-\phi\int\frac{r}{1+rx}\,\mathrm{d}P(r)-\lambda. \qquad (82)$$

Then, the following properties hold:

(1) For $\lambda=0$ and $\phi\in(1,\infty)$, there is a unique $x_0\in(0,\infty)$ such that $f(x_0)=0$. The function f is positive and strictly decreasing over $(0,x_0)$ and negative over (x_0,∞), with $\lim_{x\to 0^+}f(x)=\infty$ and $\lim_{x\to\infty}f(x)=0$.

(2) For $\lambda>0$ and $\phi\in(0,\infty)$, there is a unique $x_0^\lambda\in(0,\infty)$ such that $f(x_0^\lambda)=0$. The function f is positive and strictly decreasing over $(0,x_0^\lambda)$ and negative over (x_0^λ,∞), with $\lim_{x\to 0^+}f(x)=\infty$ and $\lim_{x\to\infty}f(x)=-\lambda$.

(3) For $\lambda=0$ and $\phi\in(1,\infty)$, f is differentiable on $(0,\infty)$ and its derivative f' is strictly increasing over $(0,x_0)$, with $\lim_{x\to 0^+}f'(x)=-\infty$ and $f'(x_0)<0$.

(4) For $\lambda>0$ and $\phi\in(0,\infty)$, f is differentiable on $(0,\infty)$ and its derivative f' is strictly increasing over $(0,\infty)$, with $\lim_{x\to 0^+}f'(x)=-\infty$ and $f'(x_0^\lambda)<0$.

The continuity and limiting behavior of the function $\phi \mapsto v(-\lambda;\phi)$ is given for ridge regression ($\lambda > 0$) in Lemma F.10 and for ridgeless regression ($\lambda = 0$) in Lemma F.11.

Lemma F.10 (Continuity in the aspect ratio for ridge regression, adapted from Patil et al. (2022a))**.** *Let $\lambda, a > 0$ and $b < \infty$ be real numbers. Let P be a probability measure supported on $[a, b]$. Consider the function $v(-\lambda;\cdot) : \phi \mapsto v(-\lambda;\phi)$, over $(0,\infty)$, where $v(-\lambda;\phi) > 0$ is the unique solution to the fixed-point equation*

$$\frac{1}{v(-\lambda;\phi)} = \lambda + \phi \int \frac{r}{1 + rv(-\lambda;\phi)} \,\mathrm{d}P(r) \tag{83}$$

Then, the following properties hold:

(1) The range of the function $v(-\lambda;\cdot)$ is a subset of $(0, \lambda^{-1})$.

(2) The function $v(-\lambda;\cdot)$ is continuous and strictly decreasing over $(0,\infty)$. Furthermore, $\lim_{\phi\to 0^+} v(-\lambda;\phi) = \lambda^{-1}$, and $\lim_{\phi\to\infty} v(-\lambda;\phi) = 0$.

(3) The function $\widetilde{v}_v(-\lambda;\cdot) : \phi \mapsto \widetilde{v}_v(-\lambda;\phi)$, where

$$\widetilde{v}_v(-\lambda;\phi) = \left(v(-\lambda;\phi)^{-2} - \int \phi r^2 (1 + rv(-\lambda;\phi))^{-2} \,\mathrm{d}P(r) \right)^{-1},$$

is positive and continuous over $(0,\infty)$. Furthermore, $\lim_{\phi\to 0^+} \widetilde{v}_v(-\lambda;\phi) = \lambda^{-2}$, and $\lim_{\phi\to\infty} \widetilde{v}_v(-\lambda;\phi) = 0$.

(4) The function $\widetilde{v}_b(-\lambda;\cdot) : \phi \mapsto \widetilde{v}_b(-\lambda;\phi)$, where

$$\widetilde{v}_b(-\lambda;\phi) = \widetilde{v}_v(-\lambda;\phi) \int \phi r^2 (1 + v(-\lambda;\phi) r)^{-2} \,\mathrm{d}P(r),$$

is positive and continuous over $(0,\infty)$. Furthermore, $\lim_{\phi\to 0^+} \widetilde{v}_b(-\lambda;\phi) = \lim_{\phi\to\infty} \widetilde{v}_b(-\lambda;\phi) = 0$.

Lemma F.11 (Continuity in the aspect ratio for ridgeless regression, adapted from Patil et al. (2022b))**.** *Let $a > 0$ and $b < \infty$ be real numbers. Let P be a probability measure supported on $[a, b]$. Consider the function $v(0;\cdot) : \phi \mapsto v(0;\phi)$, over $(1,\infty)$, where $v(0;\phi) > 0$ is the unique solution to the fixed-point equation*

$$\frac{1}{\phi} = \int \frac{v(0;\phi) r}{1 + v(0;\phi) r} \,\mathrm{d}P(r). \tag{84}$$

Then, the following properties hold:

(1) The function $v(0;\cdot)$ is continuous and strictly decreasing over $(1,\infty)$. Furthermore, $\lim_{\phi\to 1^+} v(0;\phi) = \infty$, and $\lim_{\phi\to\infty} v(0;\phi) = 0$.

(2) The function $\phi \mapsto (\phi v(0;\phi))^{-1}$ is strictly increasing over $(1,\infty)$. Furthermore, $\lim_{\phi\to 1^+} (\phi v(0;\phi))^{-1} = 0$ and $\lim_{\phi\to\infty} (\phi v(0;\phi))^{-1} = 1$.

(3) The function $\widetilde{v}_v(0;\cdot) : \phi \mapsto \widetilde{v}_v(0;\phi)$, where

$$\widetilde{v}_v(0;\phi) = \left(v(0;\phi)^{-2} - \phi \int r^2 (1 + rv(0;\phi))^{-2} \,\mathrm{d}P(r) \right)^{-1},$$

is positive and continuous over $(1,\infty)$. Furthermore, $\lim_{\phi\to 1^+} \widetilde{v}_v(0;\phi) = \infty$, and $\lim_{\phi\to\infty} \widetilde{v}_v(0;\phi) = 0$.

(4) The function $\widetilde{v}_b(0;\cdot) : \phi \mapsto \widetilde{v}_b(0;\phi)$, where

$$\widetilde{v}_b(0;\phi) = \widetilde{v}_v(0;\phi) \int r^2 (1 + v(0;\phi) r)^{-2} \,\mathrm{d}P(r),$$

is positive and continuous over $(1,\infty)$. Furthermore, $\lim_{\phi\to 1^+} \widetilde{v}_b(0;\phi) = \infty$, and $\lim_{\phi\to\infty} \widetilde{v}_b(0;\phi) = 0$.

The continuity and differentiabilty of the function $\lambda \mapsto v(-\lambda;\phi)$ on a closed interval $[0, \lambda_{\max}]$ for some constant $\lambda_{\max}$ is given for $\phi \in (1,\infty)$ in Lemma F.12 adapted from Patil et al. (2022b). This ensures that $v(0;\phi) = \lim_{\lambda\to 0^+} v(-\lambda;\phi)$ is well-defined for $\phi > 1$ and also implies that related functions are bounded.

Lemma F.12 (Differentiability in the regularization parameter)**.** *Let $0 < a \le b < \infty$ be real numbers. Let P be a probability measure supported on $[a,b]$. Let $\phi > 0$ be a real number. Let $\Lambda = [0, \lambda_{\max}]$ for some constant $\lambda_{\max} \in (0,\infty)$. For $\lambda \in \Lambda$, let $v(-\lambda;\phi) > 0$ denote the solution to the fixed-point equation*

$$\frac{1}{v(-\lambda;\phi)} = \lambda + \phi \int \frac{r}{v(-\lambda;\phi)r + 1}\,\mathrm{d}P(r).$$

When $\lambda = 0$ and $\phi \in (0,1]$, $v(-\lambda;\phi) := +\infty$. Then, the following properties hold:

(1) (Monotonicity) For $\phi \in (0,\infty)$, the function $\lambda \mapsto v(-\lambda;\phi)$ is strictly decreasing in $\lambda \in [0,\infty)$.

(2) (Differentiability) For $\phi \in (1,\infty)$, the function $\lambda \mapsto v(-\lambda;\phi)$ is twice differentiable over Λ.

(3) (Boundedness of the second derivative) For $\phi \in (1,\infty)$, $v(-\lambda;\phi)$, $\partial/\partial\lambda[v(-\lambda;\phi)]$, and $\partial^2/\partial\lambda^2[v(-\lambda;\phi)]$ are bounded over Λ.

Proof of Lemma F.12. Start by re-writing the fixed-point equation as

$$\lambda = \frac{1}{v(-\lambda;\phi)} - \phi \int \frac{r}{v(-\lambda;\phi)r + 1}\,\mathrm{d}P(r).$$

Define a function f by

$$f(x) = \frac{1}{x} - \phi \int \frac{r}{xr+1}\,\mathrm{d}P(r).$$

Observe that $v(-\lambda;\phi) = f^{-1}(\lambda)$. We next study various properties of f and prove the different parts in the statment.

Part (1) Properties of f and f^{-1}:

Observe that

$$f(x) = \frac{1}{x} - \phi \int \frac{r}{xr+1}\,\mathrm{d}P(r) = \frac{1}{x}\left(1 - \phi \int \frac{xr}{xr+1}\,\mathrm{d}P(r)\right).$$

The function $g : x \mapsto 1/x$ is positive and strictly decreasing over $(0,\infty)$ with $\lim_{x\to 0^+} g(x) = \infty$ and $\lim_{x\to\infty} g(x) = 0$, while the function

$$h : x \mapsto 1 - \phi \int \frac{xr}{xr+1}\,\mathrm{d}P(r)$$

is strictly decreasing over $(0,\infty)$ with $h(0) = 1$ and $\lim_{x\to\infty} h(x) = 1 - \phi$.

Thus, there is a unique $0 < x_0 < \infty$ when $\phi > 1$ such that $h(x_0) = 0$, and consequently $f(x_0) = 0$; and $x_0 = +\infty$ when $\phi \in (0,1]$ such that $g(x_0) = 0$, and consequently $f(x_0) = 0$. Because h and g are positive over $[0, x_0)$, f, a product of two positive strictly decreasing functions, is strictly decreasing over $(0, x_0)$, with $\lim_{x\to 0^+} f(x) = \infty$ and $f(x_0) = 0$.

Because f is strictly decreasing over $(0, x_0)$, f^{-1} is strictly decreasing (see, e.g., Problem 2, Chapter 5 of Rudin (1976)). Since $f(x_0) = 0$, $f^{-1}(0) = x_0$, and since $\lim_{x\to 0^+} f(x) = \infty$, $\lim_{y\to\infty} f^{-1}(y) = 0$. Hence, f^{-1} is strictly decreasing over $[0,\infty)$ for all $\phi > 0$ and bounded above by $x_0 < \infty$ for all $\phi > 1$.

Parts (2) and (3) We will prove the remaining two parts together.

Properties of f' and $(f^{-1})'$:

The derivative f' at x is given by

$$f'(x) = -\frac{1}{x^2} + \phi \int \frac{r^2}{(xr+1)^2}\,\mathrm{d}P(r) = -\frac{1}{x^2}\left(1 - \phi \int \left(\frac{xr}{xr+1}\right)^2 \mathrm{d}P(r)\right).$$

The function $g : x \mapsto 1/x^2$ is positive and strictly decreasing over $(0, \infty)$ with $\lim_{x\to 0^+} g(x) = \infty$ and $\lim_{x\to\infty} g(x) = 0$. On the other hand, the function

$$h : x \mapsto 1 - \phi \int \left(\frac{xr}{xr+1}\right)^2 \mathrm{d}P(r)$$

is strictly decreasing over $(0, \infty)$ with $h(0) = 1$ and $h(x_0) > 0$. This follows because for $x \in [0, x_0]$,

$$\begin{aligned} \phi \int \left(\frac{xr}{xr+1}\right)^2 \mathrm{d}P(r) &\leq \left(\frac{x_0 b}{x_0 b + 1}\right) \phi \int \left(\frac{xr}{xr+1}\right) \mathrm{d}P(r) \\ &< \phi \int \frac{xr}{xr+1} \,\mathrm{d}P(r) \leq \phi \int \frac{x_0 r}{x_0 r + 1} \,\mathrm{d}P(r) = 1, \end{aligned} \tag{85}$$

where the first inequality in the chain above follows as the support of P is $[a, b]$, and the last inequality follows since $f(x_0) = 0$ and $x_0 > 0$, which implies that

$$\frac{1}{x_0} = \phi \int \frac{r}{x_0 r + 1} \,\mathrm{d}P(r), \quad \text{ or equivalently that } \quad 1 = \phi \int \frac{x_0 r}{x_0 r + 1} \,\mathrm{d}P(r).$$

Thus, $-f'$, a product of two positive strictly decreasing functions, is strictly decreasing, and in turn, f' is strictly increasing. Moreover, $\lim_{x\to 0^+} f'(x) = -\infty$; when $\phi > 1$, $f'(x_0) < 0$ and when $\phi \in (0, 1]$, $f'(x)$ approaches zero from below as $x \to +\infty$.

When $\phi > 1$, because $f'(x) \neq$ over $(0, x_0)$, by the inverse function theorem, $(f^{-1})'$, we have

$$\left|(f^{-1})'(f(x))\right| = \left|\frac{1}{f'(x)}\right| < \left|\frac{1}{f'(x_0)}\right| = \frac{1}{\frac{1}{x_0^2}\left(1 - \phi \int \left(\frac{xr}{xr+1}\right)^2 \mathrm{d}P(r)\right)} < \infty,$$

where the first inequality uses the fact that $|f'(x_0)| < |f'(x)|$ for $x \in (0, x_0]$ from Part 1, and the last inequality uses the bound from (85).

Properties of f'' and $(f^{-1})''$:

The second derivative f'' at x is given by

$$f''(x) = \frac{2}{x^3} - 2\phi \int \frac{r^3}{(xr+1)^3} \,\mathrm{d}P(r) = \frac{2}{x^3}\left(1 - \phi \int \left(\frac{xr}{xr+1}\right)^3 \mathrm{d}P(r)\right).$$

The rest of the arguments are similar to those in Part 2. The function $g : x \mapsto 1/x^3$ is positive and strictly decreasing over $(0, \infty)$ with $\lim_{x\to 0^+} g(x) = \infty$ and $\lim_{x\to\infty} g(x) = 0$, while the function

$$h : x \mapsto 1 - \phi \int \left(\frac{xr}{xr+1}\right)^3 \mathrm{d}P(r)$$

is strictly decreasing over $(0, \infty)$ with $h(0) = 1$ and $h(x_0) > 0$ as

$$\begin{aligned} \phi \int \left(\frac{xr}{xr+1}\right)^3 \mathrm{d}P(r) &\leq \left(\frac{x_0 b}{x_0 b + 1}\right)^2 \phi \int \left(\frac{xr}{xr+1}\right) \mathrm{d}P(r) \\ &< \phi \int \frac{xr}{xr+1} \,\mathrm{d}P(r) \leq \phi \int \frac{x_0 r}{x_0 r + 1} \,\mathrm{d}P(r) = 1. \end{aligned} \tag{86}$$

It then follows that f'' is strictly decreasing, with $\lim_{x\to 0^+} f''(x) = \infty$; when $\phi > 1$, $f''(x_0) > 0$ and when $\phi \in (0, 1]$, $f''(x)$ approaches zero from above as $x \to +\infty$.

When $\phi > 1$, by inverse function theorem, we have

$$\left|(f^{-1})''(f(x))\right| = \left|\frac{f''(x)}{f'(x)^3}\right| = \frac{\frac{2}{x^3}\left(1 - \phi \int \left(\frac{xr}{xr+1}\right)^3 \mathrm{d}P(r)\right)}{\frac{1}{x^6}\left(1 - \phi \int \left(\frac{xr}{xr+1}\right)^2 \mathrm{d}P(r)\right)^3} \leq \frac{2x_0^3}{\left(1 - \phi \int \left(\frac{xr}{xr+1}\right)^2 \mathrm{d}P(r)\right)^3} < \infty,$$

where the first inequality uses the bound from (86), and the second inequality uses the bound from (85).

This finishes all the parts and concludes the proof. □

G. Helper concentration results

G.1. Size of the intersection of randomly sampled datasets

In this section, we collect various helper results concerned with concentrations and convergences. Below we recall the definition of a hypergeometric random variable, along with its mean and variance. See, e.g., Greene & Wellner (2017) for more related details.

Definition G.1 (Hypergeometric random variable)**.** A random variable X follows the hypergeometric distribution $X \sim \mathrm{Hypergeometric}(n, K, N)$ if its probability mass function is given by

$$\mathbb{P}(X = k) = \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}}, \qquad \max\{0, n + K - N\} \le k \le \min\{n, K\}.$$

The expectation and variance of X are given by

$$\mathbb{E}[X] = \frac{nK}{N}, \qquad \mathrm{Var}(X) = \frac{nK(N-K)(N-n)}{N^2(N-1)}.$$

The following lemma characterize the limiting proportions of shared observations in two simple random samples under proportional asymptotics, when both the subsample size and the full data size tend to infinity, which is adapted from Patil et al. (2022a).

Lemma G.2 (Asymptotic proportions of shared observations)**.** *For $n \in \mathbb{N}$, define $\mathcal{I}_k := \{\{i_1, i_2, \ldots, i_k\} : 1 \le i_1 < i_2 < \ldots < i_k \le n\}$. Let $I_1, I_2 \overset{\texttt{SRSWR}}{\sim} \mathcal{I}_k$, define the random variable $i_0^{\texttt{SRSWR}} := |I_1 \cap I_2|$ to be the number of shared samples, and define $i_0^{\texttt{SRSWOR}}$ accordingly. Let $\{k_m\}_{m=1}^{\infty}$ and $\{n_m\}_{m=1}^{\infty}$ be two sequences of positive integers such that n_m is strictly increasing in m, $n_m^{\nu} \le k_m \le n_m$ for some constant $\nu \in (0, 1)$. Then, $i_0^{\texttt{SRSWR}}/k_m - k_m/n_m \xrightarrow{\text{a.s.}} 0$, and $i_0^{\texttt{SRSWOR}}/k_m - k_m/n_m \xrightarrow{\text{a.s.}} 0$.*

G.2. Convergence of random linear and quadratic forms

In this section, we collect helper lemmas on the concentration of linear and quadratic forms of random vectors.

The following lemma provides the concentration of a linear form of a random vector with independent components. It follows from a moment bound from Lemma 7.8 of Erdős & Yau (2017), along with the Borel-Cantelli lemma, and is adapted from Lemma S.8.5 of Patil et al. (2022b).

Lemma G.3 (Concentration of linear form with independent components)**.** *Let $\boldsymbol{z}_p \in \mathbb{R}^p$ be a sequence of random vector with i.i.d. entries z_{pi}, $i = 1, \ldots, p$ such that for each i, $\mathbb{E}[z_{pi}] = 0$, $\mathbb{E}[z_{pi}^2] = 1$, $\mathbb{E}[|z_{pi}|^{4+\alpha}] \le M_\alpha$ for some $\alpha > 0$ and constant $M_\alpha < \infty$. Let $\boldsymbol{a}_p \in \mathbb{R}^p$ be a sequence of random vectors independent of $\boldsymbol{z}_p$ such that $\limsup_p \|\boldsymbol{a}_p\|^2/p \le M_0$ almost surely for a constant $M_0 < \infty$. Then, $\boldsymbol{a}_p^\top \boldsymbol{z}_p/p \to 0$ almost surely as $p \to \infty$.*

The following lemma provides the concentration of a quadratic form of a random vector with independent components. It follows from a moment bound from Lemma B.26 of Bai & Silverstein (2010), along with the Borel-Cantelli lemma, and is adapted from Lemma S.8.6 of Patil et al. (2022b).

Lemma G.4 (Concentration of quadratic form with independent components)**.** *Let $\boldsymbol{z}_p \in \mathbb{R}^p$ be a sequence of random vector with i.i.d. entries z_{pi}, $i = 1, \ldots, p$ such that for each i, $\mathbb{E}[z_{pi}] = 0$, $\mathbb{E}[z_{pi}^2] = 1$, $\mathbb{E}[|z_{pi}|^{4+\alpha}] \le M_\alpha$ for some $\alpha > 0$ and constant $M_\alpha < \infty$. Let $\boldsymbol{D}_p \in \mathbb{R}^{p \times p}$ be a sequence of random matrix such that $\limsup \|\boldsymbol{D}_p\|_{\mathrm{op}} \le M_0$ almost surely as $p \to \infty$ for some constant $M_0 < \infty$. Then, $\boldsymbol{z}_p^\top \boldsymbol{D}_p \boldsymbol{z}_p/p - \mathrm{tr}[\boldsymbol{D}_p]/p \to 0$ almost surely as $p \to \infty$.*

G.3. Convergence of Cesàro-type mean and max for triangular array

In this section, we collect a helper lemma on deducing almost sure convergence of a Cesàro-type mean from almost sure convergence of the original sequence, which is adapted from Patil et al. (2022a).

Lemma G.5 (Convergence of conditional expectation)**.** *For $n \in \mathbb{N}$, suppose $\{R_{n,\ell}\}_{\ell=1}^{N_n}$ is a set of N_n random variables defined over the probability space $(\Omega, \mathcal{F}, \mathbb{P})$, with $1 < N_n < \infty$ almost surely. If there exists a constant c such that $R_{n,p_n} \xrightarrow{\text{a.s.}} c$ for all deterministic sequences $\{p_n \in [N_n]\}_{n=1}^{\infty}$, then the following statements hold:*

(1) $\max_{\ell \in [N_n]} |R_{n,\ell}(\omega) - c| \xrightarrow{\text{a.s.}} 0$.

(2) $N_n^{-1} \sum_{\ell=1}^{N_n} R_{n,\ell} \xrightarrow{\text{a.s.}} c$.

Lemma G.6 (Convergence of conditional expectation over simple random sampling)**.** *For $n \in \mathbb{N}$ and $k = k_n \in \mathcal{K}_n$, let $M_n = |\mathcal{I}_k|$ and suppose $\{R_{n,1}(I_\ell)\}_{\ell \in [M_n]}$ and $\{R_{n,2}(I_m, I_\ell)\}_{m,\ell \in [M_n], m \neq \ell}$ are sets of M_n and $M_n(M_n - 1)$ random variables, such that $R_{n,2}(I_m, I_\ell) \leq (R_{n,1}(I_m) + R_{n,2}(I_\ell))/2$. Then, the following holds:*

(1) If there exists a constant c_1 such that $R_{n,1}(I_{\ell_n}) \xrightarrow{\text{a.s.}} c_1$ for all deterministic sequences $\{\ell_n \in [M_n]\}_{n=1}^{\infty}$, then $\max_{\ell \in [M_n]} |R_{n,\ell}(I_\ell) - c| \xrightarrow{\text{a.s.}} 0$ and $\mathbb{E}_{I_\ell \overset{\text{SRS}}{\sim} \mathcal{I}_k}[|R_{n,\ell}(I_\ell) - c|] \xrightarrow{\text{a.s.}} 0$.

(2) Further, if there exists a constant c_2 such that $R_{n,2}(I_{m_n}, I_{\ell_n}) \xrightarrow{\text{a.s.}} c_2$ for all sequences of simple random samples $\{(I_{m_n}, I_{\ell_n}) \overset{\text{SRS}}{\sim} \mathcal{I}_{k_n}\}_{n=1}^{\infty}$, then $\max_{(I_m, I_\ell) \overset{\text{SRS}}{\sim} \mathcal{I}_k} |R_{n,2}(I_m, I_\ell) - c_2| \xrightarrow{\text{a.s.}} 0$ and $\mathbb{E}_{(I_m, I_\ell) \overset{\text{SRS}}{\sim} \mathcal{I}_k}[|R_{n,2}(I_{m_n}, I_{\ell_n}) - c_2|] \xrightarrow{\text{a.s.}} 0$.

Proof of Lemma G.6. We split the proof into two cases.

Part (1) The conclusion directly follows from Lemma G.5.

Part (2) Observe that

$$R_{n,2}(I_m, I_\ell) \leq \frac{1}{2}(R_{n,1}(I_m) + R_{n,2}(I_\ell)). \tag{87}$$

From (1), we have that $\mathbb{E}_I[R_{n,1}(I)] \xrightarrow{\text{a.s.}} c_1$, where the expectation is taken with respect to the uniform distribution over $\mathcal{I}_k$. From the condition, we have $R_{n,2}(I_m, I_\ell) \xrightarrow{\text{a.s.}} c_2$ for any $I_m, I_\ell \overset{\text{SRS}}{\sim} \mathcal{I}_k$. Then, by Pratt's lemma (see, e.g., Gut, 2005, Theorem 5.5), the conclusion follows. □

H. GCV correction for arbitrary M

Note that the asymptotic limit of the training error for arbitrary $M \in \mathbb{N}$ is given by

$$\mathscr{T}_M^\lambda = 2\mathscr{E}_{k,2}^\lambda - \mathscr{E}_{k,1}^\lambda + \frac{2}{M}(\mathscr{E}_{k,1}^\lambda - \mathscr{E}_{k,2}^\lambda),$$

where $\mathscr{E}_{k,j}^\lambda = c_{k,M,j}\mathscr{T}_{k,j}^\lambda + (1 - c_{k,M,j})\mathscr{R}_{k,j}^\lambda$. Here, $c_{k,M,j}$ is the limiting proportion of the distinct number of observations from j simple random samples to the distinct number of observations from M simple random samples of size k. Roughly speaking, the proportion of unseen observations from M simple random samples of size k is $(n-k)^M/n^M$ and thus

$$c_{k,M,j} = \lim \frac{1 - (n-k)^j/n^j}{1 - (n-k)^M/n^M} = \frac{1 - (1 - \phi/\phi_s)^j}{1 - (1 - \phi/\phi_s)^M}.$$

From the expression, one knows for certain that the GCV asymptotics will not match the risk of the estimator in general. In addition, the form of the expression also leads to an approach to correct the GCV estimator for general M that we will discuss below. What we prove in Theorem 3.1 is that the difference between the two asymptotics vanishes as $M \to \infty$. We expect the difference to scale as $1/M$. The explicit analysis of the finite-ensemble effect requires carefully analyzing the coefficients $c_{k,M,j}$, and even for the isotropic design, the expression for the GCV asymptotics for general appears to be very involved. It is in principle possible to perform this analysis, but we did not pursue it further in the paper given our primary focus on the full-ensemble estimator. Numerically, we observe that the bias is small for a moderate M (e.g., for $M = 10$) and a reasonable data model with SNR (SNR = 0.6) from Figure 4. In general, we expect this to be the case for either moderate k or M and typical real-world SNR ranges. We will consider adding more numerical illustrations of the finite-ensemble effect in the revision under different settings.

We aim to define the corrected GCV as

$$\overline{\mathrm{gcv}}^{\lambda}_{k,M} := \frac{a_1 T^{\lambda}_{k,M} + a_2 \bar{R}^{\lambda}_{k,M}}{D^{\lambda}_{k,M}},$$

where a_1 and a_2 are two unknown parameters to be determined. To determine the unknown parameters, we need to match the limiting GCV with the true risk. Since

$$\mathscr{T}_1^{\lambda}(\phi,\phi_s) = \mathscr{D}_1^{\lambda}(\phi,\phi_s)\mathscr{R}_1^{\lambda}(\phi,\phi_s), \quad \text{and} \quad \mathscr{T}_2^{\lambda}(\phi,\phi_s) = b_1\mathscr{R}_1^{\lambda}(\phi,\phi_s) + b_2\mathscr{R}_2^{\lambda}(\phi,\phi_s),$$

for some known constants b_1 and b_2 which can be derived in the proof of Proposition 3.3, the adjustment is given by

$$\begin{aligned}
&a_1\Bigg[-\left(1-\frac{2}{M}\right)(c_{k,M,1}\mathscr{D}_1^{\lambda}(\phi,\phi_s) + 1 - c_{k,M,1})\mathscr{R}_1^{\lambda}(\phi,\phi_s) \\
&\quad + 2\left(1+\frac{1}{M}\right)(c_{k,M,2}b_1\mathscr{R}_1^{\lambda}(\phi,\phi_s) + (1-c_{k,M,2}+b_2)\mathscr{R}_2^{\lambda}(\phi,\phi_s))\Bigg] \\
&\quad + \mathscr{D}_M^{\lambda}(\phi,\phi_s)a_2\left(-\left(1-\frac{2}{M}\right)\mathscr{R}_1^{\lambda}(\phi,\phi_s) + 2\left(1+\frac{1}{M}\right)\mathscr{R}_2^{\lambda}(\phi,\phi_s)\right) \\
&= -\left(1-\frac{2}{M}\right)[a_1(c_{k,M,1}\mathscr{D}_1^{\lambda}(\phi,\phi_s) + 1 - c_{k,M,1}) + a_2\mathscr{D}_M^{\lambda}(\phi,\phi_s)]\mathscr{R}_1^{\lambda}(\phi,\phi_s) \\
&\quad + 2\left(1+\frac{1}{M}\right)[a_1(c_{k,M,2}b_1\mathscr{R}_1^{\lambda}(\phi,\phi_s) + (1-c_{k,M,2}+b_2)\mathscr{R}_2^{\lambda}(\phi,\phi_s)) + a_2\mathscr{D}_M^{\lambda}(\phi,\phi_s)]\mathscr{R}_2^{\lambda}(\phi,\phi_s)),
\end{aligned}$$

which implies that

$$\begin{aligned}
a_1(c_{k,M,1}\mathscr{D}_1^{\lambda}(\phi,\phi_s) + 1 - c_{k,M,1}) + a_2\mathscr{D}_M^{\lambda}(\phi,\phi_s) &= \mathscr{D}_M^{\lambda}(\phi,\phi_s) \\
a_1(c_{k,M,2}b_1\mathscr{R}_1^{\lambda}(\phi,\phi_s) + (1-c_{k,M,2}+b_2)\mathscr{R}_2^{\lambda}(\phi,\phi_s)) + a_2\mathscr{D}_M^{\lambda}(\phi,\phi_s)\mathscr{R}_2^{\lambda}(\phi,\phi_s) &= \mathscr{D}_M^{\lambda}(\phi,\phi_s)\mathscr{R}_2^{\lambda}(\phi,\phi_s).
\end{aligned}$$

Solving the above linear system for $a_1 > 0$ and $a_2 \in \mathbb{R}$ gives the correct weights for defining a consistent GCV estimate. The solutions will depend on $\mathscr{D}_M^{\lambda}(\phi,\phi_s)$, $\mathscr{R}_1^{\lambda}(\phi,\phi_s)$ and $\mathscr{R}_2^{\lambda}(\phi,\phi_s)$. For the denominator, (12) is a consistent estimate for $\mathscr{D}_M^{\lambda}(\phi,\phi_s)$. For the prediction risks of $M = 1, 2$, we can use out-of-bag observations to estimate $\mathscr{R}_1^{\lambda}(\phi,\phi_s)$ and $\mathscr{R}_2^{\lambda}(\phi,\phi_s)$.

I. Additional details for numerical experiments

The covariance matrix of an auto-regressive process of order 1 (AR(1)) is given by $\mathbf{\Sigma}_{\mathrm{ar1}}$, where $(\mathbf{\Sigma}_{\mathrm{ar1}})_{ij} = \rho_{\mathrm{AR1}}^{|i-j|}$ for some parameter $\rho_{\mathrm{AR1}} \in (0,1)$, and the AR(1) data model is defined as:

$$\begin{aligned}
y_i &= \boldsymbol{x}_i^{\top}\boldsymbol{\beta}_0 + \epsilon_i, \quad \boldsymbol{x}_i \sim \mathcal{N}(0, \mathbf{\Sigma}_{\mathrm{ar1}}), \\
\boldsymbol{\beta}_0 &= \frac{1}{5}\sum_{j=1}^{5}\boldsymbol{w}_{(j)}, \quad \epsilon_i \sim \mathcal{N}(0,\sigma^2),
\end{aligned} \tag{M-AR1}$$

where $\boldsymbol{w}_{(j)}$ is the eigenvector of $\mathbf{\Sigma}_{\mathrm{ar1}}$ associated with the top jth eigenvalue $r_{(j)}$. From Grenander & Szegö (1958, pp. 69-70), the top j-th eigenvalue can be written as $r_{(j)} = (1-\rho_{\mathrm{AR1}}^2)/(1 - 2\rho_{\mathrm{AR1}}\cos\theta_{jp} + \rho_{\mathrm{AR1}}^2)$ for some $\theta_{jp} \in ((j-1)\pi/(p+1), j\pi/(p+1))$. Then, under model (M-AR1), the signal strength ρ^2 defined in Assumption 2.2 is $5^{-1}(1-\rho_{\mathrm{AR1}}^2)/(1-\rho_{\mathrm{AR1}})^2$, which is the limit of $25^{-1}\sum_{j=1}^{5} r_{(j)}$. Thus, model (M-AR1) parameterized by two parameters ρ_{AR1} and σ^2 satisfies Assumption 2.1-2.2.