Supplement

This document serves as a supplement to the paper "Correcting generalized cross-validation for arbitrary ensembles of penalized estimators." The initial (unnumbered) section outlines the supplement's structure and summarizes the general notation used in the main paper and supplement in Table N1.

Organization

The content of this appendix is organized as follows.

- In Section A, we provide proofs of results in Section 4.

Section	**Content**	**Purpose**
Section A.1	Lemma 7	Preparatory derivative formulae
Section A.2	Lemmas 8–11	Proof of Theorem 1
Section A.3	Lemmas 12–13	Proof of Theorem 2
Section A.4		Proof of Theorem 3
Section A.5	Lemmas 14–19	Helper lemmas (and their proofs) used in the proofs of Theorems 1–3
Section A.6	Lemmas 20–22	Miscellaneous useful facts used in the proofs of Theorems 1–3
Section A.7	Theorem 23	Relaxing Assumption D in the underparameterized regime

- In Section B, we provide proofs of results in Section 5.

Section	**Content**	**Purpose**
Section B.1		Preparatory definitions
Section B.2	Lemmas 24–29	Proofs of Theorems 4 and 5 (and proofs of Lemmas 24–29)
Section B.3		Proof of Theorem 6
Section B.4	Lemmas 30–32	Helper lemmas (and their proofs) used in the proofs of Theorems 4–6

- In Section C, we provide additional illustrations for Section 4 with Gaussian data.

Section	**Content**	**Purpose**
Section C.1	Figures 9–10	Intermediate ovlp- vs. full-estimators in k and λ for elastic net and lasso
Section C.2	Figures 11–13	CGCV ovlp- vs. full-estimators in k and λ for ridge, elastic net, and lasso
Section C.3	Figures 14–16	CGCV vs. GCV in k for ridge, elastic net, and lasso
Section C.4	Figures 17–19	CGCV vs. GCV in M for ridge, elastic net, and lasso
Section C.5	Figures 20–22	CGCV vs. GCV in λ for ridge, elastic net, and lasso
Section C.6	Figure 23	CGCV vs. GCV with a large ensemble size M for ridge, elastic net, and lasso
Section C.7	Figure 24	Inconsistency for the GCV variant (8)

- In Section D, we provide additional illustrations for Section 4 with non-Gaussian data.

Section	**Content**	**Purpose**
Section D.1	Figures 25–28	Experiments with Rademacher distribution
Section D.2	Figures 29–32	Experiments with uniform distribution

- In Section E, we provide additional numerical illustrations for Section 5 with non-Gaussian data.

Section	**Content**	**Purpose**
Section E.1	Figures 33 and 34	CGCV vs. GCV in λ for ridge, elastic net, and lasso

Notation

The notational conventions used in this paper are as follows. Any section-specific notation is introduced in-line as needed.

Table N1. Summary of general notation used in the paper and the supplement.

Notation	**Description**
Non-bold lower or upper case	Denotes scalars (e.g., k, λ, M)
Bold lower case	Denotes vectors (e.g., $\boldsymbol{x}$, $\boldsymbol{y}$, $\boldsymbol{\beta}_0$)
Bold upper case	Denotes matrices (e.g., $\boldsymbol{X}$, $\boldsymbol{\Sigma}$, $\boldsymbol{L}$)
Script font	Denotes certain limiting functions (e.g., $\mathscr{R}$).
$\mathbb{R}$, $\mathbb{R}_{\geq 0}$	Set of real and non-negative real numbers
$[n]$	Set $\{1, \dots, n\}$ for a natural number n
$\lvert I \rvert$	Cardinality of a set I
$(x)_+$	Positive part of a real number x
∇f, $\nabla^2 f$	Gradient and Hessian of a function f
$\mathbb{1}_A$, $\mathbb{P}(A)$	Indicator random variable associated with an event A and probability of A
$\mathbb{E}[X], \mathrm{Var}(X)$	Expectation and variance of a random variable X
$\mathcal{V}^\perp$	Orthogonal complement of a vector space $\mathcal{V}$
$\mathrm{tr}[\boldsymbol{A}]$, $\boldsymbol{A}^{-1}$	Trace and inverse (if invertible) of a square matrix $\boldsymbol{A} \in \mathbb{R}^{p \times p}$
$\mathrm{rank}(\boldsymbol{B})$, $\boldsymbol{B}^\top$, $\boldsymbol{B}^+$	Rank, transpose and Moore-Penrose inverse of matrix $\boldsymbol{B} \in \mathbb{R}^{n \times p}$
$\boldsymbol{C}^{1/2}$	Principal square root of a positive semidefinite matrix $\boldsymbol{C}$
$\boldsymbol{I}_p$ or $\boldsymbol{I}$	The $p \times p$ identity matrix
$\langle \boldsymbol{u}, \boldsymbol{v} \rangle$	Inner product of vectors $\boldsymbol{u}$ and $\boldsymbol{v}$
$\lVert \boldsymbol{u} \rVert_p$	The ℓ_p norm of a vector $\boldsymbol{u}$ for $p \geq 1$
$\lVert f \rVert_{L_p}$	The L_p norm of a function f for $p \geq 1$
$\lVert \boldsymbol{A} \rVert_{\mathrm{op}}$	Operator (or spectral) norm of a real matrix $\boldsymbol{A}$
$\lVert \boldsymbol{A} \rVert_{\mathrm{tr}}$	Trace (or nuclear) norm of a real matrix $\boldsymbol{A}$
$\lVert \boldsymbol{A} \rVert_F$	Frobenius norm of a real matrix $\boldsymbol{A}$
$X \lesssim Y$	$X \leq CY$ for some absolute constant C
$X \lesssim_\alpha Y$	$X \leq C_\alpha Y$ for some constant C_α that may depend on ambient parameter α
$X = \mathcal{O}_\alpha(Y)$	$\lvert X \rvert \leq C_\alpha Y$ for some constant C_α that may depend on ambient parameter α
$\boldsymbol{u} \leq \boldsymbol{v}$	Lexicographic ordering for vectors $\boldsymbol{u}$ and $\boldsymbol{v}$
$\boldsymbol{A} \preceq \boldsymbol{B}$	Loewner ordering for symmetric matrices $\boldsymbol{A}$ and $\boldsymbol{B}$
$\mathcal{O}_{\mathbb{P}}$	Probabilistic big-O notation
$o_{\mathbb{P}}$	Probabilistic little-o notation
$\boldsymbol{C} \simeq \boldsymbol{D}$	Asymptotic equivalence of matrices $\boldsymbol{C}$ and $\boldsymbol{D}$ (see Section B for more details)
$\overset{\mathrm{d}}{=}$	Equality in distribution
$\xrightarrow{\mathrm{d}}$	Denotes convergence in distribution
$\xrightarrow{\mathrm{p}}$	Denotes convergence in probability
$\xrightarrow{\mathrm{a.s.}}$	Denotes almost sure convergence

A. Proofs in Section 4

Throughout, C, C' denote positive absolute constants. If no subscript is present for norm $\|\boldsymbol{u}\|$ of a vector $\boldsymbol{u}$, then it is assumed to the ℓ_2 norm of $\boldsymbol{u}$.

A.1. Preparatory derivative formulae

For a pair of possibly overlapping subsamples (I_m, I_ℓ) (written as $(I, \widetilde{I})$ in this proof for easy of writing) of $\{1, 2, \ldots, n\}$ with $|I_m| = |I_\ell| = k \le n$, we define $\widehat{\boldsymbol{\beta}}$ and $\widetilde{\boldsymbol{\beta}}$ as the estimators using subsamples $(\boldsymbol{x}_i, y_i)_{i\in I}$ and $(\boldsymbol{x}_i, y_i)_{i\in\widetilde{I}}$ respectively, i.e.,

$$\widehat{\boldsymbol{\beta}} := \operatorname*{argmin}_{\boldsymbol{\beta}\in\mathbb{R}^p} \frac{1}{k}\sum_{i\in I}(y_i - \boldsymbol{x}_i^\top\boldsymbol{\beta})^2 + g(\boldsymbol{\beta}), \quad \text{and} \quad \widetilde{\boldsymbol{\beta}} := \operatorname*{argmin}_{\boldsymbol{\beta}\in\mathbb{R}^p} \frac{1}{k}\sum_{i\in \widetilde{I}}(y_i - \boldsymbol{x}_i^\top\boldsymbol{\beta})^2 + \widetilde{g}(\boldsymbol{\beta}),$$

and we denote the corresponding residual and the degree of freedom by

$$\boldsymbol{r} := \boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}, \quad \widetilde{\boldsymbol{r}} := \boldsymbol{y} - \boldsymbol{X}\widetilde{\boldsymbol{\beta}}, \quad \widehat{\mathsf{df}} := \operatorname{tr}[\frac{\partial}{\partial \boldsymbol{y}}\boldsymbol{X}\widehat{\boldsymbol{\beta}}], \quad \widetilde{\mathsf{df}} := \operatorname{tr}[\frac{\partial}{\partial \boldsymbol{y}}\boldsymbol{X}\widetilde{\boldsymbol{\beta}}].$$

For clarity, we use the notation below throughout Section A:

$$\boldsymbol{Z} := \begin{bmatrix}\boldsymbol{X}\boldsymbol{\Sigma}^{-1/2} \,\big|\, \sigma^{-1}\boldsymbol{\epsilon}\end{bmatrix} \in \mathbb{R}^{n\times(p+1)}, \quad \boldsymbol{h} := \begin{pmatrix}\boldsymbol{\Sigma}^{1/2}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)\\ -\sigma\end{pmatrix} \in \mathbb{R}^{p+1}, \quad \widetilde{\boldsymbol{h}} := \begin{pmatrix}\boldsymbol{\Sigma}^{1/2}(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)\\ -\sigma\end{pmatrix} \in \mathbb{R}^{p+1}$$

Under Assumptions B and C, the matrix $\boldsymbol{Z}$ has i.i.d. $\mathcal{N}(0,1)$ entries. The individual risk component $R_{m,\ell}$ in (6) and the residuals $(\boldsymbol{r}, \widetilde{\boldsymbol{r}})$ can be written as functions of $(\boldsymbol{Z}, \boldsymbol{h}, \widetilde{\boldsymbol{h}})$ as follows:

$$(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)^\top\boldsymbol{\Sigma}(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) + \sigma^2 = \boldsymbol{h}^\top\widetilde{\boldsymbol{h}}, \quad \boldsymbol{r} = -\boldsymbol{Z}\boldsymbol{h}, \quad \widetilde{\boldsymbol{r}} = -\boldsymbol{Z}\widetilde{\boldsymbol{h}}.$$

Our first lemma below provides the derivative formula of $\boldsymbol{h}$ with respect to $\boldsymbol{Z}$, which is an important component of our proof. Similar derivative formulas have been studied in M-estimation (Bellec and Shen, 2022), multi-task linear model (Bellec and Romon, 2021; Tan et al., 2022), and multinomial logistic model (Tan and Bellec, 2023).

Lemma 7. [Variant of Bellec and Shen (2022, Theorem 1)] Suppose the penalty g satisfies Assumption D with a constant $\mu > 0$. Then there exists a matrix $\boldsymbol{B} \in \mathbb{R}^{(p+1)\times(p+1)}$ depending on $(\boldsymbol{z}_i)_{i\in I}$ such that

$$\|\boldsymbol{B}\|_{\mathrm{op}} \le (k\mu)^{-1}, \quad \operatorname{rank}(\boldsymbol{B}) \le p, \quad \operatorname{tr}[\boldsymbol{B}] \ge 0 \tag{28}$$

and the derivative of $\boldsymbol{h}$ with respect to $\boldsymbol{Z} = (z_{ij})_{i\in[n], j\in[p+1]}$ is given by

$$\text{for all } i \in [n] \text{ and } j \in [p+1], \quad \frac{\partial \boldsymbol{h}}{\partial z_{ij}} = \boldsymbol{B}\boldsymbol{e}_j\boldsymbol{e}_i^\top\boldsymbol{L}\boldsymbol{r} - \boldsymbol{B}\boldsymbol{Z}^\top\boldsymbol{L}\boldsymbol{e}_i\boldsymbol{e}_j^\top\boldsymbol{h} \quad \text{with} \quad \boldsymbol{L} = \boldsymbol{L}_I = \sum_{i\in I}\boldsymbol{e}_i\boldsymbol{e}_i^\top. \tag{29}$$

Furthermore, $\boldsymbol{V} := \boldsymbol{I}_n - \boldsymbol{Z}\boldsymbol{B}\boldsymbol{Z}^\top\boldsymbol{L}$ satisfies the following:

$$\widehat{\mathsf{df}} = k - \operatorname{tr}[\boldsymbol{L}\boldsymbol{V}], \quad 0 < k\Big(1 + \frac{\|\boldsymbol{L}\boldsymbol{G}\|_{op}^2}{k\mu}\Big)^{-1} \le \operatorname{tr}[\boldsymbol{L}\boldsymbol{V}] \le k \quad \text{and} \quad \|\boldsymbol{L}\boldsymbol{V}\|_{\mathrm{op}} \le 1, \tag{30}$$

where $\boldsymbol{G} = \boldsymbol{X}\boldsymbol{\Sigma}^{-1/2} \in \mathbb{R}^{n\times p}$ is a standard Gaussian matrix.

Lemma 7 is proved in Section A.5.1. Note that Lemma 7 holds for the derivatives of $\widetilde{\boldsymbol{h}}$, using $\widetilde{\boldsymbol{B}}$, $\widetilde{\boldsymbol{L}} = \boldsymbol{L}_{\widetilde{I}}$, and $\widetilde{\boldsymbol{V}}$ instead.

A.2. Proof of Theorem 1 (restated here for convenience)

Theorem 1 (Finite-sample bounds for intermediate estimators) *Suppose Assumptions A–D hold. Let $c = k/n$, $\gamma = \max(1, p/n)$, and $\tau = \min(1, \mu)$. Then there exists an absolute constant $C > 0$ such that the following holds:*

$$\mathbb{E}\Big[\Big|\frac{d^{\rm ovlp}_{m,\ell}}{n}\cdot\frac{\widehat{R}^{\rm ovlp}_{m,\ell}-R_{m,\ell}}{\sqrt{R_{m,m}R_{\ell,\ell}}}\Big|\Big] \le C\frac{\gamma^{7/2}}{\tau^2c^3\sqrt{n}}, \qquad \mathbb{E}\Big[\Big|\frac{d^{\rm full}_{m,\ell}}{n}\cdot\frac{\widehat{R}^{\rm full}_{m,\ell}-R_{m,\ell}}{\sqrt{R_{m,m}R_{\ell,\ell}}}\Big|\Big] \le C\frac{\gamma^{5/2}}{\tau^2c^2\sqrt{n}}. \tag{20}$$

Furthermore, if the same penalty g_m is used for subsample estimate (2) *across all $m \in [M]$, for any $\epsilon \in (0, 1)$, we have*

$$\mathbb{P}\Big(\Big|1-\frac{\widetilde{R}^{\rm ovlp}_M}{R_M}\Big| > \epsilon\Big) \le C\frac{M^3\gamma^{11/2}}{\epsilon\tau^4c^7\sqrt{n}}, \qquad \mathbb{P}\Big(\Big|1-\frac{\widetilde{R}^{\rm full}_M}{R_M}\Big| > \epsilon\Big) \le C\frac{M^3\gamma^{9/2}}{\epsilon\tau^4c^2\sqrt{n}}. \tag{21}$$

Thus, if $(M, \mu^{-1}, p/n, n/k)$ are bounded from above by a constant independent of n, we have

$$\widetilde{R}^{\rm ovlp}_M/R_M = 1 + \mathcal{O}_{\mathbb{P}}(n^{-1/2}), \qquad \widetilde{R}^{\rm full}_M/R_M = 1 + \mathcal{O}_{\mathbb{P}}(n^{-1/2}).$$

A.2.1. Part 1: Proof of Equation (20)

Proof for the ovlp-estimator. Using the notation in Section A.1 with I and $\widetilde{I}$ meaning I_m and I_ℓ, the inequality we want to show can be written as

$$\mathbb{E}\Big[\Big|\frac{\boldsymbol{r}^\top\boldsymbol{L}\widetilde{\boldsymbol{L}}\widetilde{\boldsymbol{r}} - |I\cap\widetilde{I}|(1-\widehat{\mathsf{df}}/k)(1-\widetilde{\mathsf{df}}/k)\boldsymbol{h}^\top\widetilde{\boldsymbol{h}}}{\|\boldsymbol{h}\|\|\widetilde{\boldsymbol{h}}\|}\Big|\Big] \lesssim \sqrt{n}\tau^{-2}c^{-3}\gamma^{7/2}, \tag{31}$$

where $c = k/n, \tau = \min(1, \mu), \gamma = \max(1, p/n)$. Using $\mathrm{tr}[\boldsymbol{L}\boldsymbol{V}] = k - \widehat{\mathsf{df}}$ and $\mathrm{tr}[\widetilde{\boldsymbol{L}}\widetilde{\boldsymbol{V}}] = k - \widetilde{\mathsf{df}}$ by (30), we can rewrite the error inside the expectation in the left-hand side as the sum of three terms

$$\frac{\boldsymbol{r}^\top\boldsymbol{L}\widetilde{\boldsymbol{L}}\widetilde{\boldsymbol{r}} - |I\cap\widetilde{I}|(1-\widehat{\mathsf{df}}/k)(1-\widetilde{\mathsf{df}}/k)\boldsymbol{h}^\top\widetilde{\boldsymbol{h}}}{\|\boldsymbol{h}\|\|\widetilde{\boldsymbol{h}}\|} = \mathsf{Rem}_1 + \mathsf{Rem}_2 + \mathsf{Rem}_3, \tag{32}$$

where $\mathsf{Rem}_1, \mathsf{Rem}_2, \mathsf{Rem}_3$ are defined as

$$\mathsf{Rem}_1 = \frac{\boldsymbol{h}^\top\widetilde{\boldsymbol{h}}}{\|\boldsymbol{h}\|\|\widetilde{\boldsymbol{h}}\|}\frac{\mathrm{tr}[\widetilde{\boldsymbol{L}}\widetilde{\boldsymbol{V}}]}{k}\Big(\mathrm{tr}[\widetilde{\boldsymbol{L}}\boldsymbol{L}\boldsymbol{V}] - \frac{|I\cap\widetilde{I}|\,\mathrm{tr}[\boldsymbol{L}\boldsymbol{V}]}{k}\Big), \mathsf{Rem}_2 = \frac{\mathrm{tr}[\widetilde{\boldsymbol{L}}\widetilde{\boldsymbol{V}}]}{k}\frac{\boldsymbol{r}^\top\widetilde{\boldsymbol{L}}\boldsymbol{L}\boldsymbol{r}(1+\mathrm{tr}[\widetilde{\boldsymbol{B}}]) - \mathrm{tr}[\widetilde{\boldsymbol{L}}\boldsymbol{L}\boldsymbol{V}]\boldsymbol{h}^\top\widetilde{\boldsymbol{h}}}{\|\boldsymbol{h}\|\|\widetilde{\boldsymbol{h}}\|},$$

$$\mathsf{Rem}_3 = \frac{\boldsymbol{r}^\top\widetilde{\boldsymbol{L}}\boldsymbol{L}\widetilde{\boldsymbol{r}}}{\|\boldsymbol{h}\|\|\widetilde{\boldsymbol{h}}\|}\frac{k - \mathrm{tr}[\widetilde{\boldsymbol{L}}\widetilde{\boldsymbol{V}}](1+\mathrm{tr}[\widetilde{\boldsymbol{B}}])}{k}.$$

Next, we bound the moment of $|\mathsf{Rem}_1|, |\mathsf{Rem}_2|, |\mathsf{Rem}_3|$ one by one.

Control of Rem_1. Since $\mathrm{tr}[\widetilde{\boldsymbol{L}}\widetilde{\boldsymbol{V}}] \in (0, k]$ by (30) and $|\widetilde{\boldsymbol{h}}^\top\boldsymbol{h}| \le \|\boldsymbol{h}\|\|\widetilde{\boldsymbol{h}}\|$ by the Cauchy–Schwarz inequality, we have

$$|\mathsf{Rem}_1| \le |\,\mathrm{tr}[\widetilde{\boldsymbol{L}}\boldsymbol{L}\boldsymbol{V}] - k^{-1}|I\cap\widetilde{I}|\,\mathrm{tr}[\boldsymbol{L}\boldsymbol{V}]| = \mathrm{RHS}.$$

Below we bound the second moment of RHS. Here, the key fact is that conditionally on $(|I\cap\widetilde{I}|, I)$, $I\cap\widetilde{I}$ is uniformly distributed over all subsets of I of size $|I\cap\widetilde{I}|$. Then, the variance formula of the sampling without replacement (see Lemma 21 with $x_i = V_{ii}$, $M = |I| = k$, and $m = |I\cap\widetilde{I}|$) implies that the conditional expectation of $(\mathrm{RHS})^2$ given $(|I\cap\widetilde{I}|, I, \boldsymbol{V})$ is bounded from above as

$$\mathbb{E}\Big[\Big(\mathrm{tr}[\widetilde{\boldsymbol{L}}\boldsymbol{L}\boldsymbol{V}] - \frac{|I\cap\widetilde{I}|}{k}\,\mathrm{tr}[\boldsymbol{L}\boldsymbol{V}]\Big)^2 \;\Big|\; |I\cap\widetilde{I}|, I, \boldsymbol{V}\Big] = |I\cap\widetilde{I}|^2\mathbb{E}\Big[\Big(\frac{\sum_{i\in I\cap\widetilde{I}}V_{ii}}{|I\cap\widetilde{I}|} - \frac{\sum_{i\in I}V_{ii}}{|I|}\Big)^2 \;\Big|\; |I\cap\widetilde{I}|, I, \boldsymbol{V}\Big]$$

$$= |I\cap\widetilde{I}|^2\,\mathrm{Var}\Big[\frac{\sum_{i\in I\cap\widetilde{I}}V_{ii}}{|I\cap\widetilde{I}|} \;\Big|\; |I\cap\widetilde{I}|, I, \boldsymbol{V}\Big] \le |I\cap\widetilde{I}|^2\frac{\sum_{i\in I}V_{ii}^2}{|I||I\cap\widetilde{I}|} = |I\cap\widetilde{I}|\frac{\sum_{i\in I}V_{ii}^2}{|I|}.$$

Note that the above inequality holds even when $|I \cap \widetilde{I}| = 0$, since the both sides are then 0. Note in passing that $\sum_{i\in I} V_{ii}^2/|I| \le \|\boldsymbol{L}\boldsymbol{V}\|_F^2/|I| \le \|\boldsymbol{L}\boldsymbol{V}\|_{\mathrm{op}}^2 \le 1$ by (30). Then, we obtain

$$\mathbb{E}\Big[\Big|\mathrm{tr}[\widetilde{\boldsymbol{L}}\boldsymbol{L}\boldsymbol{V}] - \frac{|I\cap\widetilde{I}|\,\mathrm{tr}[\boldsymbol{L}\boldsymbol{V}]}{k}\Big|^2\Big] \le \mathbb{E}\Big[|I\cap\widetilde{I}|\frac{\sum_{i\in I} V_{ii}^2}{|I|}\Big] \le \mathbb{E}[|I\cap\widetilde{I}|] \le k. \tag{33}$$

Thus, (33) and the Cauchy–Schwarz inequality $\mathbb{E}[\cdot]^2 \le \mathbb{E}[(\cdot)^2]$ yield

$$\mathbb{E}|\mathsf{Rem}_1| \le \sqrt{\mathbb{E}[|\mathsf{Rem}_1|^2]} \le \sqrt{\mathbb{E}[(\mathrm{tr}[\widetilde{\boldsymbol{L}}\boldsymbol{L}\boldsymbol{V}] - k^{-1}|I\cap\widetilde{I}|\,\mathrm{tr}[\boldsymbol{L}\boldsymbol{V}])^2]} \le \sqrt{k}. \tag{34}$$

Control of Rem_2. Since $\mathrm{tr}[\widetilde{\boldsymbol{L}}\widetilde{\boldsymbol{V}}] \in (0, k]$ by (30), we have

$$|\mathsf{Rem}_2| \le |\boldsymbol{r}^\top\widetilde{\boldsymbol{L}}\boldsymbol{L}\boldsymbol{r}(1+\mathrm{tr}[\widetilde{\boldsymbol{B}}]) - \mathrm{tr}[\widetilde{\boldsymbol{L}}\boldsymbol{L}\boldsymbol{V}]\boldsymbol{h}^\top\widetilde{\boldsymbol{h}}|/(\|\boldsymbol{h}\|\|\widetilde{\boldsymbol{h}}\|).$$

The lemma below gives the bound of the second moment of the right-hand side.

Lemma 8. For any subset $J \subset [n]$ that is independent of $\boldsymbol{Z}$, we have

$$\mathbb{E}\Big[\Big(\frac{-\boldsymbol{r}^\top\boldsymbol{L}_J\widetilde{\boldsymbol{r}} - \widetilde{\boldsymbol{r}}^\top\widetilde{\boldsymbol{L}}\boldsymbol{L}_J\boldsymbol{r}\,\mathrm{tr}[\widetilde{\boldsymbol{B}}] + \mathrm{tr}[\boldsymbol{L}_J\boldsymbol{V}]\boldsymbol{h}^\top\widetilde{\boldsymbol{h}}}{\|\boldsymbol{h}\|\|\widetilde{\boldsymbol{h}}\|}\Big)^2 \ \Big|\ J\Big] \lesssim \tau^{-2}(|J| + \frac{p(|J|+p)^2}{k^2})$$

where $\tau = \min(1,\mu)$ and $\boldsymbol{L}_J = \sum_{i\in J}\boldsymbol{e}_i\boldsymbol{e}_i^\top$.

Lemma 8 is proved in Section A.5.2. A sketch is as follows: Lemma 8 follows from the derivative formula (29) and the second order Stein's formula, where $-\boldsymbol{r}^\top\boldsymbol{L}_J\widetilde{\boldsymbol{r}} = \boldsymbol{r}^\top\boldsymbol{L}_J\boldsymbol{Z}\widetilde{\boldsymbol{h}}$ is regarded as a component-wise inner product of the Gaussian matrix $\boldsymbol{L}_J\boldsymbol{Z} \in \mathbb{R}^{n\times(p+1)}$ and $\boldsymbol{L}_J\boldsymbol{r}\widetilde{\boldsymbol{h}}^\top \in \mathbb{R}^{n\times(p+1)}$.

Lemma 8 with $J = I\cap\widetilde{I}$ and $\mathbb{E}|\mathsf{Rem}_2| \le \sqrt{\mathbb{E}[|\mathsf{Rem}_2|^2]}$ lead to

$$\mathbb{E}|\mathsf{Rem}_2| \lesssim \tau^{-1}\sqrt{\mathbb{E}[|I\cap\widetilde{I}| + p(|I\cap\widetilde{I}|+p)^2k^{-2}]} \le \tau^{-1}\sqrt{k+p(k+p)^2k^{-2}} \lesssim \sqrt{k}\tau^{-1}(1+p/k)^{3/2} \tag{35}$$

thanks to $\tau = \min(1,\mu)$.

Control of Rem_3. Note

$$|\mathsf{Rem}_3| = \frac{|\boldsymbol{r}^\top\widetilde{\boldsymbol{L}}\boldsymbol{L}\widetilde{\boldsymbol{r}}|}{\|\boldsymbol{h}\|\|\widetilde{\boldsymbol{h}}\|}\cdot\frac{|k-\mathrm{tr}[\widetilde{\boldsymbol{L}}\widetilde{\boldsymbol{V}}](1+\mathrm{tr}[\widetilde{\boldsymbol{B}}])|}{k} \le \|\widetilde{\boldsymbol{L}}\boldsymbol{Z}\|_{\mathrm{op}}^2\frac{|k-\mathrm{tr}[\widetilde{\boldsymbol{L}}\widetilde{\boldsymbol{V}}](1+\mathrm{tr}[\widetilde{\boldsymbol{B}}])|}{k},$$

thanks to $|\boldsymbol{r}^\top\widetilde{\boldsymbol{L}}\boldsymbol{L}\widetilde{\boldsymbol{r}}| = |\boldsymbol{h}^\top\boldsymbol{Z}^\top\widetilde{\boldsymbol{L}}\boldsymbol{L}\boldsymbol{Z}\widetilde{\boldsymbol{h}}| \le \|\boldsymbol{h}\|\|\widetilde{\boldsymbol{h}}\|\|\widetilde{\boldsymbol{L}}\boldsymbol{L}\boldsymbol{Z}\|_{\mathrm{op}}^2 \le \|\boldsymbol{h}\|\|\widetilde{\boldsymbol{h}}\|\|\widetilde{\boldsymbol{L}}\boldsymbol{Z}\|_{\mathrm{op}}^2$. Let us define the random variable Y and the event Ω as

$$Y = k^{-1}(k - \mathrm{tr}[\widetilde{\boldsymbol{L}}\widetilde{\boldsymbol{V}}](1+\mathrm{tr}[\widetilde{\boldsymbol{B}}])), \quad \Omega = \{\|\widetilde{\boldsymbol{L}}\boldsymbol{Z}\|_{\mathrm{op}} \le 2(\sqrt{k}+\sqrt{p+1})\}.$$

Then, $\mathbb{E}[|\mathsf{Rem}_3|]$ is bounded from above as

$$\begin{aligned}\mathbb{E}|\mathsf{Rem}_3| &\le \mathbb{E}[\|\widetilde{\boldsymbol{L}}\boldsymbol{Z}\|_{\mathrm{op}}^2|Y|] \le \|Y\|_\infty\mathbb{E}[\mathbb{1}_{\Omega^c}\|\widetilde{\boldsymbol{L}}\boldsymbol{Z}\|_{\mathrm{op}}^2] + [2(\sqrt{k}+\sqrt{p+1})]^2\mathbb{E}|Y| \\ &\lesssim \|Y\|_\infty\cdot\mathbb{P}(\Omega^c)\cdot\mathbb{E}[\|\widetilde{\boldsymbol{L}}\boldsymbol{Z}\|_{\mathrm{op}}^4]^{1/2} + (k+p)\mathbb{E}|Y| := \mathsf{Rem}_3^1 + \mathsf{Rem}_3^2\end{aligned}$$

thanks to Hölder's inequality. Below we bound Rem_3^1 and Rem_3^2.

1. Bound of Rem_3^1: Since $\|\widetilde{\boldsymbol{L}}\boldsymbol{Z}\|_{\mathrm{op}} =^d \|\boldsymbol{G}\|_{\mathrm{op}}$ with $\boldsymbol{G} \in \mathbb{R}^{k\times(p+1)}$ being the standard Gaussian matrix, the concentration of the operator norm (see Lemma 20) implies

$$\mathbb{P}(\Omega^c) \le e^{-(\sqrt{k}+\sqrt{p+1})^2/2} \le e^{-(k+p)/2} \text{ and } \mathbb{E}\|\widetilde{\boldsymbol{L}}\boldsymbol{Z}\|_{\mathrm{op}}^4 \lesssim (k+p)^2.$$

By contrast, $\|\widetilde{\boldsymbol{B}}\|_{\mathrm{op}} \le (k\mu)^{-1}$ and $\mathrm{rank}(\widetilde{\boldsymbol{B}}) \le p$ by (28) and $\mathrm{tr}[\widetilde{\boldsymbol{L}}\widetilde{\boldsymbol{V}}] \in (0,k]$ by (30) lead to

$$\|Y\|_\infty = |1 - k^{-1}\mathrm{tr}[\widetilde{\boldsymbol{L}}\widetilde{\boldsymbol{V}}](1+\mathrm{tr}[\widetilde{\boldsymbol{B}}])| \le 1 + |k^{-1}\mathrm{tr}[\widetilde{\boldsymbol{L}}\widetilde{\boldsymbol{V}}]|(1+|\mathrm{tr}[\widetilde{\boldsymbol{B}}]|) \le 2 + p(k\mu)^{-1} \le \tau^{-1}(1+p/k)$$

with $\tau = \min(1,\mu)$. Therefore, we have

$$\mathsf{Rem}_3^1 = \|Y\|_\infty\mathbb{P}(\Omega^c)\mathbb{E}[\|\widetilde{\boldsymbol{L}}\boldsymbol{Z}\|_{\mathrm{op}}^4]^{1/2} \lesssim \tau^{-1}(1+p/k)e^{-(k+p)/2}(k+p) \lesssim \sqrt{k}\tau^{-1}(1+p/k)^{3/2},$$

thanks to $e^x \geq \sqrt{x/2}$ for all $x > 0$.

2. Bound of Rem_3^2: The following lemma bounds $\mathsf{Rem}_3^2 = (k+p)\mathbb{E}[|Y|]$.

Lemma 9. Let $\xi_1 = (1+\mathrm{tr}[\boldsymbol{B}])\|\boldsymbol{L}\boldsymbol{r}\|^2/\|\boldsymbol{h}\|^2 - \mathrm{tr}[\boldsymbol{L}\boldsymbol{V}]$ and $\xi_2 = k - (1+\mathrm{tr}[\boldsymbol{B}])^2\|\boldsymbol{L}\boldsymbol{r}\|^2/\|\boldsymbol{h}\|^2$. Then

$$\mathbb{E}|k - \mathrm{tr}[\boldsymbol{L}\boldsymbol{V}](1+\mathrm{tr}[\boldsymbol{B}])| = \mathbb{E}|(1+\mathrm{tr}[\boldsymbol{B}])\xi_1 + \xi_2| \lesssim \sqrt{k}\tau^{-2}(1+p/k)^{5/2}.$$

Lemma 9 is proved in Section A.5.3. A sketch is as follows: The quantity ξ_1 is bounded using Lemma 8 with $J = \widetilde{I} = I$. For ξ_2, we use a chi-square type moment inequality (Bellec, 2023, Theorem 7.1).

Recall $\mathsf{Rem}_3^2 = (k+p)\mathbb{E}|Y|$ with $Y := k^{-1}(k - \mathrm{tr}[\widetilde{\boldsymbol{L}}\widetilde{\boldsymbol{V}}](1+\mathrm{tr}[\widetilde{\boldsymbol{B}}]))$. Then, Lemma 9 implies

$$\mathsf{Rem}_3^2 = (1+p/k)\mathbb{E}[|k - \mathrm{tr}[\widetilde{\boldsymbol{L}}\widetilde{\boldsymbol{V}}](1+\mathrm{tr}[\widetilde{\boldsymbol{B}}])|] \lesssim \sqrt{k}\tau^{-2}(1+p/k)^{7/2}.$$

Combining the bounds of Rem_3^1 and Rem_3^2, we have

$$\mathbb{E}|\mathsf{Rem}_3| \lesssim \mathsf{Rem}_3^1 + \mathsf{Rem}_3^2 \lesssim \sqrt{k}\tau^{-2}(1+p/k)^{7/2} + \sqrt{k}\tau^{-1}(1+p/k)^{3/2} \lesssim \sqrt{k}\tau^{-2}(1+p/k)^{7/2}, \tag{36}$$

thanks to $\tau = \min(1,\mu) \leq 1$.

We have now controlled each of $\mathbb{E}|\mathsf{Rem}_1|, \mathbb{E}|\mathsf{Rem}_2|, \mathbb{E}|\mathsf{Rem}_3|$. (34), (35) and (36) result in

$$\begin{aligned}\mathbb{E}[|\mathsf{Rem}_1|] + \mathbb{E}[|\mathsf{Rem}_2|] + \mathbb{E}[|\mathsf{Rem}_3|] &\lesssim \sqrt{k}\tau^{-1}(1+p/k)^{3/2} + \sqrt{k}\tau^{-2}(1+p/k)^{7/2} + \sqrt{k}\\ &\lesssim \sqrt{k}\tau^{-2}(1+p/k)^{7/2} \lesssim \sqrt{n}\tau^{-2}c^{-3}\gamma^{7/2}\end{aligned}$$

thanks to $\gamma = \max(1, p/n) \geq 1$, $c = k/n \leq 1$ and $\tau = \min(1,\mu) \leq 1$. This finishes the proof of (20) for the ovlp-estimator. $\square$

Proof for the full*-estimator.* Using the notation in Section A.1, our goal here is to show

$$\mathbb{E}\Big[\Big|\frac{(n - \widehat{\mathsf{df}} - \widetilde{\mathsf{df}} + \frac{|I\cap\widetilde{I}|}{k^2}\widehat{\mathsf{df}}\widetilde{\mathsf{df}})\boldsymbol{h}^\top\widetilde{\boldsymbol{h}} - \boldsymbol{r}^\top\widetilde{\boldsymbol{r}}}{\|\boldsymbol{h}\|\|\widetilde{\boldsymbol{h}}\|}\Big|\Big] \lesssim \sqrt{n}\tau^{-2}c^{-2}\gamma^{5/2}.$$

Here, $\mathrm{tr}[\boldsymbol{V}] = n - \widehat{\mathsf{df}}$, $\mathrm{tr}[\boldsymbol{L}\boldsymbol{V}] = k - \widehat{\mathsf{df}}$ and $\mathrm{tr}[\widetilde{\boldsymbol{L}}\boldsymbol{V}] - \mathrm{tr}[\widetilde{\boldsymbol{L}}\boldsymbol{L}\boldsymbol{V}] = k - |I\cap\widetilde{I}|$ implies

$$\Big(n - \widehat{\mathsf{df}} - \widetilde{\mathsf{df}} + \frac{|I\cap\widetilde{I}|}{k^2}\widehat{\mathsf{df}}\widetilde{\mathsf{df}}\Big)\boldsymbol{h}^\top\widetilde{\boldsymbol{h}} = \Big(\mathrm{tr}[\boldsymbol{V}] - \frac{\widetilde{\mathsf{df}}}{k}\big\{\mathrm{tr}[\widetilde{\boldsymbol{L}}\boldsymbol{V}] - \mathrm{tr}[\widetilde{\boldsymbol{L}}\boldsymbol{L}\boldsymbol{V}] + \frac{|I\cap\widetilde{I}|}{k}\mathrm{tr}[\boldsymbol{L}\boldsymbol{V}]\big\}\Big)\boldsymbol{h}^\top\widetilde{\boldsymbol{h}},$$

so that, by simple algebra, the error can be decomposed as

$$\Big(n - \widehat{\mathsf{df}} - \widetilde{\mathsf{df}} + \frac{|I\cap\widetilde{I}|}{k^2}\widehat{\mathsf{df}}\widetilde{\mathsf{df}}\Big)\boldsymbol{h}^\top\widetilde{\boldsymbol{h}} - \boldsymbol{r}^\top\widetilde{\boldsymbol{r}} = \|\boldsymbol{h}\|\|\widetilde{\boldsymbol{h}}\|(\mathsf{Rem}_1 + \mathsf{Rem}_2 + \mathsf{Rem}_3 + \mathsf{Rem}_4),$$

where the four following remainders will be bounded separately:

$$\mathsf{Rem}_1 = \frac{-\boldsymbol{r}^\top\widetilde{\boldsymbol{r}} - \widetilde{\boldsymbol{r}}^\top\widetilde{\boldsymbol{L}}\boldsymbol{r}\,\mathrm{tr}[\widetilde{\boldsymbol{B}}] + \mathrm{tr}[\boldsymbol{V}]\boldsymbol{h}^\top\widetilde{\boldsymbol{h}}}{\|\boldsymbol{h}\|\|\widetilde{\boldsymbol{h}}\|}, \qquad \mathsf{Rem}_2 = \frac{\mathrm{tr}[\widetilde{\boldsymbol{B}}]}{1+\mathrm{tr}[\widetilde{\boldsymbol{B}}]}\frac{(1+\mathrm{tr}[\widetilde{\boldsymbol{B}}])\widetilde{\boldsymbol{r}}^\top\widetilde{\boldsymbol{L}}\boldsymbol{r} - \mathrm{tr}[\widetilde{\boldsymbol{L}}\boldsymbol{V}]\boldsymbol{h}^\top\widetilde{\boldsymbol{h}}}{\|\boldsymbol{h}\|\|\widetilde{\boldsymbol{h}}\|},$$

$$\mathsf{Rem}_3 = \frac{(1+\mathrm{tr}[\widetilde{\boldsymbol{B}}])\,\mathrm{tr}[\widetilde{\boldsymbol{L}}\widetilde{\boldsymbol{V}}] - k}{(1+\mathrm{tr}[\widetilde{\boldsymbol{B}}])}\frac{\mathrm{tr}[\widetilde{\boldsymbol{L}}\boldsymbol{V}]}{k}\frac{\boldsymbol{h}^\top\widetilde{\boldsymbol{h}}}{\|\boldsymbol{h}\|\|\widetilde{\boldsymbol{h}}\|}, \qquad \mathsf{Rem}_4 = \frac{\widetilde{\mathsf{df}}}{k}(\mathrm{tr}[\widetilde{\boldsymbol{L}}\boldsymbol{L}\boldsymbol{V}] - \frac{|I\cap\widetilde{I}|}{k}\mathrm{tr}(\boldsymbol{L}\boldsymbol{V}))\frac{\boldsymbol{h}^\top\widetilde{\boldsymbol{h}}}{\|\boldsymbol{h}\|\|\widetilde{\boldsymbol{h}}\|}.$$

Control of Rem_1 and Rem_2. Lemma 8 with $K = [n]$ implies

$$\mathbb{E}[|\mathsf{Rem}_1|] \le \sqrt{\mathbb{E}[|\mathsf{Rem}_1|^2]} \lesssim \sqrt{\tau^{-2}(n + p(n+p)^2/k^2)} \le \tau^{-1}c^{-1}\sqrt{n + p(1+p/n)^2} \lesssim \sqrt{n}\tau^{-1}c^{-1}(1+p/n)^{3/2},$$

thanks to $k/n = c \in (0, 1]$, while $\mathrm{tr}[\widetilde{\boldsymbol{B}}] \ge 0$ by (28) and Lemma 8 with $J = \widetilde{I}$ lead to

$$\mathbb{E}[|\mathsf{Rem}_2| \le \sqrt{\mathbb{E}[|\mathsf{Rem}_2|^2]} \lesssim \sqrt{\tau^{-2}(k + p(k+p)^2/k^2)} \lesssim \tau^{-1}\sqrt{k}(1+p/k)^{3/2} \lesssim \sqrt{n}\tau^{-1}c^{-1}(1+p/n)^{3/2}.$$

Control of Rem_3. According to (30), it holds that $\mathrm{tr}[\widetilde{\boldsymbol{L}}\boldsymbol{V}] = \mathrm{tr}[\widetilde{\boldsymbol{L}}\boldsymbol{L}\boldsymbol{V}] + k - |I \cap \widetilde{I}|$ and $\|\boldsymbol{L}\boldsymbol{V}\|_{\mathrm{op}} \le 1$. Thus,

$$k^{-1}|\,\mathrm{tr}[\widetilde{\boldsymbol{L}}\boldsymbol{V}]| = k^{-1}|\,\mathrm{tr}[\widetilde{\boldsymbol{L}}\boldsymbol{L}\boldsymbol{V}]| + k^{-1}|k - |I \cap \widetilde{I}|| \le \|\boldsymbol{L}\boldsymbol{V}\|_{\mathrm{op}} + 1 \le 2.$$

Combining the above display and $\mathrm{tr}[\widetilde{\boldsymbol{B}}] \ge 0$ by (28), as well as Lemma 9, we obtain

$$\mathbb{E}[|\mathsf{Rem}_3|] \le 2\mathbb{E}[|(1 + \mathrm{tr}[\widetilde{\boldsymbol{B}}])\,\mathrm{tr}[\widetilde{\boldsymbol{L}}\widetilde{\boldsymbol{V}}] - k|] \lesssim \tau^{-2}k^{1/2}(1+p/k)^{5/2} \lesssim \sqrt{n}\tau^{-2}c^{-2}(1+p/n)^{5/2}.$$

Control of Rem_4. Equation (33) implies

$$\mathbb{E}[|\mathsf{Rem}_4|] \le \mathbb{E}|\,\mathrm{tr}[\widetilde{\boldsymbol{L}}\boldsymbol{L}\boldsymbol{V}] - k^{-1}|I \cap \widetilde{I}|\,\mathrm{tr}[\boldsymbol{L}\boldsymbol{V}]| \le \sqrt{k} \le \sqrt{n}.$$

From the above displays, we observe that the dominating upper bound is that of $\mathbb{E}[|\mathsf{Rem}_3|]$. Thus, we obtain

$$\mathbb{E}[|\mathsf{Rem}_1|] + \mathbb{E}[|\mathsf{Rem}_2|] + \mathbb{E}[|\mathsf{Rem}_3|] + \mathbb{E}[|\mathsf{Rem}_4|] \lesssim \sqrt{n}\tau^{-2}c^{-2}(1+p/n)^{5/2} \lesssim \sqrt{n}\tau^{-2}c^{-2}\gamma^{5/2}$$

thanks to $\gamma = \max(1, p/n)$. This finishes the proof of (20) for the full-estimator. □

A.2.2. Part 2: Proof of Equation (21)

For $\# =$ ovlp and full, by algebra, the relative error is bounded from above as

$$\begin{aligned}
\Big|\frac{\widetilde{R}_M^{\#}}{R_M} - 1\Big| &= \frac{|\sum_{m,\ell}(\widehat{R}^{\#}_{m,\ell} - \boldsymbol{h}_m^\top \boldsymbol{h}_\ell)|}{\|\sum_{m'} \boldsymbol{h}_{m'}\|^2} \le \frac{(\sum_{m'}\|\boldsymbol{h}_{m'}\|)^2}{\|\sum_{m'} \boldsymbol{h}_{m'}\|^2} \cdot \sum_{m,\ell} \frac{\|\boldsymbol{h}_m\|\|\boldsymbol{h}_\ell\|}{(\sum_{m'}\|\boldsymbol{h}_{m'}\|)^2} \cdot \frac{|\widehat{R}^{\#}_{m,\ell} - \boldsymbol{h}_m^\top \boldsymbol{h}_\ell|}{\|\boldsymbol{h}_m\|\|\boldsymbol{h}_\ell\|} \\
&\le \frac{(\sum_{m'}\|\boldsymbol{h}_{m'}\|)^2}{\|\sum_{m'} \boldsymbol{h}_{m'}\|^2} \cdot \max_{m,\ell} \frac{|\widehat{R}^{\#}_{m,\ell} - \boldsymbol{h}_m^\top \boldsymbol{h}_\ell|}{\|\boldsymbol{h}_m\|\|\boldsymbol{h}_\ell\|} \le M \cdot \frac{\sum_{m'}\|\boldsymbol{h}_{m'}\|^2}{\|\sum_{m'} \boldsymbol{h}_{m'}\|^2} \cdot \max_{m,\ell} \frac{|\widehat{R}^{\#}_{m,\ell} - \boldsymbol{h}_m^\top \boldsymbol{h}_\ell|}{\|\boldsymbol{h}_m\|\|\boldsymbol{h}_\ell\|} \\
&= M \cdot \mathrm{Ratio} \cdot \max_{m,\ell} \frac{n}{|d^{\#}_{m,\ell}|}|E^{\#}_{m,\ell}|.
\end{aligned} \tag{37}$$

Here, we have defined Ratio and $E^{\#}_{m,\ell}$ by

$$\mathrm{Ratio} := \frac{\sum_{m'}\|\boldsymbol{h}_{m'}\|^2}{\|\sum_{m'} \boldsymbol{h}_{m'}\|^2}, \quad E^{\#}_{m,\ell} := \frac{d^{\#}_{m,\ell}(\widehat{R}^{\#}_{m,\ell} - \boldsymbol{h}_m^\top \boldsymbol{h}_\ell)}{n\|\boldsymbol{h}_m\|\|\boldsymbol{h}_\ell\|}, \tag{38}$$

where $d^{\#}_{m,\ell}$ is the denominator of the corresponding estimator for $\# \in \{\mathrm{ovlp}, \mathrm{full}\}$ defined in (19). We restate their expressions for convenience:

$$d^{\mathrm{ovlp}}_{m,\ell} = |I_m \cap I_\ell|(1 - \widehat{\mathsf{df}}_m/k)(1 - \widehat{\mathsf{df}}_\ell/k), \quad d^{\mathrm{full}}_{m,\ell} = n - \widehat{\mathsf{df}}_m - \widehat{\mathsf{df}}_\ell + k^{-2}|I_m \cap I_\ell|\widehat{\mathsf{df}}_m\widehat{\mathsf{df}}_\ell \tag{39}$$

Below we bound $|E^{\#}_{m,\ell}|$, $|d^{\#}_{m,\ell}|^{-1}$, and Ratio.

Control of $E^{\#}_{m,\ell}$. Markov's inequality applied with the moment bound (20) yields

$$\text{for all } \epsilon > 0, \quad \mathbb{P}(|E^{\#}_{m,\ell}| > \epsilon) \le \epsilon^{-1}\mathbb{E}[|E^{\#}_{m,\ell}|] \lesssim \frac{1}{\epsilon\sqrt{n}\tau^2} \times \begin{cases} c^{-3}\gamma^{7/2} & \# = \mathrm{ovlp} \\ c^{-2}\gamma^{5/2} & \# = \mathrm{full} \end{cases} \tag{40}$$

Control of Ratio. Here the key lemma to bound $\mathrm{Ratio} := \sum_{m'}\|\boldsymbol{h}_{m'}\|^2/\|\sum_{m'}\boldsymbol{h}_{m'}\|^2$ is the concentration of the correlation $\boldsymbol{h}_m^\top \boldsymbol{h}_\ell/\|\boldsymbol{h}_m\|\|\boldsymbol{h}_\ell\|$:

Lemma 10. If the same penalty is used for all $m \in [M]$, i.e., $g_m = g$ in (2), then

$$\eta_{m,\ell} := \mathbb{E}\Big[\frac{\boldsymbol{h}_m^\top \boldsymbol{h}_\ell}{\|\boldsymbol{h}_m\|\|\boldsymbol{h}_\ell\|} \mid (\boldsymbol{z}_i)_{i\in I_m\cap I_\ell}\Big] \ge 0, \quad \text{and} \quad \mathbb{E}\Big[\Big(\frac{\boldsymbol{h}_m^\top \boldsymbol{h}_\ell}{\|\boldsymbol{h}_m\|\|\boldsymbol{h}_\ell\|} - \eta_{m,\ell}\Big)^2\Big] \lesssim \frac{\gamma}{nc^2\tau^2}.$$

Lemma 10 is proved in Section A.5.4. The proof is based on a symmetry argument for the non-negativity of $\eta_{m,\ell}$ and the Gaussian Poincaré inequality for the upper bound.
Now we provide a high-probability bound for the Ratio. Let $U_{m,\ell} := -\boldsymbol{h}_m^\top \boldsymbol{h}_\ell / \|\boldsymbol{h}_m\|\|\boldsymbol{h}_\ell\| + \eta_{m,\ell}$. Then, using the positiveness of $\eta_{m,\ell}$ by Lemma 10, we find

$$\begin{aligned}\sum_m \|\boldsymbol{h}_m\|^2 - \|\sum_m \boldsymbol{h}_m\|^2 &= \sum_{m\neq\ell}(-\boldsymbol{h}_m^\top\boldsymbol{h}_\ell + \eta_{m,\ell}\|\boldsymbol{h}_m\|\|\boldsymbol{h}_\ell\| - \eta_{m,\ell}\|\boldsymbol{h}_m\|\|\boldsymbol{h}_\ell\|) \le \sum_{m\neq\ell} U_{m,\ell}\|\boldsymbol{h}_m\|\|\boldsymbol{h}_\ell\| \\ &\le \Big(\sum_{m\neq\ell} U_{m\ell}^2\Big)^{1/2}\Big(\sum_{m\neq\ell}\|\boldsymbol{h}_m\|^2\|\boldsymbol{h}_\ell\|^2\Big)^{1/2} \le \Big(\sum_{m,\ell} U_{m\ell}^2\Big)^{1/2}\sum_m \|\boldsymbol{h}_m\|^2.\end{aligned}$$

Dividing both sides by $\sum_m \|\boldsymbol{h}_m\|^2$, we have

$$\frac{\|\sum_m \boldsymbol{h}_m\|^2}{\sum_m \|\boldsymbol{h}_m\|^2} \ge 1 - \Big(\sum_{m,\ell} U_{m,\ell}^2\Big)^{1/2}.$$

Then, Markov's inequality applied with $\mathbb{E}[U_{m,\ell}^2] \lesssim \gamma/(nc^2\tau^2)$ by Lemma 10 yields

$$\mathbb{P}\Big(\frac{\sum_m \|\boldsymbol{h}_m\|^2}{\|\sum_m \boldsymbol{h}_m\|^2} \ge 2\Big) \le \mathbb{P}\Big(1 - \Big(\sum_{m,\ell} U_{m,\ell}^2\Big)^{1/2} \le \frac{1}{2}\Big) \le \mathbb{P}\Big(\sum_{m,\ell} U_{m,\ell}^2 \ge 1/4\Big) \le 4\mathbb{E}\Big[\sum_{m,\ell} U_{m,\ell}^2\Big] \lesssim \frac{M^2\gamma}{nc^2\tau^2}, \tag{41}$$

which gives the upper bound of Ratio $= \sum_m \|\boldsymbol{h}_m\|^2 / \|\sum_m \boldsymbol{h}_m\|^2$.
Control of $|d_{m,\ell}^{\#}|^{-1}$. The lemma below gives an lower bound of $d_{m,\ell}^{\#}$ for $\# \in \{\text{ovlp}, \text{full}\}$.

Lemma 11. Let $d_{m,\ell}^{\text{ovlp}}$ and $d_{m,\ell}^{\text{full}}$ be the random variables defined in (39). Then, there exists a positive absolute constant $C > 0$ such that for $\# \in \{\text{ovlp}, \text{full}\}$,

$$\mathbb{P}\Big(\frac{d_{m,\ell}^{\#}}{n} < Cd^{\#}(\tau, c, \gamma)\Big) \lesssim \frac{\gamma^4}{nc^2\tau^4} \text{ where } d^{\text{ovlp}}(\tau, c, \gamma) = \tau^2 c^4 \gamma^{-2}, \text{ and } d^{\text{full}}(\tau, c, \gamma) = \tau^2\gamma^{-2}.$$

Lemma 11 is proved in Section A.5.5. The proof uses (30) and some property of the hypergeometric distribution (Lemma 22).

Combining (37), (40), (41), and Lemma 11 yield the following for all $\epsilon > 0$:

$$\begin{aligned}\mathbb{P}\Big(\Big|\frac{\widetilde{R}_M^{\#}}{R_M} - 1\Big| > \epsilon\Big) &\le \mathbb{P}\Big(\text{Ratio} \cdot \max_{m,\ell} \frac{n}{|d_{m,\ell}^{\#}|}|E_{m,\ell}^{\#}| > \frac{\epsilon}{M}\Big) \\ &\lesssim \frac{M^2\gamma}{n\tau^2c^2} + \mathbb{P}\Big(2\max_{m,\ell}\frac{n}{|d_{m,\ell}^{\#}|}|E_{m,\ell}^{\#}| > \frac{\epsilon}{M}\Big) \\ &\le \frac{M^2\gamma}{n\tau^2c^2} + \sum_{m,\ell}\mathbb{P}\Big(\frac{n}{|d_{m,\ell}^{\#}|}|E_{m,\ell}^{\#}| > \frac{\epsilon}{2M}\Big) \\ &\le \frac{M^2\gamma}{n\tau^2c^2} + \sum_{m,\ell}\Big\{\mathbb{P}\Big(\frac{d_{m,\ell}^{\#}}{n} < Cd^{\#}(\tau, c, \gamma)\Big) + \mathbb{P}\Big(|E_{m,\ell}^{\#}| > \frac{\epsilon C d^{\#}(\tau, c, \gamma)}{2M}\Big)\Big\} \\ &\lesssim \frac{M^2\gamma}{n\tau^2c^2} + M^2\Big[\frac{\gamma^4}{nc^2\tau^4} + \frac{2M}{\epsilon C d^{\#}(\tau, c, \gamma)}\frac{1}{\sqrt{n}\tau^2}\cdot\begin{cases} c^{-3}\gamma^{7/2} & \# = \text{ovlp} \\ c^{-2}\gamma^{5/2} & \# = \text{full},\end{cases}\Big]\end{aligned}$$

where $C > 0$ is an absolute constant, and $d^{\#}(c,\tau,\gamma)$ is given by $d^{\mathrm{ovlp}}(\tau,c,\gamma) = \tau^2 c^4 \gamma^{-2}$ and $d^{\mathrm{full}}(\tau,c,\gamma) = \tau^2\gamma^{-2}$. Substituting this expression to the last bound, we obtain

$$\text{for all } \epsilon > 0, \quad \mathbb{P}\Big(\Big|\frac{\widetilde{R}_M^{\#}}{R_M} - 1\Big| > \epsilon\Big) \lesssim \frac{M^2\gamma}{n\tau^2c^2} + \frac{M^2\gamma^4}{nc^2\tau^4} + \frac{M^3}{\epsilon\sqrt{n}\tau^{2+2}} \cdot \begin{cases} c^{-3-4}\gamma^{7/2+2} & \# = \mathrm{ovlp} \\ c^{-2}\gamma^{5/2+2} & \# = \mathrm{full} \end{cases}$$
$$\lesssim \frac{M^2\gamma^4}{nc^2\tau^4} + \frac{M^3}{\epsilon\sqrt{n}\tau^4} \times \begin{cases} c^{-7}\gamma^{11/2} & \# = \mathrm{ovlp} \\ c^{-2}\gamma^{9/2} & \# = \mathrm{full} \end{cases}, \tag{42}$$

thereby when $\epsilon \in (0,1)$, the second term dominates the first term. This concludes the proof of (21).

A.3. Proof of Theorem 2 (restated here for convenience)

Theorem 2 (Finite-sample bounds for corrected GCV) *Suppose Assumptions A–D hold. Let $c = k/n$, $\gamma = \max(1, p/n)$, and $\tau = \min(1,\mu)$. Assume that the same penalty g_m is used for each $m \in [M]$ in (2). Then, for any $\epsilon \in (0,1)$, there exists an absolute constant $C > 0$ such that the following holds:*

$$\mathbb{P}\Big(\Big|1 - \frac{\widetilde{R}_M^{\mathrm{cgcv,ovlp}}}{R_M}\Big| > \epsilon\Big) \le C\frac{M^4\gamma^{15/2}}{\epsilon\tau^6c^6\sqrt{n}}, \qquad \mathbb{P}\Big(\Big|1 - \frac{\widetilde{R}_M^{\mathrm{cgcv,full}}}{R_M}\Big| > \epsilon\Big) \le C\frac{M^4\gamma^{13/2}}{\epsilon\tau^6c^4\sqrt{n}}. \tag{22}$$

Thus, if $(M, \mu^{-1}, p/n, n/k)$ are bounded by constants independent of n, we have

$$\widetilde{R}_M^{\mathrm{cgcv,ovlp}}/R_M = 1 + \mathcal{O}_{\mathbb{P}}(n^{-1/2}), \qquad \widetilde{R}_M^{\mathrm{cgcv,full}}/R_M = 1 + \mathcal{O}_{\mathbb{P}}(n^{-1/2}).$$

The goal in this section is to show $\widetilde{R}_M^{\mathrm{cgcv},\#}/R_M \approx 1$ for $\# \in \{\mathrm{ovlp}, \mathrm{full}\}$. The definition of $\widetilde{R}_M^{\mathrm{cgcv},\#}$ is recalled for convenience:

$$\widetilde{R}_M^{\mathrm{cgcv},\#} = \frac{\|M^{-1}\sum_m \boldsymbol{r}_m\|^2}{n(1-\widetilde{\mathsf{df}}_M/n)^2} - (c^{-1}-1)\frac{(\widetilde{\mathsf{df}}_M/n)^2}{(1-\widetilde{\mathsf{df}}_M/n)^2}\frac{1}{M^2}\sum_m \widehat{R}_{m,m}^{\#}. \tag{43}$$

Now, we define $\widetilde{R}_M^{\mathrm{cgcv}}$ by

$$\widetilde{R}_M^{\mathrm{cgcv}} := \frac{\|M^{-1}\sum_m \boldsymbol{r}_m\|^2}{n(1-\widetilde{\mathsf{df}}_M/n)^2} - (c^{-1}-1)\frac{(\widetilde{\mathsf{df}}_M/n)^2}{(1-\widetilde{\mathsf{df}}_M/n)^2}\frac{1}{M^2}\sum_m \|\boldsymbol{h}_m\|^2. \tag{44}$$

The difference between $\widetilde{R}_M^{\mathrm{cgcv}}$ and $\widetilde{R}_M^{\mathrm{cgcv},\#}$ is whether it uses $\|\boldsymbol{h}_m\|^2$ or its estimate $\widehat{R}_{m,m}^{\#}$ in the rightmost sum. Considering $\widehat{R}_{m,m}^{\#} \approx \|\boldsymbol{h}_m\|^2$ from (20), $\widetilde{R}_M^{\mathrm{cgcv}}$ is naturally expected to be close to $\widetilde{R}_M^{\mathrm{cgcv},\#}$. We state this approximation more precisely as follows:

$$\text{for all } \epsilon \in (0,1), \quad \mathbb{P}\Big(\Big|\frac{\widetilde{R}_M^{\mathrm{cgcv}} - \widetilde{R}_M^{\mathrm{cgcv},\#}}{R_M}\Big| > \epsilon\Big) \lesssim \frac{M^2}{\sqrt{n}\epsilon\tau^6} \times \begin{cases} c^{-6}\gamma^{15/2} & \# = \mathrm{ovlp} \\ c^{-2}\gamma^{13/2} & \# = \mathrm{full} \end{cases}. \tag{45}$$

Equation (45) is proved in Section A.3.1. The second task is to show $\widetilde{R}_M^{\mathrm{cgcv}} \approx R_M$.

Lemma 12. For $\widetilde{R}_M^{\mathrm{cgcv}}$ defined in (44), we have

$$\text{for all } \epsilon \in (0,1), \quad \mathbb{P}\Big(\Big|\frac{\widetilde{R}_M^{\mathrm{cgcv}} - R_M}{R_M}\Big| > \epsilon\Big) \lesssim \frac{M^4}{\sqrt{n}\epsilon\tau^5} \cdot c^{-4}\gamma^{13/2}.$$

Lemma 12 is proved in Section A.3.2, where we use some concentration of $\widehat{\mathsf{df}}_m$ around its average $\widetilde{\mathsf{df}}_M = M^{-1}\sum_{m=1}^M \widehat{\mathsf{df}}_m$. Now we assume Lemma 12. Then, Equation (45) and Lemma 12 together yield

$$\mathbb{P}\Big(|\widetilde{R}_M^{\mathrm{cgcv},\#}/R_M - 1| > \epsilon\Big) \le \mathbb{P}\Big(|(\widetilde{R}_M^{\mathrm{cgcv},\#} - \widetilde{R}_M^{\mathrm{cgcv}})/R_M| > \epsilon/2\Big) + \mathbb{P}\Big(|\widetilde{R}_M^{\mathrm{cgcv}}/R_M - 1| > \epsilon/2\Big)$$

$$\lesssim \frac{M^4}{\sqrt{n}\epsilon\tau^5}\cdot c^{-4}\gamma^{13/2} + \frac{M}{\sqrt{n}\epsilon\tau^6}\cdot\begin{cases} c^{-6}\gamma^{15/2} & \# = \text{ovlp} \\ c^{-2}\gamma^{13/2} & \# = \text{full}\end{cases} \lesssim \frac{M^4}{\sqrt{n}\epsilon\tau^6}\cdot\begin{cases} c^{-6}\gamma^{15/2} & \# = \text{ovlp} \\ c^{-4}\gamma^{13/2} & \# = \text{full}\end{cases},$$

for all $\epsilon \in (0,1)$. This completes the proof of Theorem 2. In the following sections, we prove Equation (45) and Lemma 12.

A.3.1. Proof of Equation (45)

By the definition of $\widetilde{R}^{\text{cgcv},\#}$ in (43) and $\widetilde{R}^{\text{cgcv}}$ in (44),

$$\frac{\widetilde{R}^{\text{cgcv}}_M - \widetilde{R}^{\text{cgcv},\#}_M}{R_M} = (c^{-1}-1)\Big(\frac{\widetilde{\mathsf{df}}_M/n}{1-\widetilde{\mathsf{df}}_M/n}\Big)^2 \frac{\sum_m (\widehat{R}^{\#}_{mm} - \|\boldsymbol{h}_m\|^2)}{\|\sum_m \boldsymbol{h}_m\|^2}.$$

Here, thanks to $(c^{-1}-1)(\widetilde{\mathsf{df}}_M/n)^2 \le (c^{-1}-1)(k/n)^2 = (c^{-1}-1)c^2 \le c$ by (30), we have

$$\Big|\frac{\widetilde{R}^{\text{cgcv}}_M - \widetilde{R}^{\text{cgcv},\#}_M}{R_M}\Big| \le \frac{c}{(1-\widetilde{\mathsf{df}}_M/n)^2}\frac{\sum_m \|\boldsymbol{h}_m\|^2}{\|\sum_m \boldsymbol{h}_m\|^2}\max_m \Big|\frac{\widehat{R}^{\#}_{mm}}{\|\boldsymbol{h}_m\|^2} - 1\Big| = c\cdot U\cdot \max_m V^{\#}_m,$$

where U and $V^{\#}_m$ are defined as

$$U := \frac{1}{(1-\widetilde{\mathsf{df}}_M/n)^2}\frac{\sum_m \|\boldsymbol{h}_m\|^2}{\|\sum_m \boldsymbol{h}_m\|^2}, \qquad V^{\#}_m := \Big|\frac{\widehat{R}^{\#}_{mm}}{\|\boldsymbol{h}_m\|^2} - 1\Big|. \tag{46}$$

Control of U. By Lemma 18 (introduced later), there exists an absolute constant $C \in (0,1)$ such that

$$\mathbb{P}(1-\widehat{\mathsf{df}}_m/n \le C\tau\gamma^{-1}) \le e^{-nc/2},$$

Applying the above display to $1-\widetilde{\mathsf{df}}_M/n = M^{-1}\sum_{m=1}^M (1-\widehat{\mathsf{df}}_m/n)$ with the union bound, we have

$$\mathbb{P}\Big(\frac{1}{(1-\widetilde{\mathsf{df}}_M/n)^2} \ge \frac{\gamma^2}{C^2\tau^2}\Big) = \mathbb{P}(1-\widetilde{\mathsf{df}}_M/n \le C\tau\gamma^{-1}) \le \sum_{m=1}^M \mathbb{P}(1-\widehat{\mathsf{df}}_m/n \le C\tau\gamma^{-1}) \lesssim Me^{-nc/2}.$$

Combining the above display and the upper bound of $\sum_m \|\boldsymbol{h}_m\|^2/\|\sum_m \boldsymbol{h}_m\|^2$ given by (41), we have

$$\mathbb{P}\Big(U > \frac{2\gamma^2}{C^2\tau^2}\Big) \le \mathbb{P}\Big(\frac{\sum_m \|\boldsymbol{h}_m\|^2}{\|\sum_m \boldsymbol{h}_m\|^2} > 2\Big) + \mathbb{P}\Big(\frac{1}{(1-\widetilde{\mathsf{df}}_M/n)^2} \ge \frac{\gamma^2}{C^2\tau^2}\Big) \lesssim \frac{M^2\gamma^2}{n\tau^2c^2} + \frac{M}{e^{nc/2}} \lesssim \frac{M^2\gamma^2}{n\tau^2c^2} \tag{47}$$

Control of $V^{\#}_m$. Equation (42) with $M=1$ implies

$$\text{for all } \epsilon > 0, \quad \mathbb{P}(V^{\#}_m > \epsilon) \lesssim \frac{\gamma^4}{nc^2\tau^4} + \frac{1}{\epsilon\sqrt{n}\tau^4}\cdot\begin{cases} c^{-7}\gamma^{11/2} & \# = \text{ovlp} \\ c^{-2}\gamma^{9/2} & \# = \text{full}\end{cases} \tag{48}$$

Now we have controlled U and $V^{\#}_m$. (47) and (48) together yield, for all $\epsilon \in (0,1)$,

$$\begin{aligned}
\mathbb{P}\Big(\Big|\frac{\widetilde{R}^{\text{cgcv}}_M - \widetilde{R}^{\text{cgcv},\#}_M}{R_M}\Big| > \epsilon\Big) &\le \mathbb{P}(cU\max_m V^{\#}_m > \epsilon) \le \mathbb{P}\Big(U > \frac{2\gamma^2}{C^2\tau^2}\Big) + \mathbb{P}\Big(c\frac{2\gamma^2}{C^2\tau^2}\max_m V^{\#}_M > \epsilon\Big) \\
&\lesssim \frac{M^2\gamma^2}{n\tau^2c^2} + \sum_m \mathbb{P}\Big(V^{\#}_m > \frac{\epsilon C^2\tau^2}{2c\gamma^2}\Big) \text{ (by (47))} \\
&\lesssim \frac{M^2\gamma^2}{n\tau^2c^2} + \frac{M\gamma^4}{n\tau^4c^2} + \frac{M}{\sqrt{n}\tau^4}\Big(\frac{\epsilon C^2\tau^2}{2c\gamma^2}\Big)^{-1}\cdot\begin{cases} c^{-7}\gamma^{11/2} & \# = \text{ovlp} \\ c^{-2}\gamma^{9/2} & \# = \text{full}\end{cases} \text{ (by (48))} \\
&\lesssim \frac{M^2\gamma^4}{n\tau^4c^2} + \frac{M}{\sqrt{n}\epsilon\tau^6}\times\begin{cases} c^{-6}\gamma^{15/2} & \# = \text{ovlp} \\ c^{-1}\gamma^{13/2} & \# = \text{full}\end{cases} \text{ (thanks to } \tau\in(0,1] \text{ and } \gamma \ge 1) \\
&\lesssim \frac{M^2}{\sqrt{n}\epsilon\tau^6}\times\begin{cases} c^{-6}\gamma^{15/2} & \# = \text{ovlp} \\ c^{-2}\gamma^{13/2} & \# = \text{full}\end{cases} \text{ (thanks to } \epsilon\in(0,1))
\end{aligned}$$

This concludes the proof.

A.3.2. Proof of Lemma 12 (restated here for convenience)

Lemma 12. For $\widetilde{R}_M^{\text{cgcv}}$ defined in (44), we have

$$\text{for all } \epsilon \in (0,1), \quad \mathbb{P}\Big(\Big|\frac{\widetilde{R}_M^{\text{cgcv}} - R_M}{R_M}\Big| > \epsilon\Big) \lesssim \frac{M^4}{\sqrt{n}\epsilon\tau^5} \cdot c^{-4}\gamma^{13/2}.$$

Proof. If we define $d_{m,\ell}^{\text{cgcv}}$ by

$$d_{m,\ell}^{\text{cgcv}} := n\Big\{\Big(1 - \frac{\widetilde{\mathsf{df}}_M}{n}\Big)^2 + (c^{-1} - 1)\,\mathbb{1}_{\{m=\ell\}}\,\Big(\frac{\widetilde{\mathsf{df}}_M}{n}\Big)^2\Big\}$$

the relative error can be written as

$$\frac{\widetilde{R}_M^{\text{cgcv}}}{R_M} - 1 = \frac{\|\sum_m \boldsymbol{r}_m\|^2 - n(c^{-1}-1)(\widetilde{\mathsf{df}}_M/n)^2\sum_m\|\boldsymbol{h}_m\|^2}{n(1-\widetilde{\mathsf{df}}_M/n)^2\|\sum_m \boldsymbol{h}_m\|^2} - 1 = \frac{\sum_{m,\ell}(\boldsymbol{r}_m^\top\boldsymbol{r}_\ell - d_{m,\ell}^{\text{cgcv}}\boldsymbol{h}_m^\top\boldsymbol{h}_\ell)}{n(1-\widetilde{\mathsf{df}}_M/n)^2\|\sum_m \boldsymbol{h}_m\|^2}.$$

By the same argument in (37), the relative error is bounded from above as

$$\begin{aligned}
\Big|\frac{\widetilde{R}_M^{\text{cgcv}}}{R_M} - 1\Big| &\le \frac{1}{(1-\widetilde{\mathsf{df}}_M/n)^2}\frac{(\sum_{m=1}^M\|\boldsymbol{h}_m\|)^2}{\|\sum_{m=1}^M\boldsymbol{h}_m\|^2}\cdot\max_{m,\ell}\frac{|\boldsymbol{r}_m^\top\boldsymbol{r}_\ell - d_{m,\ell}^{\text{cgcv}}\boldsymbol{h}_m^\top\boldsymbol{h}_\ell|}{n\|\boldsymbol{h}_m\|\|\boldsymbol{h}_\ell\|}\\
&\le M\cdot\frac{1}{(1-\widetilde{\mathsf{df}}_M/n)^2}\frac{\sum_m\|\boldsymbol{h}_m\|^2}{\|\sum_m\boldsymbol{h}_m\|^2}\cdot\max_{m,\ell}\Big(\frac{|\boldsymbol{r}_m^\top\boldsymbol{r}_\ell - d_{m,\ell}^{\text{full}}\boldsymbol{h}_m^\top\boldsymbol{h}_\ell|}{n\|\boldsymbol{h}_m\|\|\boldsymbol{h}_\ell\|} + \frac{|d_{m,\ell}^{\text{cgcv}} - d_{m,\ell}^{\text{full}}|}{n}\Big)\\
&= M\cdot U\cdot\max_{m,\ell}\Big(|E_{m,\ell}^{\text{full}}| + |d_{m,\ell}^{\text{cgcv}} - d_{m,\ell}^{\text{full}}|/n\Big),
\end{aligned}$$

where $E_{m,\ell}^{\text{full}}$ and U were defined before by (38) and (46). Their expressions are recalled here for convenience:

$$U := \frac{1}{(1-\widetilde{\mathsf{df}}_M/n)^2}\frac{\sum_m\|\boldsymbol{h}_m\|^2}{\|\sum_m\boldsymbol{h}_m\|^2}, \quad E_{m,\ell}^{\text{full}} := \frac{\boldsymbol{r}_m^\top\boldsymbol{r}_\ell - d_{m,\ell}^{\text{full}}\boldsymbol{h}_m^\top\boldsymbol{h}_\ell}{n\|\boldsymbol{h}_m\|\|\boldsymbol{h}_\ell\|} = \frac{d_{m,\ell}^{\text{full}}(\widehat{R}_{m,\ell}^{\text{full}} - \boldsymbol{h}_m^\top\boldsymbol{h}_\ell)}{n\|\boldsymbol{h}_m\|\|\boldsymbol{h}_\ell\|},$$

where $d_{m,\ell}^{\text{full}} = n - \widehat{\mathsf{df}}_m - \widehat{\mathsf{df}}_\ell + k^{-2}|I_m\cap I_\ell|\widehat{\mathsf{df}}_m\widehat{\mathsf{df}}_\ell$ is the denominator in the full-estimator. Note in passing that U and $E_{m,\ell}^{\text{full}}$ have been already bounded by (40) with $\# = $ full and (47): there exists a absolute constant C such that

$$\mathbb{P}(U > C\tau^{-2}\gamma^2) \lesssim \frac{M^2\gamma^2}{n\tau^2c^2}, \qquad \mathbb{P}(|E_{m,\ell}^{\text{full}}| > \epsilon) \lesssim \frac{\tau^{5/2}}{\epsilon\sqrt{n}\tau^2c^2} \text{ for all } \epsilon > 0. \tag{49}$$

It remains to bound $|d_{m,\ell}^{\text{cgcv}} - d_{m,\ell}^{\text{full}}|/n$. Here, we will argue that

$$\text{for all } m,\ell\in[M], \quad \text{for all } \epsilon > 0, \quad \mathbb{P}\Big(\frac{|d_{m,\ell}^{\text{full}} - d_{m,\ell}^{\text{cgcv}}|}{n} > \epsilon\Big) \lesssim M\Big(e^{-nc/2} + \frac{\gamma^{9/2}}{\epsilon\sqrt{n}\tau^3c^4}\Big). \tag{50}$$

The proof of (50) is given at the end of this section. Now we assume it. Then (49) and (50) together yield, for all $\epsilon\in(0,1)$,

$$\begin{aligned}
\mathbb{P}\Big(\Big|\frac{\widetilde{R}_M^{\text{cgcv}}}{R_M} - 1\Big| > \epsilon\Big) &\le \mathbb{P}\Big(U\cdot\max_{m,\ell}(|E_{m,\ell}^{\text{full}}| + \frac{|d_{m,\ell}^{\text{cgcv}} - d_{m,\ell}^{\text{full}}|}{n}) > \frac{\epsilon}{M}\Big)\\
&\le \mathbb{P}\Big(U > \frac{C\gamma^2}{\tau^2}\Big) + \mathbb{P}\Big(\max_{m,\ell}(|E_{m,\ell}^{\text{full}}| + \frac{|d_{m,\ell}^{\text{cgcv}} - d_{m,\ell}^{\text{full}}|}{n}) > \frac{\epsilon}{M}\frac{\tau^2}{C\gamma^2}\Big)\\
&\lesssim \frac{M^2\gamma^2}{n\tau^2c^2} + \sum_{m,\ell}\mathbb{P}\Big(|E_{m,\ell}^{\text{full}}| > \frac{\epsilon\tau^2}{2CM\gamma^2}\Big) + \mathbb{P}\Big(\frac{|d_{m,\ell}^{\text{cgcv}} - d_{m,\ell}^{\text{full}}|}{n} > \frac{\epsilon\tau^2}{2CM\gamma^2}\Big)
\end{aligned}$$

$$
\begin{aligned}
&\lesssim \frac{M^2\gamma^2}{n\tau^2c^2} + M^2\Big\{\frac{2CM\gamma^2}{\epsilon\tau^2}\cdot\frac{\gamma^{5/2}}{\sqrt{n}\tau^2c^2} + Me^{-nc/2} + \frac{2CM\gamma^2}{\epsilon\tau^2}\cdot\frac{M\gamma^{9/2}}{\sqrt{n}\tau^3c^4}\Big\}\\
&\lesssim \frac{M^4\gamma^{13/2}}{\sqrt{n}\epsilon\tau^5c^4} \quad \text{(thanks to } \epsilon, c, \tau\in(0,1] \text{ and } \gamma\geq 1),
\end{aligned}
$$

which concludes the proof of Lemma 12. $\square$

Proof of Equation (50)**.** The expressions of $d^{\text{full}}_{m,\ell}$ and $d^{\text{cgcv}}_{m,\ell}$ are recalled here for convenience:

$$
\frac{d^{\text{full}}_{m,\ell}}{n} = 1 - \frac{\widehat{\mathsf{df}}_m}{n} - \frac{\widehat{\mathsf{df}}_\ell}{n} + \frac{|I_m\cap I_\ell|}{nk^2}\widehat{\mathsf{df}}_m\widehat{\mathsf{df}}_\ell, \qquad \frac{d^{\text{cgcv}}_{m,\ell}}{n} = \Big(1-\frac{\widetilde{\mathsf{df}}_M}{n}\Big)^2 + (c^{-1}-1)\,\mathbb{1}_{\{m=\ell\}}\Big(\frac{\widetilde{\mathsf{df}}_M}{n}\Big)^2.
$$

Below we prove $d^{\text{full}}_{m,\ell}/n \approx d^{\text{cgcv}}_{m,\ell}/n$. The key lemma is the concentration of $\widehat{\mathsf{df}}_m$ around its average $\widetilde{\mathsf{df}}_M = \sum_m \widehat{\mathsf{df}}_m/M$.

Lemma 13. Suppose the same penalty is used for $(\boldsymbol{h}_m)_{m=1}^M$. Then, we have

$$
\text{for all } m\in[M], \quad \text{for all } \epsilon>0, \quad \mathbb{P}\Big(\frac{|\widehat{\mathsf{df}}_m - \widetilde{\mathsf{df}}_M|}{n} > \epsilon\Big) \lesssim M\Big(e^{-nc/2} + \frac{\gamma^{9/2}}{\epsilon\sqrt{n}\tau^3c^4}\Big).
$$

The proof of Lemma 13 is given in Section A.5.6. From Lemma 13, it suffices to bound $|d^{\text{full}}_{m,\ell} - d^{\text{cgcv}}_{m,\ell}|/n$ from above by $|\widehat{\mathsf{df}}_m - \widetilde{\mathsf{df}}_M|/n$ and $|\widehat{\mathsf{df}}_\ell - \widetilde{\mathsf{df}}_M|/n$ up to an absolute constant. Below, we prove (50) for $m=\ell$ and $m\neq\ell$ separately.

When $m=\ell$. Letting f be the function $f(x) = 1-2x+c^{-1}x^2 = (1-x)^2 + (c^{-1}-1)x^2$, we have

$$
\frac{d^{\text{full}}_{m,m}}{n} = 1 - 2\frac{\widehat{\mathsf{df}}_m}{n} + \frac{\widehat{\mathsf{df}}_m^2}{nk} = f(\frac{\widehat{\mathsf{df}}_m}{n}), \qquad \frac{d^{\text{cgcv}}_{m,m}}{n} = (1-\frac{\widetilde{\mathsf{df}}_M}{n})^2 + (c^{-1}-1)(\frac{\widetilde{\mathsf{df}}_M}{n})^2 = f(\frac{\widetilde{\mathsf{df}}_M}{n}).
$$

Here, $\widehat{\mathsf{df}}_m/n, \widetilde{\mathsf{df}}_M/n \in [0,c)$ by (30), while f is 2-Lipschitz on $[0,c]$ since $\sup_{x\in[0,c]}|f'(x)| = \sup_{x\in[0,c]} 2(1-x/c) = 2$. Thus, we obtain that, for all $\epsilon>0$,

$$
\mathbb{P}\Big(\frac{|d^{\text{full}}_{m,m} - d^{\text{cgcv}}_{m,m}|}{n} > \epsilon\Big) = \mathbb{P}\Big(\Big|f(\frac{\widehat{\mathsf{df}}_m}{n}) - f(\frac{\widetilde{\mathsf{df}}_M}{n})\Big| > \epsilon\Big) \leq \mathbb{P}\Big(2\Big|\frac{\widehat{\mathsf{df}}_m}{n} - \frac{\widetilde{\mathsf{df}}_M}{n}\Big| > \epsilon\Big) \lesssim M\Big(e^{-nc/2} + \frac{\gamma^{9/2}}{\epsilon\sqrt{n}\tau^3c^4}\Big),
$$

thanks to the Lipschitz property of f and Lemma 13. This completes the proof of (50) for $m=\ell$.

When $m\neq\ell$. Letting g be the function $g(x,y) = (1-x)(1-y)$, we have

$$
\frac{d^{\text{full}}_{m,\ell}}{n} = g\Big(\frac{\widehat{\mathsf{df}}_m}{n}, \frac{\widehat{\mathsf{df}}_\ell}{n}\Big) + \frac{\widehat{\mathsf{df}}_m}{n}\frac{\widehat{\mathsf{df}}_\ell}{n}\Big(\frac{|I_m\cap I_\ell|\cdot n}{k^2} - 1\Big), \qquad \frac{d^{\text{cgcv}}_{m,\ell}}{n} = g\Big(\frac{\widetilde{\mathsf{df}}_M}{n}, \frac{\widetilde{\mathsf{df}}_M}{n}\Big).
$$

Here, $g(x,y) = (1-x)(1-y)$ satisfies the following inequality: for all $x, x', y, y' \in [0,1]$,

$$
|g(x,y) - g(x',y')| \leq |g(x,y) - g(x',y)| + |g(x',y) - g(x',y')| \leq |x-x'| + |y-y'|.
$$

From this property of g and $\widehat{\mathsf{df}}_m/n, \widehat{\mathsf{df}}_\ell/n, \widetilde{\mathsf{df}}_M/n \in [0,c)\subset[0,1]$ by (30), we find

$$
\begin{aligned}
\frac{|d^{\text{full}}_{m,\ell} - d^{\text{cgcv}}_{m,\ell}|}{n} &\leq \frac{\widehat{\mathsf{df}}_m}{n}\frac{\widehat{\mathsf{df}}_\ell}{n}\Big|\frac{|I_m\cap I_\ell|\cdot n}{k^2} - 1\Big| + \Big|g\Big(\frac{\widehat{\mathsf{df}}_m}{n}, \frac{\widehat{\mathsf{df}}_\ell}{n}\Big) - g\Big(\frac{\widetilde{\mathsf{df}}_M}{n}, \frac{\widetilde{\mathsf{df}}_M}{n}\Big)\Big|\\
&\leq c^2\Big|\frac{|I_m\cap I_\ell|\cdot n}{k^2} - 1\Big| + \frac{|\widehat{\mathsf{df}}_m - \widetilde{\mathsf{df}}_M|}{n} + \frac{|\widehat{\mathsf{df}}_\ell - \widetilde{\mathsf{df}}_M|}{n}.
\end{aligned}
$$

Here, thanks to the bound of of $\mathrm{Var}[|I_m \cap I_\ell|]$ (see Lemma 22), the moment of the first term on the right-hand side is bounded from above as

$$\mathbb{E}\Big[c^2\Big|\frac{|I_m \cap I_\ell|\cdot n}{k^2}-1\Big|\Big] = c^2\frac{n}{k^2}\mathbb{E}\Big[\Big||I_m\cap I_\ell| - \frac{k^2}{n}\Big|\Big] \le \frac{1}{n}\sqrt{\mathbb{E}\Big[\Big||I_m\cap I_\ell| - \frac{k^2}{n}\Big|^2\Big]} \le \frac{1}{n}\sqrt{\frac{k^2}{n}} = \frac{c}{\sqrt{n}}.$$

Therefore, Markov's inequality applied with the above moment bound Lemma 13 results in

$$\begin{aligned}\mathbb{P}\Big(\frac{|d^{\mathrm{full}}_{m,\ell} - d^{\mathrm{cgcv}}_{m,\ell}|}{n} > \epsilon\Big) &\le \mathbb{P}\Big(c^2\Big|\frac{|I_m\cap I_\ell|\cdot n}{k^2} - 1\Big| > \frac{\epsilon}{3}\Big) + \mathbb{P}\Big(\frac{|\widehat{\mathsf{df}}_m - \widetilde{\mathsf{df}}_M|}{n} > \frac{\epsilon}{3}\Big) + \mathbb{P}\Big(\frac{|\widehat{\mathsf{df}}_\ell - \widetilde{\mathsf{df}}_M|}{n} > \frac{\epsilon}{3}\Big)\\ &\lesssim \frac{c}{\epsilon\sqrt{n}} + M\Big(e^{-nc/2} + \frac{\gamma^{9/2}}{\epsilon\sqrt{n}\tau^3c^4}\Big) \lesssim M\Big(e^{-nc/2} + \frac{\gamma^{9/2}}{\epsilon\sqrt{n}\tau^3 c^4}\Big),\end{aligned}$$

for all $\epsilon > 0$. This completes the proof of Equation (50) for $m \neq \ell$. □

A.4. Proof of Theorem 3 (restated here for convenience)

Theorem 3 (Inconsistency of GCV for finite ensembles) *Suppose Assumptions A–D hold, and assume that the same penalty g_m is used for each $m \in [M]$ in (2). Let $c = k/n$, $\gamma = \max(1, p/n)$, $\tau = \min(1,\mu)$. Then, there exists an absolute constant $C > 0$ such that the following holds: The probability that $\widetilde{R}^{\mathrm{gcv}}_M$ in (9) over-estimates the risk R_M is lower bounded as:*

$$\text{for all } \delta\in(0,1),\quad \mathbb{P}\Big(\frac{\widetilde{R}^{\mathrm{gcv}}_M}{R_M} \ge 1 + \delta^2\frac{c(1-c)}{2M}\Big) \ge \mathbb{P}\Big(\frac{\widetilde{\mathsf{df}}_M}{k}\ge\delta\Big) - C\frac{M^5\gamma^{15/2}}{\sqrt{n}\tau^5c^5(1-c)\delta^2}.$$

Therefore, if there exists a positive constant δ_0 independent of n such that $\mathbb{P}(\widetilde{\mathsf{df}}_M/k \ge \delta_0) \ge \delta_0$, and $\{M, \mu^{-1}, p/n, c^{-1}, (1-c)^{-1}\}$ are bounded by a constant independent of n, $\widetilde{R}^{\mathrm{gcv}}_M$ is inconsistent and over-estimates the risk R_M with positive probability, in the sense that

$$\liminf_{n\to\infty}\ \mathbb{P}\Big(\frac{\widetilde{R}^{\mathrm{gcv}}_M}{R_M} \ge 1 + \frac{\delta_0^2 c(1-c)}{2M}\Big) \ge \delta_0.$$

Proof. Define Correction as

$$\mathrm{Correction} := (c^{-1}-1)\frac{(\widetilde{\mathsf{df}}_M/n)^2}{(1-\widetilde{\mathsf{df}}_M/n)^2}\frac{\sum_m\|\boldsymbol{h}_m\|^2}{\|\sum_m \boldsymbol{h}_m\|^2},$$

so that Lemma 12 can be written as

$$\mathbb{P}\Big(\Big|\frac{\widetilde{R}^{\mathrm{gcv}}_M}{R_M} - \mathrm{Correction} - 1\Big| > \epsilon\Big) \lesssim \frac{M^4\gamma^{13/2}}{\epsilon\sqrt{n}\tau^5c^4}\ \text{ for all } \epsilon\in(0,1). \tag{51}$$

Since $\|\sum_m \boldsymbol{h}_m\|^2 \le (\sum_m\|\boldsymbol{h}_m\|)^2 \le M\sum_m\|\boldsymbol{h}_m\|^2$ by the triangle inequality, it holds that

$$\forall\delta\in(0,1),\quad \Big\{\widetilde{\mathsf{df}}_M/k \ge \delta\Big\} = \Big\{\widetilde{\mathsf{df}}_M/n\ge c\delta\Big\} \subset \Big\{\mathrm{Correction} \ge (c^{-1}-1)\Big(\frac{c\delta}{1-c\delta}\Big)^2\frac{1}{M} \ge \frac{\delta^2c(1-c)}{M}\Big\}.$$

Therefore, we obtain that, for all $\delta\in(0,1)$,

$$\begin{aligned}\mathbb{P}\Big(\frac{\widetilde{\mathsf{df}}_M}{k}\ge\delta\Big) &\le \mathbb{P}\Big(\mathrm{Correction} \ge \frac{\delta^2c(1-c)}{M}\Big)\\ &\le \mathbb{P}\Big(\frac{\widetilde{R}^{\mathrm{gcv}}_M}{R_M} - \mathrm{Correction} - 1 \le -\frac{\delta^2c(1-c)}{2M}\Big) + \mathbb{P}\Big(\frac{\widetilde{R}^{\mathrm{gcv}}_M}{R_M} - 1 > \frac{\delta^2c(1-c)}{2M}\Big)\\ &\le C\Big(\frac{\delta^2c(1-c)}{2M}\Big)^{-1}\frac{M^4\gamma^{13/2}}{\sqrt{n}\tau^5c^4} + \mathbb{P}\Big(\frac{\widetilde{R}^{\mathrm{gcv}}_M}{R_M} > 1 + \frac{c(1-c)\delta^2}{2M}\Big)\ \Big(\text{by (51) with } \epsilon = \frac{\delta^2c(1-c)}{2M}\Big)\\ &\le C'\frac{M^5}{\sqrt{n}\tau^5\delta^2}\frac{\gamma^{13/2}}{c^5(1-c)} + \mathbb{P}\Big(\frac{\widetilde{R}^{\mathrm{gcv}}_M}{R_M} > 1 + \frac{c(1-c)\delta^2}{2M}\Big),\end{aligned}$$

where $C, C' > 0$ are absolute constants. This finishes the proof. □

A.5. Technical lemmas and their proofs

A.5.1. Proof of Lemma 7 (restated here for convenience)

Lemma 7. [Variant of Bellec and Shen (2022, Theorem 1)] Suppose the penalty g satisfies Assumption D with a constant $\mu > 0$. Then there exists a matrix $\boldsymbol{B} \in \mathbb{R}^{(p+1)\times(p+1)}$ depending on $(\boldsymbol{z}_i)_{i\in I}$ such that

$$\|\boldsymbol{B}\|_{\mathrm{op}} \le (k\mu)^{-1}, \quad \mathrm{rank}(\boldsymbol{B}) \le p, \quad \mathrm{tr}[\boldsymbol{B}] \ge 0 \tag{28}$$

and the derivative of $\boldsymbol{h}$ with respect to $\boldsymbol{Z} = (z_{ij})_{i\in[n],j\in[p+1]}$ is given by

$$\text{for all } i \in [n] \text{ and } j \in [p+1], \quad \frac{\partial \boldsymbol{h}}{\partial z_{ij}} = \boldsymbol{B}\boldsymbol{e}_j\boldsymbol{e}_i^\top \boldsymbol{L}\boldsymbol{r} - \boldsymbol{B}\boldsymbol{Z}^\top \boldsymbol{L}\boldsymbol{e}_i\boldsymbol{e}_j^\top \boldsymbol{h} \quad \text{with} \quad \boldsymbol{L} = \boldsymbol{L}_I = \sum_{i\in I} \boldsymbol{e}_i\boldsymbol{e}_i^\top. \tag{29}$$

Furthermore, $\boldsymbol{V} := \boldsymbol{I}_n - \boldsymbol{Z}\boldsymbol{B}\boldsymbol{Z}^\top \boldsymbol{L}$ satisfies the following:

$$\widehat{\mathsf{df}} = k - \mathrm{tr}[\boldsymbol{L}\boldsymbol{V}], \quad 0 < k\left(1 + \frac{\|\boldsymbol{L}\boldsymbol{G}\|_{op}^2}{k\mu}\right)^{-1} \le \mathrm{tr}[\boldsymbol{L}\boldsymbol{V}] \le k \quad \text{and} \quad \|\boldsymbol{L}\boldsymbol{V}\|_{\mathrm{op}} \le 1, \tag{30}$$

where $\boldsymbol{G} = \boldsymbol{X}\boldsymbol{\Sigma}^{-1/2} \in \mathbb{R}^{n\times p}$ is a standard Gaussian matrix.

Proof. By the change of variable $\boldsymbol{\beta} \mapsto \boldsymbol{u} = \boldsymbol{\Sigma}^{1/2}(\boldsymbol{\beta} - \boldsymbol{\beta}_0)$ and $\boldsymbol{G} = \boldsymbol{X}\boldsymbol{\Sigma}^{-1/2}$, we have

$$\boldsymbol{\Sigma}^{1/2}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) = \widehat{\boldsymbol{u}}, \quad \boldsymbol{r} = \boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}} = \boldsymbol{\epsilon} - \boldsymbol{G}\widehat{\boldsymbol{u}}, \quad \widehat{\mathsf{df}} = \mathrm{tr}[\boldsymbol{X}(\partial/\partial\boldsymbol{y})\widehat{\boldsymbol{\beta}}] = \mathrm{tr}[\boldsymbol{G}(\partial/\partial\boldsymbol{\epsilon})\widehat{\boldsymbol{u}}],$$

where $\widehat{\boldsymbol{u}}$ is a penalized estimator with an isotropic design $\boldsymbol{G} = \boldsymbol{X}\boldsymbol{\Sigma}^{-1/2}$:

$$\widehat{\boldsymbol{u}} = \widehat{\boldsymbol{u}}(\boldsymbol{\epsilon}, \boldsymbol{G}) := \operatorname*{argmin}_{\boldsymbol{u}\in\mathbb{R}^p} \frac{1}{k}\sum_{i\in I}(\epsilon_i - \boldsymbol{g}_i^\top \boldsymbol{u})^2 + f(\boldsymbol{u}) \text{ with } f(\boldsymbol{u}) := g(\boldsymbol{\Sigma}^{-1/2}\boldsymbol{u} + \boldsymbol{\beta}_0).$$

Note in passing that the map $\boldsymbol{u} \in \mathbb{R}^p \mapsto f(\boldsymbol{u}) - \mu\|\boldsymbol{u}\|^2/2$ is convex thanks to Assumption D. Then, Bellec and Shen (2022, Theorem 1) with $\boldsymbol{\Sigma} = \boldsymbol{I}_p$ implies the followings: there exists a matrix $\boldsymbol{A} \in \mathbb{R}^{p\times p}$ depending on $(\epsilon_i, \boldsymbol{g}_i)_{i\in I}$ such that

$$\|\boldsymbol{A}\|_{\mathrm{op}} \le (|I|\mu)^{-1} = (k\mu)^{-1},$$

and the derivative of $\widehat{\boldsymbol{u}}$ with respect to $\boldsymbol{G} = (g_{ij})_{ij} \in \mathbb{R}^{n\times p}$ and $\boldsymbol{\epsilon} = (\epsilon_i) \in \mathbb{R}^n$ are given by

$$\text{for all } i \in [n], j \in [p], \quad \frac{\partial \widehat{\boldsymbol{u}}}{\partial g_{ij}}(\boldsymbol{\epsilon}, \boldsymbol{G}) = \begin{cases} \boldsymbol{A}[\boldsymbol{e}_j\boldsymbol{e}_i^\top \boldsymbol{r} - \boldsymbol{G}^\top \boldsymbol{e}_i\boldsymbol{e}_j\widehat{\boldsymbol{u}}] & (i \in I) \\ \boldsymbol{0}_p & (i \notin I) \end{cases} = \boldsymbol{A}[\boldsymbol{e}_j(\boldsymbol{L}\boldsymbol{r})_i - \boldsymbol{G}^\top \boldsymbol{L}\boldsymbol{e}_i\widehat{u}_j],$$

$$\text{for all } i \in [n], \quad \frac{\partial \widehat{\boldsymbol{u}}}{\partial \epsilon_i}(\boldsymbol{u}, \boldsymbol{G}) = \begin{cases} \boldsymbol{A}\boldsymbol{G}^\top \boldsymbol{e}_i & (i \in I) \\ \boldsymbol{0}_p & (i \notin I) \end{cases} = \boldsymbol{A}\boldsymbol{G}^\top \boldsymbol{L}\boldsymbol{e}_i,$$

where $\boldsymbol{L} = \sum_{i\in I} \boldsymbol{e}_i\boldsymbol{e}_i^\top$. Furthermore, equation (D.8) (and the following argument) in Bellec and Shen (2022) imply that $\boldsymbol{V} := \boldsymbol{I}_n - \boldsymbol{G}\boldsymbol{A}\boldsymbol{G}^\top \boldsymbol{L}$ satisfies

$$0 < k(1 + \|\boldsymbol{G}\|_{\mathrm{op}}^2/(k\mu))^{-1} \le \mathrm{tr}[\boldsymbol{L}\boldsymbol{V}] = k - \widehat{\mathsf{df}} \le k, \quad \|\boldsymbol{L}\boldsymbol{V}\|_{\mathrm{op}} \le 1.$$

Thus, if we define the matrix $\boldsymbol{B} \in \mathbb{R}^{(p+1)\times(p+1)}$ as $\boldsymbol{B} := \begin{pmatrix} \boldsymbol{A} & \boldsymbol{0}_p \\ \boldsymbol{0}_p^\top & 0 \end{pmatrix}$, the derivative of $\boldsymbol{h} = (\widehat{\boldsymbol{u}}^\top, -\sigma)^\top \in \mathbb{R}^{p+1}$ with respect to $\boldsymbol{Z} = [\boldsymbol{G}|\sigma^{-1}\boldsymbol{\epsilon}] = (z_{ij})_{ij} \in \mathbb{R}^{n\times(p+1)}$ can be written as

$$i \in [n], \ 1 \le j \le p, \quad \frac{\partial \boldsymbol{h}}{\partial z_{ij}} = \frac{\partial}{\partial g_{ij}}\begin{pmatrix} \widehat{\boldsymbol{u}} \\ -\sigma \end{pmatrix} = \begin{pmatrix} \boldsymbol{A}[\boldsymbol{e}_j(\boldsymbol{L}\boldsymbol{r})_i - \boldsymbol{G}^\top \boldsymbol{L}\boldsymbol{e}_i\widehat{h}_j] \\ 0 \end{pmatrix} = \boldsymbol{B}\Big[\begin{pmatrix} \boldsymbol{e}_j \\ 0 \end{pmatrix}(\boldsymbol{L}\boldsymbol{r})_i - \boldsymbol{Z}^\top \boldsymbol{L}\boldsymbol{e}_i\widehat{h}_j\Big],$$

$$i \in [n], \quad \frac{\partial \boldsymbol{h}}{\partial z_{i,p+1}} = \sigma \frac{\partial}{\partial \epsilon_i} \begin{pmatrix} \widehat{\boldsymbol{u}} \\ -\sigma \end{pmatrix} = \begin{pmatrix} \sigma \boldsymbol{A}\boldsymbol{G}^\top \boldsymbol{L}\boldsymbol{e}_i \\ 0 \end{pmatrix} = \boldsymbol{B}\Big[\begin{pmatrix} \boldsymbol{0}_p \\ 1 \end{pmatrix} (\boldsymbol{L}\boldsymbol{r})_i - \boldsymbol{Z}^\top \boldsymbol{L}\boldsymbol{e}_i \widehat{h}_{p+1}\Big].$$

This finishes the proof of the derivative formula (29). It remains to prove $\operatorname{tr}[\boldsymbol{B}] \geq 0$ in (28). By the definition of $\boldsymbol{B}$, it suffices to show $\operatorname{tr}[\boldsymbol{A}] \geq 0$. Equation (7.4) in Bellec (2022) implies $\boldsymbol{v}^\top \boldsymbol{A}\boldsymbol{v} \geq 0$ for all $\boldsymbol{v} \in \ker(\boldsymbol{A})^\perp$. Then, letting $\boldsymbol{V} = (\boldsymbol{v}_1, \boldsymbol{v}_2, \dots, \boldsymbol{v}_p) \in \mathbb{R}^{p\times p}$ be an orthogonal matrix with its columns including an orthogonal basis of $\ker(\boldsymbol{A})^\perp$, we have $0 \leq \sum_{i=1}^p \boldsymbol{v}_i^\top \boldsymbol{A}\boldsymbol{v}_i = \operatorname{tr}[\boldsymbol{V}^\top \boldsymbol{A}\boldsymbol{V}] = \operatorname{tr}[\boldsymbol{A}\boldsymbol{V}\boldsymbol{V}^\top] = \operatorname{tr}[\boldsymbol{A}]$. This completes the proof. □

A.5.2. Proof of Lemma 8 (restated here for convenience)

Lemma 8. For any subset $J \subset [n]$ that is independent of $\boldsymbol{Z}$, we have

$$\mathbb{E}\Big[\Big(\frac{-\boldsymbol{r}^\top \boldsymbol{L}_J \widetilde{\boldsymbol{r}} - \widetilde{\boldsymbol{r}}^\top \widetilde{\boldsymbol{L}}\boldsymbol{L}_J \boldsymbol{r} \operatorname{tr}[\widetilde{\boldsymbol{B}}] + \operatorname{tr}[\boldsymbol{L}_J \boldsymbol{V}]\boldsymbol{h}^\top \widetilde{\boldsymbol{h}}}{\|\boldsymbol{h}\|\|\widetilde{\boldsymbol{h}}\|}\Big)^2 \ \Big|\ J\Big] \lesssim \tau^{-2}(|J| + \frac{p(|J|+p)^2}{k^2})$$

where $\tau = \min(1, \mu)$ and $\boldsymbol{L}_J = \sum_{i\in J} \boldsymbol{e}_i \boldsymbol{e}_i^\top$.

Proof. The proof is based on the moment inequality in Lemma 14 ahead. We will bound Ξ_J in Lemma 14 using the derivative formula (29) and the bound of the operator norm $\|\boldsymbol{B}\|_{\mathrm{op}} \leq (k\mu)^{-1}$ by (28). Note in passing that the derivative formula (29) implies

$$\sum_{i\in J}\sum_{j=1}^{p+1} \Big\|\frac{\partial \boldsymbol{h}}{\partial z_{ij}}\Big\|^2 = \sum_{i\in J}\sum_{j=1}^{p+1} \|\boldsymbol{B}\boldsymbol{e}_j(\boldsymbol{L}\boldsymbol{r})_i - \boldsymbol{B}\boldsymbol{Z}^\top \boldsymbol{L}\boldsymbol{e}_i h_j\|^2 \lesssim \|\boldsymbol{B}\|_F^2 \|\boldsymbol{L}_J \boldsymbol{r}\|^2 + \|\boldsymbol{L}_J \boldsymbol{Z}\boldsymbol{B}^\top\|_F^2 \|\boldsymbol{h}\|^2,$$

so that $\boldsymbol{r} = -\boldsymbol{Z}\boldsymbol{h}$ and $\|\boldsymbol{B}\|_F^2 \leq \operatorname{rank}(\boldsymbol{B})\|\boldsymbol{B}\|_{\mathrm{op}}^2 \leq p(k\mu)^{-2}$ lead to

$$\Xi_J = \sum_{i\in J} \|\boldsymbol{h}\|^{-2} \sum_{j=1}^{p+1} \Big\|\frac{\partial \boldsymbol{h}}{\partial z_{ij}}\Big\|^2 \lesssim \|\boldsymbol{B}\|_F^2 \|\boldsymbol{L}_J \boldsymbol{Z}\|_{\mathrm{op}}^2 \lesssim p(k\mu)^{-2}\|\boldsymbol{L}_J \boldsymbol{Z}\|_{\mathrm{op}}^2. \tag{52}$$

Thus, Lemma 14 and $\mathbb{E}[\|\boldsymbol{L}_J \boldsymbol{Z}\|_{\mathrm{op}}^4] \lesssim (|J|+p)^2$ by (20) yield

$$\mathbb{E}\Big[\frac{1}{\|\boldsymbol{h}\|^2\|\widetilde{\boldsymbol{h}}\|^2}\Big(\boldsymbol{r}\boldsymbol{L}_J \widetilde{\boldsymbol{r}} + \sum_{i\in J}\sum_{j=1}^{q} \frac{\partial r_i \widetilde{h}_j}{\partial z_{ij}}\Big)^2 \ \Big|\ J\Big] \lesssim |J| + p(|J|+p)^2(k\mu)^{-2} \lesssim \tau^{-2}(|J| + p(|J|+p)^2/k^2),$$

thanks to $\tau = \min(1, \mu)$. It remains to bound the error inside the square in the LHS; the derivative formula (29) leads to

$$\frac{1}{\|\boldsymbol{h}\|\|\widetilde{\boldsymbol{h}}\|}\sum_{i\in K}\sum_{j=1}^{p+1}\Big(\frac{\partial(\widetilde{h}_j r_i)}{\partial z_{ij}} - \widetilde{\boldsymbol{r}}^\top \widetilde{\boldsymbol{L}}\boldsymbol{L}_J \boldsymbol{r} \operatorname{tr}[\widetilde{\boldsymbol{B}}] + \operatorname{tr}[\boldsymbol{L}_J \boldsymbol{V}]\widetilde{\boldsymbol{h}}^\top \boldsymbol{h}\Big) = \frac{-\widetilde{\boldsymbol{h}}^\top \widetilde{\boldsymbol{B}}\boldsymbol{Z}^\top \widetilde{\boldsymbol{L}}\boldsymbol{L}_J \boldsymbol{r} - \boldsymbol{r}^\top \boldsymbol{L}\boldsymbol{L}_J \boldsymbol{Z}\boldsymbol{B}\widetilde{\boldsymbol{h}}}{\|\boldsymbol{h}\|\|\widetilde{\boldsymbol{h}}\|}.$$

Here, the square of the RHS can be bounded from above by $4\mu^{-2}k^{-2}\|\boldsymbol{L}_J \boldsymbol{Z}\|_{\mathrm{op}}^4$. From this and $\mathbb{E}[\|\boldsymbol{L}_J \boldsymbol{Z}\|_{\mathrm{op}}^4] \lesssim (|J|+p)^2$ by (20), we find that the expectation of the RHS is bounded from above by $(|J|+p)^2/(k\mu)^{-2}$ up to some absolute constant. This finishes the proof. □

Lemma 14 (Variant of Bellec (2023, Proposition 6.1))**.** Assume that $\boldsymbol{h}$ and $\widetilde{\boldsymbol{h}}$ are locally Lipschitz function from $\mathbb{R}^{n\times q} \to \mathbb{R}^q$ such that $\|\boldsymbol{h}\|^2, \|\widetilde{\boldsymbol{h}}\|^2 \neq 0$, and let $\boldsymbol{r} = -\boldsymbol{Z}\boldsymbol{h}$ and $\widetilde{\boldsymbol{r}} = -\boldsymbol{Z}\widetilde{\boldsymbol{h}}$, where $\boldsymbol{Z} \in \mathbb{R}^{n\times q}$ has i.i.d. $\mathcal{N}(0,1)$ entries. If $J \subset [n]$ is independent of $\boldsymbol{Z}$, we have

$$\mathbb{E}\Big[\frac{1}{\|\boldsymbol{h}\|^2\|\widetilde{\boldsymbol{h}}\|^2}\Big(\boldsymbol{r}\boldsymbol{L}_J \widetilde{\boldsymbol{r}} + \sum_{i\in J}\sum_{j=1}^{q} \frac{\partial r_i \widetilde{h}_j}{\partial z_{ij}}\Big)^2 \ \Big|\ J\Big] \lesssim |J| + \mathbb{E}[\|\boldsymbol{L}_J \boldsymbol{Z}\|_{\mathrm{op}}^2(1 + \Xi_J + \widetilde{\Xi}_J)],$$

where $\boldsymbol{L}_J = \sum_{i\in J} \boldsymbol{e}_i \boldsymbol{e}_i^\top$, $\Xi_J = \sum_{i\in J}\sum_{j=1}^q \|\boldsymbol{h}\|^{-2}\Big\|\frac{\partial \boldsymbol{h}}{\partial z_{ij}}\Big\|^2$, and $\widetilde{\Xi}_J = \sum_{i\in J}\sum_{j=1}^q \|\widetilde{\boldsymbol{h}}\|^{-2}\Big\|\frac{\partial \widetilde{\boldsymbol{h}}}{\partial z_{ij}}\Big\|^2$.

Proof. Let $\boldsymbol{\varrho} = \boldsymbol{L}_J \boldsymbol{r}/\|\boldsymbol{h}\| = -\boldsymbol{L}_J \boldsymbol{Z}\boldsymbol{h}/\|\boldsymbol{h}\| \in \mathbb{R}^n$ and $\widetilde{\boldsymbol{\eta}} := \widetilde{\boldsymbol{h}}/\|\widetilde{\boldsymbol{h}}\| \in \mathbb{R}^q$. Then, Proposition 6.1. in Bellec (2023) implies

$$\mathbb{E}\Big[\Big(-\boldsymbol{\varrho}^\top \boldsymbol{Z}\widetilde{\boldsymbol{\eta}} + \sum_{i,j}\frac{\partial(\rho_i\widetilde{\eta}_j)}{\partial z_{ij}}\Big)^2\Big] \le \mathbb{E}[\|\boldsymbol{\varrho}\|^2\|\widetilde{\boldsymbol{\eta}}\|^2] + \sum_{ij}\mathbb{E}\Big[\|\boldsymbol{\varrho}\|^2\Big\|\frac{\partial \widetilde{\boldsymbol{\eta}}}{\partial z_{ij}}\Big\| + \|\widetilde{\boldsymbol{\eta}}\|^2\Big\|\frac{\partial \boldsymbol{\varrho}}{\partial z_{ij}}\Big\|^2\Big], \tag{53}$$

where $\mathbb{E}[\|\boldsymbol{\varrho}\|^2\|\widetilde{\boldsymbol{\eta}}\|^2] \le \mathbb{E}[\|\boldsymbol{L}_J\boldsymbol{Z}\|_{\text{op}}^2]$ thanks to $\|\widetilde{\boldsymbol{\eta}}\|^2 = 1$ and $\|\boldsymbol{\varrho}\|^2 \le \|\boldsymbol{L}_J\boldsymbol{Z}\|_{\text{op}}^2$. It remains to bound the second term. Here, we use the following identity: if $\boldsymbol{h}: \mathbb{R}^{n\times q} \to \mathbb{R}^q$ is locally Lipschitz with $\|\boldsymbol{h}\|^2 \ne 0$, we have

$$\frac{\partial}{\partial z_{ij}}\Big(\frac{\boldsymbol{h}}{\|\boldsymbol{h}\|}\Big) = \frac{\boldsymbol{P}^\perp}{\|\boldsymbol{h}\|}\frac{\partial \boldsymbol{h}}{\partial z_{ij}} \text{ with } \boldsymbol{P}^\perp = \boldsymbol{I}_{p+1} - \frac{\boldsymbol{h}\boldsymbol{h}^\top}{\|\boldsymbol{h}\|^2} \text{ so that } \Big\|\frac{\partial}{\partial z_{ij}}\Big(\frac{\boldsymbol{h}}{\|\boldsymbol{h}\|}\Big)\Big\|^2 \le \frac{1}{\|\boldsymbol{h}\|^2}\Big\|\frac{\partial \boldsymbol{h}}{\partial z_{ij}}\Big\|^2, \tag{54}$$

where the inequality follows from $\|\boldsymbol{P}^\perp\|_{\text{op}} \le 1$. Then, (54) and $\|\boldsymbol{\varrho}\|^2 \le \|\boldsymbol{L}_J\boldsymbol{Z}\|_{\text{op}}^2$ yield

$$\sum_{i\in J}\sum_{j=1}^q \|\boldsymbol{\varrho}\|^2\Big\|\frac{\partial \widetilde{\boldsymbol{\eta}}}{\partial z_{ij}}\Big\| \le \|\boldsymbol{L}_J\boldsymbol{Z}\|_{\text{op}}^2 \sum_{i\in J}\sum_j \frac{1}{\|\widetilde{\boldsymbol{h}}\|^2}\Big\|\frac{\partial \widetilde{\boldsymbol{h}}}{\partial z_{ij}}\Big\|^2 = \|\boldsymbol{L}_J\boldsymbol{Z}\|_{\text{op}}^2 \Xi_J.$$

In the same way, $\boldsymbol{\varrho} = -\boldsymbol{L}_J\boldsymbol{Z}\boldsymbol{h}/\|\boldsymbol{h}\|$, $\|\widetilde{\boldsymbol{\eta}}\|^2 = 1$, and (54) lead to

$$\sum_{i\in J}\sum_{j=1}^q \|\widetilde{\boldsymbol{\eta}}\|^2\Big\|\frac{\partial \boldsymbol{\varrho}}{\partial z_{ij}}\Big\|^2 = \sum_{i\in J}\sum_{j=1}^q \Big\|\boldsymbol{L}_J\Big(\boldsymbol{e}_i\frac{h_j}{\|\boldsymbol{h}\|} + \boldsymbol{Z}\frac{\partial}{\partial z_{ij}}\Big(\frac{\boldsymbol{h}}{\|\boldsymbol{h}\|}\Big)\Big)\Big\|^2 \le 2|J| + 2\|\boldsymbol{L}_J\boldsymbol{Z}\|_{\text{op}}^2\Xi_J.$$

Therefore, the RHS of (53) is bounded from above by $|J| + \mathbb{E}[\|\boldsymbol{L}_J\boldsymbol{Z}\|_{\text{op}}^2(1 + \Xi_J + \widetilde{\Xi}_J)]$ up to some absolute constant. It remains to control the error inside the square:

$$\frac{1}{\|\boldsymbol{h}\|\|\widetilde{\boldsymbol{h}}\|}\sum_{i\in J}\sum_{j\in[q]}\frac{\partial r_i\widetilde{h}_j}{\partial z_{ij}} - \sum_{i\in J}\sum_{j\in[q]}\frac{\partial \rho_i\widetilde{\eta}_j}{\partial z_{ij}} = \sum_{i\in J}\sum_{j=1}^q \frac{r_i\widetilde{h}_j}{\|\boldsymbol{h}\|\|\widetilde{\boldsymbol{h}}\|}\Big(\frac{\boldsymbol{h}^\top}{\|\boldsymbol{h}\|^2}\frac{\partial \boldsymbol{h}}{\partial z_{ij}} + \frac{\widetilde{\boldsymbol{h}}^\top}{\|\widetilde{\boldsymbol{h}}\|^2}\frac{\partial \widetilde{\boldsymbol{h}}}{\partial z_{ij}}\Big)$$

By multiple applications of the Cauchy-Schwartz inequality, the square of the RHS is bounded from above by $2\|\boldsymbol{L}_J\boldsymbol{Z}\|_{\text{op}}^2(\Xi_J + \widetilde{\Xi}_J)$. This finishes the proof. □

A.5.3. Proof of Lemma 9 (restated here for convenience)

Lemma 9. Let $\xi_1 = (1+\text{tr}[\boldsymbol{B}])\|\boldsymbol{L}\boldsymbol{r}\|^2/\|\boldsymbol{h}\|^2 - \text{tr}[\boldsymbol{L}\boldsymbol{V}]$ and $\xi_2 = k - (1+\text{tr}[\boldsymbol{B}])^2\|\boldsymbol{L}\boldsymbol{r}\|^2/\|\boldsymbol{h}\|^2$. Then

$$\mathbb{E}|k - \text{tr}[\boldsymbol{L}\boldsymbol{V}](1+\text{tr}[\boldsymbol{B}])| = \mathbb{E}|(1+\text{tr}[\boldsymbol{B}])\xi_1 + \xi_2| \lesssim \sqrt{k}\tau^{-2}(1+p/k)^{5/2}.$$

Proof. Lemma 8 with $J = I = \widetilde{I}$ and $0 \le \text{tr}[\boldsymbol{B}] \le p\|\boldsymbol{B}\|_{\text{op}} \le p/(k\mu)$ by (28) yield

$$\mathbb{E}|(1+\text{tr}[\boldsymbol{B}])\xi_1| \lesssim \mu^{-1}(p/k)\mathbb{E}|\xi_1| \lesssim \tau^{-1}(p/k)\sqrt{k}\tau^{-1}(1+p/k)^{3/2} \lesssim \sqrt{k}\tau^{-2}(1+p/k)^{5/2},$$

while $\mathbb{E}|\xi_2| \lesssim \sqrt{k}\tau^{-2}(1+p/k)^2$ by Lemma 15 below. Thus, we obtain

$$\mathbb{E}[|k - \text{tr}[\boldsymbol{L}\boldsymbol{V}](1+\text{tr}[\boldsymbol{B}])|] \lesssim \sqrt{k}\tau^{-2}(1+p/k)^{5/2} + \sqrt{k}\tau^{-2}(1+p/k)^2 \lesssim \sqrt{k}\tau^{-2}(1+p/k)^{5/2}.$$

This finishes the proof. □

Lemma 15. We have $\mathbb{E}[|k - (1+\text{tr}[\boldsymbol{B}])^2\|\boldsymbol{L}\boldsymbol{r}\|^2/\|\boldsymbol{h}\|^2|] \lesssim k^{1/2}\tau^{-2}(1+p/k)^2$.

Proof. Define $\boldsymbol{a} \in \mathbb{R}^n$ and $\boldsymbol{b} \in \mathbb{R}^n$ as

$$\text{for all } i \in [n], \quad a_i = \frac{1}{\|\boldsymbol{h}\|}(1 + \operatorname{tr}[\boldsymbol{B}])(\boldsymbol{L}\boldsymbol{r})_i \quad \text{and} \quad b_i = \frac{1}{\|\boldsymbol{h}\|}(\boldsymbol{L}\boldsymbol{r})_i + \frac{1}{\|\boldsymbol{h}\|}\sum_{j=1}^{p+1} \frac{\partial h_j}{\partial z_{ij}}$$

so that $(1 + \operatorname{tr}[\boldsymbol{B}])^2 \|\boldsymbol{L}\boldsymbol{r}\|^2 / \|\boldsymbol{h}\|^2 = \|\boldsymbol{a}\|^2$. Lemma 17 ahead and the Cauchy–Schwarz inequality yield

$$\begin{aligned}\mathbb{E}[|k - \|\boldsymbol{a}\|^2|] &\le \mathbb{E}[2(\sqrt{k}\|\boldsymbol{a} - \boldsymbol{b}\| + |k - \|\boldsymbol{b}\|^2| + \|\boldsymbol{a} - \boldsymbol{b}\|^2] \\ &\lesssim \sqrt{k\mathbb{E}[\|\boldsymbol{a} - \boldsymbol{b}\|^2]} + \mathbb{E}[\|\boldsymbol{a} - \boldsymbol{b}\|^2] + \mathbb{E}[|k - \|\boldsymbol{b}\|^2|]. \end{aligned} \tag{55}$$

Thus, it suffices to bound $\mathbb{E}[\|\boldsymbol{a} - \boldsymbol{b}\|^2]$ and $\mathbb{E}[|k - \|\boldsymbol{b}\|^2]$. For $\mathbb{E}\|\boldsymbol{a} - \boldsymbol{b}\|^2$, (28), (29), and Lemma 20 yield

$$\mathbb{E}\|\boldsymbol{a} - \boldsymbol{b}\|^2 = \mathbb{E}[\|\boldsymbol{h}\|^{-2}\|\boldsymbol{L}\boldsymbol{Z}\boldsymbol{B}^\top \boldsymbol{h}\|^2] \le \mathbb{E}[\|\boldsymbol{B}\|_{\mathrm{op}}^2 \|\boldsymbol{L}\boldsymbol{Z}\|_{\mathrm{op}}^2] \lesssim (k\mu)^{-2}(k+p) \lesssim k^{-1}\tau^{-2}(1 + p/k),$$

where $\tau = \min(1, \mu)$. If we denote $\|\boldsymbol{h}\|^{-2}\sum_{i \in I}\sum_{j=1}^{p+1} \|(\partial/\partial z_{ij})\boldsymbol{h}\|^2$ by Ξ_I, Lemma 16 below leads to

$$\begin{aligned}\mathbb{E}[|k - \|\boldsymbol{b}\|^2|] &\lesssim \sqrt{k(1 + \mathbb{E}[\Xi_I])} + \mathbb{E}[\Xi_I] \text{ (by Lemma 16)} \\ &\lesssim \sqrt{k(1 + \mu^{-2}(1 + p/k)^2)} + \mu^{-2}(1 + p/k)^2 \text{ (thanks to (52) with } J = I) \\ &\lesssim \sqrt{k}\tau^{-2}(1 + p/k)^2 \text{ (thanks to } \tau = \min(1, \mu)).\end{aligned}$$

This concludes the proof. □

Lemma 16 (Variant of Bellec (2023, Theorem 7.1))**.** Let $\boldsymbol{h} : \mathbb{R}^{k \times q} \to \mathbb{R}^q$ be a locally Lipschitz functions. If $\boldsymbol{Z} \in \mathbb{R}^{k \times q}$ has i.i.d. $\mathcal{N}(0, 1)$ entries, we have

$$\mathbb{E}\Big[\Big|k - \frac{1}{\|\boldsymbol{h}\|^2}\sum_{i=1}^k \Big(\boldsymbol{e}_i^\top \boldsymbol{Z}\boldsymbol{h} - \sum_{j=1}^q \frac{\partial h_j}{\partial z_{ij}}\Big)^2\Big|\Big] \lesssim \sqrt{k(1 + \mathbb{E}\Xi)} + \mathbb{E}\Xi \text{ with } \Xi = \frac{1}{\|\boldsymbol{h}\|^2}\sum_{i=1}^k\sum_{j=1}^q \Big\|\frac{\partial \boldsymbol{h}}{\partial z_{ij}}\Big\|^2.$$

Proof. Define vectors $\boldsymbol{a}, \boldsymbol{b} \in \mathbb{R}^k$ by

$$\text{for all } i \in [k], \quad a_i = \frac{1}{\|\boldsymbol{h}\|}\Big(\boldsymbol{e}_i^\top \boldsymbol{Z}\boldsymbol{h} - \sum_j \frac{\partial h_j}{\partial z_{ij}}\Big) \quad \text{and} \quad b_i = \boldsymbol{e}_i^\top \boldsymbol{Z}\frac{\boldsymbol{h}}{\|\boldsymbol{h}\|} - \sum_j \frac{\partial}{\partial z_{ij}}\Big(\frac{h_j}{\|\boldsymbol{h}\|}\Big),$$

so that the LHS of the assertion is $\mathbb{E}|k - \|\boldsymbol{a}\|^2|$. The same argument in (55) leads to

$$\mathbb{E}[|k - \|\boldsymbol{a}\|^2|] \lesssim \sqrt{k\mathbb{E}[\|\boldsymbol{a} - \boldsymbol{b}\|^2]} + \mathbb{E}[\|\boldsymbol{a} - \boldsymbol{b}\|^2] + \mathbb{E}[|k - \|\boldsymbol{b}\|^2|].$$

Below we bound $\mathbb{E}[\|\boldsymbol{a} - \boldsymbol{b}\|^2]$ and $\mathbb{E}[|k - \|\boldsymbol{b}\|^2]$. For $\|\boldsymbol{a} - \boldsymbol{b}\|^2$, multiple applications of the Cauchy–Schwarz inequality lead to

$$\|\boldsymbol{a} - \boldsymbol{b}\|^2 = \sum_{i=1}^k\Big\{-\frac{1}{\|\boldsymbol{h}\|}\sum_j \frac{\partial h_j}{\partial z_{ij}} + \sum_j \frac{\partial}{\partial z_{ij}}\Big(\frac{h_j}{\|\boldsymbol{h}\|}\Big)\Big\}^2 = \sum_{i=1}^k\Big(-\sum_j \frac{\boldsymbol{h}^\top}{\|\boldsymbol{h}\|^3}\frac{\partial \boldsymbol{h}}{\partial z_{ij}}h_j\Big)^2 \le \sum_{i,j}\frac{1}{\|\boldsymbol{h}\|^2}\Big\|\frac{\partial \boldsymbol{h}}{\partial z_{ij}}\Big\|^2 = \Xi.$$

For $\mathbb{E}[|k - \|\boldsymbol{b}\|^2]$, Theorem 7.1 in Bellec (2023) applied to the unit vector $\boldsymbol{h}/\|\boldsymbol{h}\| \in \mathbb{R}^q$ implies

$$\mathbb{E}|k - \|\boldsymbol{b}\|^2| \lesssim \sqrt{k(1 + \mathbb{E}\Xi')} + \mathbb{E}\Xi' \quad \text{with} \quad \Xi' := \mathbb{E}\sum_{i=1}^k\sum_{j=1}^q \|\frac{\partial}{\partial z_{ij}}\Big(\frac{\boldsymbol{h}}{\|\boldsymbol{h}\|}\Big)\|^2.$$

Since $\Xi' \le \Xi$ by (54), we have $\mathbb{E}[|k - \|\boldsymbol{b}\|^2|] \lesssim \sqrt{k(1 + \mathbb{E}[\Xi])} + \mathbb{E}[\Xi]$. This finishes the proof. □

Lemma 17. We have $|k - \|\boldsymbol{a}\|^2| \le 2(\sqrt{k}\|\boldsymbol{a}-\boldsymbol{b}\| + |k - \|\boldsymbol{b}\|^2| + \|\boldsymbol{a}-\boldsymbol{b}\|^2)$ for all vector $\boldsymbol{a}, \boldsymbol{b}$ in the same Euclidean space.

Proof. By multiple applications of the triangle inequality and $2ab \le a^2 + b^2$, we have

$$\begin{aligned}\big||k - \|\boldsymbol{a}\|^2| - |k - \|\boldsymbol{b}\|^2|\big| &\le \|\boldsymbol{a}-\boldsymbol{b}\|\|\boldsymbol{a}+\boldsymbol{b}\| \le \|\boldsymbol{a}-\boldsymbol{b}\|^2 + 2\|\boldsymbol{a}-\boldsymbol{b}\|\|\boldsymbol{b}\| \\ &\le \|\boldsymbol{a}-\boldsymbol{b}\|^2 + 2\|\boldsymbol{a}-\boldsymbol{b}\|(\sqrt{|\|\boldsymbol{b}\|^2 - k|} + \sqrt{k}) \\ &\le 2\|\boldsymbol{a}-\boldsymbol{b}\|^2 + |\|\boldsymbol{b}\|^2 - k| + 2\|\boldsymbol{a}-\boldsymbol{b}\|\sqrt{k}.\end{aligned}$$

Adding $|\|\boldsymbol{b}\|^2 - k|$ to both sides and using the triangle inequality, we conclude the proof. □

A.5.4. Proof of Lemma 10 (restated here for convenience)

Lemma 10. If the same penalty is used for all $m \in [M]$, i.e., $g_m = g$ in (2), then

$$\eta_{m,\ell} := \mathbb{E}\Big[\frac{\boldsymbol{h}_m^\top \boldsymbol{h}_\ell}{\|\boldsymbol{h}_m\|\|\boldsymbol{h}_\ell\|} \mid (\boldsymbol{z}_i)_{i\in I_m\cap I_\ell}\Big] \ge 0, \quad \text{and} \quad \mathbb{E}\Big[\Big(\frac{\boldsymbol{h}_m^\top \boldsymbol{h}_\ell}{\|\boldsymbol{h}_m\|\|\boldsymbol{h}_\ell\|} - \eta_{m,\ell}\Big)^2\Big] \lesssim \frac{\gamma}{nc^2\tau^2}.$$

Proof. Letting $\boldsymbol{v} = \boldsymbol{h}/\|\boldsymbol{h}\|$ and $\widetilde{\boldsymbol{v}} = \widetilde{\boldsymbol{h}}/\|\widetilde{\boldsymbol{h}}\|$, the statement of Lemma 10 can be written as follows; if the same penalty is used for $\boldsymbol{h}$ and $\widetilde{\boldsymbol{h}}$, we have

$$\mathbb{E}\Big[\boldsymbol{v}^\top\widetilde{\boldsymbol{v}} \mid (\boldsymbol{z}_i)_{i\in I\cap\widetilde{I}}\Big] \ge 0, \quad \text{and} \quad \mathbb{E}\Big[\mathrm{Var}\Big[\boldsymbol{v}^\top\widetilde{\boldsymbol{v}} \mid (\boldsymbol{z}_i)_{i\in I\cap\widetilde{I}}\Big]\Big] \lesssim \frac{\gamma}{nc^2\tau^2},$$

Below we prove this claim. Here, the key fact is that conditionally on $(\boldsymbol{z}_i)_{i\in I\cap\widetilde{I}}$, the random vectors $\boldsymbol{v}$ and $\widetilde{\boldsymbol{v}}$ are independent and identically distributed. Then, it immediately follows from this fact that

$$\mathbb{E}\Big[\boldsymbol{v}^\top\widetilde{\boldsymbol{v}} \mid (\boldsymbol{z}_i)_{i\in I\cap\widetilde{I}}\Big] = \Big\|\mathbb{E}\Big[\boldsymbol{v} \mid (\boldsymbol{z}_i)_{i\in I\cap\widetilde{I}}\Big]\Big\|^2 \ge 0.$$

Next, we derive the bound of the variance: first using the Gaussian Poincaré inequality for the inequality below, second using that $\boldsymbol{v}$ does not depend on $(\boldsymbol{z}_i)_{i\in\widetilde{I}\setminus I}$ and $\widetilde{\boldsymbol{v}}$ does not depend on $(\boldsymbol{z}_i)_{i\in I\setminus\widetilde{I}}$ for the equality below,

$$\mathrm{Var}\Big[\boldsymbol{v}^\top\widetilde{\boldsymbol{v}} \mid (\boldsymbol{z}_i)_{i\in I\cap\widetilde{I}}\Big] \le \bar{\mathbb{E}} \sum_{j\in[p+1]} \sum_{i\in(I\setminus\widetilde{I})\cup(\widetilde{I}\setminus I)} \Big(\frac{\partial}{\partial z_{ij}}\boldsymbol{v}^\top\widetilde{\boldsymbol{v}}\Big)^2 = \bar{\mathbb{E}} \sum_{j\in[p+1]} \Big[\sum_{i\in I\setminus\widetilde{I}} \Big(\widetilde{\boldsymbol{v}}^\top\frac{\partial\boldsymbol{v}}{\partial z_{ij}}\Big)^2 + \sum_{i\in\widetilde{I}\setminus I}\Big(\boldsymbol{v}^\top\frac{\partial\widetilde{\boldsymbol{v}}}{\partial z_{ij}}\Big)^2\Big],$$

where $\bar{\mathbb{E}}$ is conditional expectation given $(\boldsymbol{z}_i)_{i\in I\cap\widetilde{I}}$. By the symmetry of $(\boldsymbol{h}, \widetilde{\boldsymbol{h}})$, it suffices to bound the first term. Letting $\boldsymbol{P}^\perp = \boldsymbol{I}_{p+1} - \boldsymbol{v}\boldsymbol{v}^\top$, we obtain the following from the identity (54):

$$\begin{aligned}\sum_{j\in[p+1]}\sum_{i\in I\setminus\widetilde{I}}\Big(\widetilde{\boldsymbol{v}}^\top\frac{\partial\boldsymbol{v}}{\partial z_{ij}}\Big)^2 &= \frac{1}{\|\boldsymbol{h}\|^2}\sum_{j\in[p+1]}\sum_{i\in I}(\widetilde{\boldsymbol{v}}^\top\boldsymbol{P}^\perp\boldsymbol{B}\boldsymbol{e}_j(\boldsymbol{L}\boldsymbol{r})_i - \widetilde{\boldsymbol{v}}^\top\boldsymbol{P}^\perp\boldsymbol{B}\boldsymbol{Z}^\top\boldsymbol{L}\boldsymbol{e}_i h_j)^2 \\ &\lesssim \|\boldsymbol{B}\|_{\mathrm{op}}^2\|\boldsymbol{L}\boldsymbol{Z}\|_{\mathrm{op}}^2 + \|\boldsymbol{L}\boldsymbol{Z}\boldsymbol{B}^\top\|_{\mathrm{op}}^2 \lesssim (k\mu)^{-2}\|\boldsymbol{L}\boldsymbol{Z}\|_{\mathrm{op}}^2.\end{aligned}$$

Thus, the moment bound $\mathbb{E}[\|\boldsymbol{L}\boldsymbol{Z}\|_{\mathrm{op}}^2] \lesssim (k+p)$ by Lemma 20 leads to

$$\mathbb{E}\Big[\mathrm{Var}\Big[\boldsymbol{v}^\top\widetilde{\boldsymbol{v}} \mid (\boldsymbol{z}_i)_{i\in I\cap\widetilde{I}}\Big]\Big] \lesssim (k\mu)^{-2}\mathbb{E}\Big[\|\boldsymbol{L}\boldsymbol{Z}\|_{\mathrm{op}}^2 + \|\widetilde{\boldsymbol{L}}\boldsymbol{Z}\|_{\mathrm{op}}^2\Big] \lesssim (k\mu)^{-2}(k+p) \lesssim n^{-1}\tau^{-2}c^{-2}\gamma,$$

thanks to $\tau = \min(1,\mu)$, $c = k/n \in (0,1)$, and $\gamma = \max(1, p/n)$. This completes the proof. □

A.5.5. Proof of Lemma 11 (restated here for convenience)

Lemma 11. Let $d_{m,\ell}^{\text{ovlp}}$ and $d_{m,\ell}^{\text{full}}$ be the random variables defined in (39). Then, there exists a positive absolute constant $C > 0$ such that for $\# \in \{\text{ovlp}, \text{full}\}$,

$$\mathbb{P}\Big(\frac{d_{m,\ell}^{\#}}{n} < Cd^{\#}(\tau, c, \gamma)\Big) \lesssim \frac{\gamma^4}{nc^2\tau^4} \text{ where } d^{\text{ovlp}}(\tau, c, \gamma) = \tau^2 c^4 \gamma^{-2}, \text{ and } d^{\text{full}}(\tau, c, \gamma) = \tau^2\gamma^{-2}.$$

Proof. Below, we prove this assertion for $\# = $ ovlp and $\# = $ full, separately.

Proof for $\# = $ ovlp. Recall that $d_{m,\ell}^{\text{ovlp}} = |I_m \cap I_\ell|(1 - \widehat{\mathsf{df}}_m/k)(1 - \widehat{\mathsf{df}}_\ell/k)$. Lemma 18 with $I = I_m$ and I_ℓ implies that there exists an absolute constant $C \in (0, 1)$ such that

$$\text{for all } m, \ell, \quad \mathbb{P}((1 - \widehat{\mathsf{df}}_m/k)(1 - \widehat{\mathsf{df}}_\ell/k) \le C^2\tau^2c^2\gamma^{-2}) \le 2e^{-nc/2}. \tag{56}$$

Lemma 22 (introduced later) and Markov's inequality lead to the following:

$$\text{for all } m \ne \ell, \quad \mathbb{P}(|I_m \cap I_\ell|n/k^2 - 1| > 1/2) \le 4\mathbb{E}[(I_m \cap I_\ell|n/k^2 - 1)^2] \le 4n^2k^{-4}k^2n^{-1} = 4n^{-1}c^{-2},$$

which implies

$$\mathbb{P}(|I_m \cap I_\ell| \le k^2/(2n) = 2^{-1}nc^2) \le 4n^{-1}c^{-2}.$$

Note in passing that the above inequality also holds for $m = \ell$ since $|I_m \cap I_\ell| = k \ge k^2/2n$ with probability 1. Therefore, we have, for all m, ℓ,

$$\begin{aligned}\mathbb{P}(d_{m,\ell}^{\text{ovlp}} \le 2^{-1}C^2nc^4\tau^2\gamma^{-2}) &\le \mathbb{P}(|I_m \cap I_\ell| \le 2^{-1}nc^2) + \mathbb{P}((1 - \widehat{\mathsf{df}}_m/k)(1 - \widehat{\mathsf{df}}_\ell/k) \le C^2\tau^2c^2\gamma^{-2}) \\ &\le 4n^{-1}c^{-2} + 2e^{-nc/2} \lesssim n^{-1}c^{-2}.\end{aligned}$$

Proof for $\# = $ full. Recall that $d_{m,\ell}^{\text{full}} = n - \widehat{\mathsf{df}}_m - \widehat{\mathsf{df}}_\ell + k^{-2}|I_m \cap I_\ell|\widehat{\mathsf{df}}_m\widehat{\mathsf{df}}_\ell$. We consider $m = \ell$ and $m \ne \ell$ separately.

When $m = \ell$, (56) with $m = \ell$ implies that the following holds with probability at least $1 - 2e^{-nc/2}$:

$$n^{-1}d_{m,m}^{\text{full}} = c(1 - \widehat{\mathsf{df}}_m/k)^2 + 1 - c \ge cC^2\tau^2c^2\gamma^{-2} + 1 - c \ge C^2\tau^2\gamma^{-2}(c^3 + 1 - c) \ge C'\tau^2\gamma^{-2},$$

where $C' = C^2 \min_{c\in[0,1]}(c^3 + 1 - c) = C^2(1 - 2\sqrt{3}/9)$ is a positive absolute constant.

When $m \ne \ell$, we decompose

$$\frac{d_{m,\ell}^{\text{full}}}{n} = \Big(1 - \frac{\widehat{\mathsf{df}}_m}{n}\Big)\Big(1 - \frac{\widehat{\mathsf{df}}_\ell}{n}\Big) + \frac{\widehat{\mathsf{df}}_m}{n}\frac{\widehat{\mathsf{df}}_\ell}{n}\Big(\frac{|I_m \cap I_\ell| \cdot n}{k^2} - 1\Big) =: A_{m,\ell} + B_{m,\ell}.$$

Lemma 18 implies that there exists an absolute constant $C \in (0, 1)$ such that

$$\mathbb{P}(A_{m,\ell} \le C^2\tau^2\gamma^{-2}) \le 2e^{-nc/2}.$$

For $B_{m,\ell}$, $0 < \widehat{\mathsf{df}}_m < k = nc$ and Lemma 22 imply

$$\mathbb{E}[|B_{m,\ell}|^2] \le (\frac{k^2}{n^2}\frac{n}{k^2})^2\mathbb{E}[||I_m \cap I_\ell| - \frac{k^2}{n}|^2] \le \frac{1}{n^2}\frac{k^2}{n} = \frac{c^2}{n}.$$

Thus, Markov's inequality leads to

$$\begin{aligned}\mathbb{P}(d_{m,\ell}^{\text{full}} > n2C^2\tau^2\gamma^{-2}) &\le \mathbb{P}(A_{m,\ell} > C^2\tau^2\gamma^{-2}) + \mathbb{P}(B_{m,\ell} > C^2\tau^2\gamma^{-2}) \\ &\le 2e^{-nc/2} + (C^2\tau^2\gamma^{-2})^{-2}\mathbb{E}[|B_{m,\ell}|^2] \lesssim (nc)^{-1} + \tau^{-4}\gamma^4c^2n^{-1} \\ &\lesssim n^{-1}c^{-1}\tau^{-4}\gamma^4.\end{aligned}$$

This concludes the proof. □

Lemma 18. There exists an absolute constant $C \in (0,1)$ such that

$$\mathbb{P}(1 - \widehat{\mathsf{df}}/k \geq C\tau c\gamma^{-1}) \geq 1 - e^{-nc/2}, \quad \mathbb{P}(1 - \widehat{\mathsf{df}}/n \geq C\tau\gamma^{-1}) \geq 1 - e^{-nc/2}$$

Proof. Equation (30) implies

$$(1 - \widehat{\mathsf{df}}/k)^{-1} \leq 1 + \mu^{-1}\|\boldsymbol{L}\boldsymbol{G}\|_{\mathrm{op}}^2/k \leq \tau^{-1}(1 + \|\boldsymbol{L}\boldsymbol{G}\|_{\mathrm{op}}^2/k),$$

thanks to $\tau = \min(1,\mu)$, where $\boldsymbol{G} \in \mathbb{R}^{n\times p}$ has i.i.d. Gaussian entries. Since $\|\boldsymbol{L}\boldsymbol{G}\|_{\mathrm{op}} \overset{\mathrm{d}}{=} \|\boldsymbol{G}'\|_{\mathrm{op}}$ with $\boldsymbol{G}' \in \mathbb{R}^{k\times p}$ having i.i.d. $\mathcal{N}(0,1)$ entries, Lemma 20 with $t = \sqrt{k} + \sqrt{p}$ implies

$$\mathbb{P}(\|\boldsymbol{L}\boldsymbol{Z}\|_{\mathrm{op}}^2 > \{2(\sqrt{k} + \sqrt{p})\}^2) \leq e^{-(\sqrt{k}+\sqrt{p})^2/2} \leq e^{-(k+p)/2} \leq e^{-k/2}.$$

Since $\{2(\sqrt{k} + \sqrt{p})\}^2 \leq 8(k+p)$. the following holds with probability at least $1 - e^{-k/2}$:

$$(1 - \widehat{\mathsf{df}}/k)^{-1} \leq \tau^{-1}(1 + 8(1 + p/k)) \leq 9\tau^{-1}(1 + p/k) \leq 9\tau^{-1}(1 + c^{-1}p/n) \leq 18\tau^{-1}c^{-1}\gamma,$$

thanks to $\gamma = \max(1, p/n)$. This completes the proof for $1 - \widehat{\mathsf{df}}/k$. For $1 - \widehat{\mathsf{df}}/n$, $0 < \widehat{\mathsf{df}}_m < k = nc$ and the above display lead to

$$\mathbb{P}(1 - \widehat{\mathsf{df}}_m/n \geq (1-c) \vee \{(18)^{-1}\tau c\gamma^{-1}\}) \geq 1 - e^{-nc/2},$$

where $(1-c) \vee \{(18)^{-1}\tau c\gamma^{-1}\} \geq (18)^{-1}\tau\gamma^{-1}((1-c) \vee c) \geq (18)^{-1}\tau\gamma^{-1}2^{-1}$ thanks to $c, \tau \in (0,1]$ and $\gamma > 1$. Therefore taking $C = (36)^{-1}$ concludes the proof. □

A.5.6. Proof of Lemma 13 (restated here for convenience)

Lemma 13. Suppose the same penalty is used for $(\boldsymbol{h}_m)_{m=1}^M$. Then, we have

$$\text{for all } m \in [M], \quad \text{for all } \epsilon > 0, \quad \mathbb{P}\Big(\frac{|\widehat{\mathsf{df}}_m - \widetilde{\mathsf{df}}_M|}{n} > \epsilon\Big) \lesssim M\Big(e^{-nc/2} + \frac{\gamma^{9/2}}{\epsilon\sqrt{n}\tau^3 c^4}\Big).$$

Proof. Now we suppose that exists a deterministic scalar d_n that does not depend on m such that

$$\mathbb{P}(|\widehat{\mathsf{df}}_m/n - d_n| > \epsilon) \lesssim e^{-nc/2} + \frac{\gamma^{9/2}}{\epsilon\sqrt{n}\tau^3 c^4}. \tag{57}$$

Then, multiple applications of the triangle inequality lead to

$$\begin{aligned}\mathbb{P}(|\widehat{\mathsf{df}}_m - \widetilde{\mathsf{df}}_M| > n\epsilon) &\leq \mathbb{P}\Big(\Big|\frac{\widehat{\mathsf{df}}_m}{n} - d_n\Big| > \frac{\epsilon}{2}\Big) + \mathbb{P}\Big(\Big|\frac{M^{-1}\sum_m \widehat{\mathsf{df}}_m}{n} - d_n\Big| > \frac{\epsilon}{2}\Big) \\ &\leq \mathbb{P}\Big(\Big|\frac{\widehat{\mathsf{df}}_m}{n} - d_n\Big| > \frac{\epsilon}{2}\Big) + \sum_{m=1}^M \mathbb{P}\Big(\Big|\frac{\widehat{\mathsf{df}}_m}{n} - d_n\Big| > \frac{\epsilon}{2}\Big) \lesssim M\Big(e^{-nc/2} + \frac{\gamma^{9/2}}{\epsilon\sqrt{n}\tau^3 c^4}\Big),\end{aligned}$$

which completes the proof. Below we prove (57). Let $\boldsymbol{L}_m = \boldsymbol{L}_{I_m}$ for brevity. Define (W_m, d_n) by

$$W_m := \frac{k\|\boldsymbol{L}_m \boldsymbol{r}_m\|^2}{n^2\|\boldsymbol{h}_m\|^2} = \frac{c\|\boldsymbol{L}_m \boldsymbol{r}_m\|^2}{n\|\boldsymbol{h}_m\|^2}, \quad d_n = \frac{k}{n} - \sqrt{\mathbb{E}[W_m]}.$$

Note that d_n does not depend on m by symmetry. Since $\widehat{\mathsf{df}}_m = k - \operatorname{tr}[\boldsymbol{L}_m \boldsymbol{V}_m]$, our goal is to bound

$$\widehat{\mathsf{df}}_m/n - d_n = \operatorname{tr}[\boldsymbol{L}_m \boldsymbol{V}_m]/n - \sqrt{\mathbb{E}[W_m]}.$$

Equation (31) with $I = \widetilde{I} = I_m$ implies

$$\mathbb{E}\Big[\Big|W_m - \frac{\operatorname{tr}[\boldsymbol{L}_m \boldsymbol{V}_m]^2}{n^2}\Big|\Big] = \frac{k}{n^2}\mathbb{E}\Big[\Big|\frac{\|\boldsymbol{L}_m \boldsymbol{r}_m\|^2}{\|\boldsymbol{h}_m\|^2} - k\operatorname{tr}[\boldsymbol{L}_m \boldsymbol{V}_m]^2\Big|\Big] \lesssim \frac{k}{n^2}\cdot\sqrt{n}\tau^{-2}c^{-3}\gamma^{7/2} = \frac{\gamma^{7/2}}{\sqrt{n}\tau^2 c^2},$$

while the Gaussian Poincaré inequality leads to (see Lemma 19 below)

$$\operatorname{Var}[W_m] = \operatorname{Var}\left[\frac{c}{n}\frac{\|\boldsymbol{L}_m \boldsymbol{r}_m\|^2}{\|\boldsymbol{h}_m\|^2}\right] = \frac{c^2}{n^2}\operatorname{Var}\left[\frac{\|\boldsymbol{L}_m \boldsymbol{r}_m\|^2}{\|\boldsymbol{h}_m\|^2}\right] \lesssim \frac{c^2}{n^2}\cdot\frac{n\gamma^3}{c^2\tau^2} = \frac{\gamma^3}{n\tau^2}.$$

By the above displays, we obtain

$$\begin{aligned}
\mathbb{E}\Big[\Big|\frac{\operatorname{tr}[\boldsymbol{L}_m \boldsymbol{V}_m]^2}{n^2} - \mathbb{E}[W_m]\Big|\Big] &\le \mathbb{E}\Big[\Big|\frac{\operatorname{tr}[\boldsymbol{L}_m \boldsymbol{V}_m]^2}{n^2} - W_m\Big|\Big] + \mathbb{E}\Big[\big|W_m - \mathbb{E}[W_m]\big|\Big] \\
&\le \mathbb{E}\Big[\Big|\frac{\operatorname{tr}[\boldsymbol{L}_m \boldsymbol{V}_m]^2}{n^2} - W_m\Big|\Big] + \sqrt{\operatorname{Var}[W_m]} \\
&\lesssim \frac{\gamma^{7/2}}{\sqrt{n}\tau^2 c^2} + \sqrt{\frac{\gamma^3}{n\tau^2}} \lesssim \frac{\gamma^{7/2}}{\sqrt{n}\tau^2 c^2}.
\end{aligned}$$

Now, Lemma 18 implies that there exists an absolute constant $C > 0$ such that

$$\mathbb{P}(\Omega^c) \le e^{-(k+p)/2} \le e^{-nc/2} \text{ with } \Omega = \{\operatorname{tr}[\boldsymbol{L}_m \boldsymbol{V}_m]/n \ge C\gamma^{-1}c\tau\}$$

Notice that under Ω,

$$\begin{aligned}
\Big|\frac{\operatorname{tr}[\boldsymbol{L}_m \boldsymbol{V}_m]^2}{n^2} - \mathbb{E}[W_m]\Big| &= \Big|\frac{\operatorname{tr}[\boldsymbol{L}_m \boldsymbol{V}_m]}{n} - \sqrt{\mathbb{E}[W_m]}\Big| \cdot \Big|\frac{\operatorname{tr}[\boldsymbol{L}_m \boldsymbol{V}_m]}{n} + \sqrt{\mathbb{E}[W_m]}\Big| \\
&= \Big|\frac{\widehat{\mathsf{df}}_m}{n} - d_n\Big| \cdot \Big|\frac{\operatorname{tr}[\boldsymbol{L}_m \boldsymbol{V}_m]}{n} + \sqrt{\mathbb{E}[W_m]}\Big| \ge \Big|\frac{\widehat{\mathsf{df}}_m}{n} - d_n\Big| C\gamma^{-1}c^2\tau.
\end{aligned}$$

Combining the above displays together, we obtain

$$\begin{aligned}
\mathbb{P}\Big(\Big|\frac{\widehat{\mathsf{df}}_m}{n} - d_n\Big| > \epsilon\Big) &\le \mathbb{P}(\Omega^c) + \mathbb{P}\Big(\Big|\frac{\operatorname{tr}[\boldsymbol{L}_m \boldsymbol{V}_m]^2}{n^2} - \mathbb{E}[W_m]\Big| > \epsilon C\gamma^{-1}c^2\tau\Big) \\
&\le e^{-nc/2} + \frac{\gamma}{\epsilon C c^2\tau}\mathbb{E}\Big[\Big|\frac{\operatorname{tr}[\boldsymbol{L}_m \boldsymbol{V}_m]^2}{n^2} - \mathbb{E}[W_m]\Big|\Big] \\
&\lesssim e^{-nc/2} + \frac{\gamma}{\epsilon c^2\tau}\cdot\frac{\gamma^{7/2}}{\sqrt{n}\tau^2 c^2} = e^{-nc/2} + \frac{\gamma^{9/2}}{\epsilon\sqrt{n}\tau^3 c^4}.
\end{aligned}$$

This finishes the proof. $\square$

Lemma 19. We have $\operatorname{Var}(\|\boldsymbol{L}\boldsymbol{r}\|^2/\|\boldsymbol{h}\|) \lesssim nc^{-2}\tau^{-2}\gamma^3$.

Proof. Let $\boldsymbol{u} = \boldsymbol{L}\boldsymbol{r}/\|\boldsymbol{h}\| = -\boldsymbol{L}\boldsymbol{Z}\boldsymbol{h}/\|\boldsymbol{h}\| = -\boldsymbol{L}\boldsymbol{Z}\boldsymbol{v}$ with $\boldsymbol{v} = \boldsymbol{h}/\|\boldsymbol{h}\|$. The Gaussian Poincaré inequality yields

$$\mathrm{Var}(\|\boldsymbol{u}\|^2) \le \mathbb{E}\Big[\sum_{ij}\Big(\frac{\partial\|\boldsymbol{u}\|^2}{\partial z_{ij}}\Big)^2\Big] = 4\sum_{ij}\mathbb{E}\Big(\boldsymbol{u}^\top\frac{\partial \boldsymbol{u}}{\partial z_{ij}}\Big)^2,$$

where

$$\boldsymbol{u}^\top\frac{\partial \boldsymbol{u}}{\partial z_{ij}} = \boldsymbol{v}^\top\boldsymbol{Z}^\top\boldsymbol{L}(\boldsymbol{e}_i v_j + \boldsymbol{Z}(\frac{\partial}{\partial z_{ij}}\frac{\boldsymbol{h}}{\|\boldsymbol{h}\|})) = \boldsymbol{v}^\top\boldsymbol{Z}^\top\boldsymbol{L}\boldsymbol{e}_i v_j + \boldsymbol{v}^\top\boldsymbol{Z}^\top\boldsymbol{L}\boldsymbol{Z}\boldsymbol{P}^\perp\frac{1}{\|\boldsymbol{h}\|}\frac{\partial \boldsymbol{h}}{\partial z_{ij}} = \mathsf{Rem}^1_{ij} + \mathsf{Rem}^2_{i,j}.$$

Note that $\sum_{ij}(\mathsf{Rem}^1_{ij})^2 \le \|\boldsymbol{L}\boldsymbol{Z}\|_{\mathrm{op}}^2$. For Rem^2_{ij}, the derivative formula (29) leads to

$$\begin{aligned}
\sum_{ij}(\mathsf{Rem}^2_{ij})^2 &= \frac{1}{\|\boldsymbol{h}\|^2}\sum_{ij}(\boldsymbol{v}^\top\boldsymbol{Z}^\top\boldsymbol{L}\boldsymbol{Z}\boldsymbol{P}^\perp(\boldsymbol{B}\boldsymbol{e}_j(\boldsymbol{L}\boldsymbol{r})_i - \boldsymbol{B}\boldsymbol{Z}^\top\boldsymbol{L}\boldsymbol{e}_i h_j))^2\\
&\lesssim \frac{1}{\|\boldsymbol{h}\|^2}(\|\boldsymbol{B}^\top\boldsymbol{P}^\perp\boldsymbol{Z}^\top\boldsymbol{L}\boldsymbol{Z}\boldsymbol{v}\|^2\|\boldsymbol{L}\boldsymbol{r}\|^2 + \|\boldsymbol{L}\boldsymbol{Z}\boldsymbol{B}^\top\boldsymbol{P}^\perp\boldsymbol{Z}^\top\boldsymbol{L}\boldsymbol{Z}\boldsymbol{v}\|^2\|\boldsymbol{h}\|^2)\\
&\lesssim \|\boldsymbol{B}\|_{\mathrm{op}}^2\|\boldsymbol{L}\boldsymbol{Z}\|_{\mathrm{op}}^6 \lesssim (k\mu)^{-2}\|\boldsymbol{L}\boldsymbol{Z}\|_{\mathrm{op}}^6.
\end{aligned}$$

Thus, using Lemma 20, we obtain $\mathrm{Var}(\|\boldsymbol{u}\|^2) \lesssim (k+p) + k^{-2}\mu^{-2}(k^3+p^3) \lesssim nc^{-2}\tau^{-2}\gamma^3$, thanks to $\tau = \min(1,\mu), c = k/n \in (0,1]$ and $\gamma = \max(1, p/n)$. □

A.6. Miscellaneous useful facts

Lemma 20. If $\boldsymbol{G} \in \mathbb{R}^{k\times q}$ has i.i.d. $\mathcal{N}0,1)$ entries, the tail $\mathbb{P}(\|\boldsymbol{G}\|_{\mathrm{op}} > \sqrt{k}+\sqrt{q}+t) \le \Phi(-t) \le e^{-t^2/2}$ holds for all $t \in \mathbb{R}$, where $\Phi(\cdot)$ is the CDF of the standard normal. Therefore, we have $\mathbb{E}[\|\boldsymbol{G}\|_{\mathrm{op}}^m] \le C(m)(\sqrt{k}+\sqrt{q})^r$ for all $m \ge 1$, where $C(m) > 0$ is a constant depending only on m.

Proof. See Theorem 2.13 of Davidson and Szarek (2001) for the tail bound. The moment bound is obtained by integrating the tail bound. □

Lemma 21 (Simple random sampling properties; see, e.g., page 13 of Chaudhuri (2014))**.** Fix an array $(x_i)_{i=1}^M$ of length $M \ge 1$ and let μ_M be the mean $M^{-1}\sum_{i\in[M]} x_i$ and σ_M^2 be the variance $M^{-1}\sum_{i\in M} x_i^2 - \mu_M^2$. Suppose J is uniformly distributed on $\{J \subset [M] : |J| = m\}$ for a fixed integer $m \le M$. Then, the mean and variance of the sample mean $\widehat{\mu}_J = \sum_{i\in J} x_i$ are given by

$$\mathbb{E}[\widehat{\mu}_J] = \mu_M, \quad \text{and} \quad \mathrm{Var}[\widehat{\mu}_J] = \frac{\sigma_M^2}{m}\left(1 - \frac{m-1}{M-1}\right) \le \frac{\sum_{i\in M} x_i^2}{mM}.$$

Lemma 22. If I and $\widetilde{I}$ are independent and uniformly distributed over $[n]$ with cardinality $k \le n$, we have $\mathbb{E}[(|I\cap\widetilde{I}| - n^{-1}k^2)^2] \le n^{-1}k^2$.

Proof. A random variable X is said to follow a hypergeometric distribution, denoted as $X \sim \mathrm{Hypergeometric}(n, K, N)$, if its probability mass function can be expressed as:

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

The mean and variance of X are respectively given by (e.g., Greene and Wellner (2017)):

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

Since $|I\cap\widetilde{I}| \sim \mathrm{Hypergeometric}(k,k,n)$, we have $\mathbb{E}[(|I\cap\widetilde{I}| - n^{-1}k^2)^2] = \mathrm{Var}(|I\cap\widetilde{I}|) \le k^2/n$. □

A.7. Relaxing Assumption D in the underparameterized regime

In the underparameterized regime when $p/n < 1$, the strongly convex assumption on the penalty function in Assumption D is not needed; we only need the penalty to be convex. We make this precise in this section.

Theorem 23 *Let Assumptions A–C be fulfilled and assume $p/k \le \gamma < 1$ for a constant γ independent of k, n, p. For any convex function g_m, define*

$$\check{\boldsymbol{b}}_m = \operatorname*{argmin}_{\boldsymbol{b}\in\mathbb{R}^p} \frac{1}{2|I_m|}\|\boldsymbol{L}_{I_m}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b})\|_2^2 + g_m(\boldsymbol{b}) + \mu_n\|\boldsymbol{\Sigma}^{1/2}(\boldsymbol{b} - \boldsymbol{\beta}_0)\|^2$$

for some deterministic $\mu_n \ge 0$; which is the same as $\widehat{\boldsymbol{\beta}}_m$ in (2) with an additional quadratic penalty term $\mu_n\|\boldsymbol{\Sigma}^{1/2}(\boldsymbol{b} - \boldsymbol{\beta}_0)\|^2$. Let also $\check{\mathsf{df}}_m$ be the degrees of freedom of $\check{\boldsymbol{b}}_m$. Then as long as $\mu_n \to 0$ as $n, k, p \to +\infty$ we have

$$\frac{\sigma^2 + \|\boldsymbol{\Sigma}^{1/2}(\check{\boldsymbol{b}}_m - \boldsymbol{\beta}_0)\|^2}{\sigma^2 + \|\boldsymbol{\Sigma}^{1/2}(\widehat{\boldsymbol{\beta}}_m - \widehat{\boldsymbol{\beta}}_0)\|^2} \xrightarrow{\mathrm{p}} 1, \ \frac{1 - \check{\mathsf{df}}_m/k}{1 - \widehat{\mathsf{df}}_m/k} \xrightarrow{\mathrm{p}} 1, \ \frac{\|\boldsymbol{y} - \boldsymbol{X}\check{\boldsymbol{b}}_m\|}{\|\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}_m\|} \xrightarrow{\mathrm{p}} 1, \ \frac{\|\boldsymbol{L}_{I_m}(\boldsymbol{y} - \boldsymbol{X}\check{\boldsymbol{b}}_m)\|}{\|\boldsymbol{L}_{I_m}(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}_m)\|} \xrightarrow{\mathrm{p}} 1$$

as well as

$$\frac{\|\boldsymbol{\Sigma}^{1/2}(\widehat{\boldsymbol{\beta}}_m - \check{\boldsymbol{b}}_m)\|}{\min\{\|\boldsymbol{\Sigma}^{1/2}(\widehat{\boldsymbol{\beta}}_m - \boldsymbol{\beta}_0)\|, \|\boldsymbol{\Sigma}^{1/2}(\check{\boldsymbol{b}}_m - \boldsymbol{\beta}_0)\|\}} \xrightarrow{\mathrm{p}} 0.$$

If g_m is convex and the same for all $m \in [M]$, taking $\tau = \mu_n \to 0$ sufficiently slowly so that the right-hand sides of (22) converge to 0 (for instance, $\mu_n = 1/\log n$ works), then

$$\frac{\widetilde{R}_M^{\mathrm{cgcv,ovlp}}}{R_M} \xrightarrow{\mathrm{p}} 1, \qquad \frac{\widetilde{R}_M^{\mathrm{cgcv,full}}}{R_M} \xrightarrow{\mathrm{p}} 1,$$

i.e., the proposed estimates are consistent without requiring strong convexity assumption on g_m.

Proof. By the optimality condition of $\check{\boldsymbol{b}}_m$, its objective value is smaller than that of $\widehat{\boldsymbol{\beta}}_m$, so that leaving aside $\mu_n\|\boldsymbol{\Sigma}^{1/2}(\check{\boldsymbol{b}}_m - \boldsymbol{\beta}_0)\|^2$ which is non-negative,

$$\frac{\|\boldsymbol{L}_{I_m}(\boldsymbol{y} - \boldsymbol{X}\check{\boldsymbol{b}}_n)\|_2^2}{2|I_m|} + g_m(\check{\boldsymbol{b}}_m) \le \frac{\|\boldsymbol{L}_{I_m}(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}_m)\|_2^2}{2|I_m|} + g_m(\widehat{\boldsymbol{\beta}}_m) + \mu_n\|\boldsymbol{\Sigma}^{1/2}(\widehat{\boldsymbol{\beta}}_m - \boldsymbol{\beta}_0)\|^2.$$

Similarly by optimality of $\widehat{\boldsymbol{\beta}}_m$, noting that $\widehat{\boldsymbol{\beta}}_m$ still minimizes the convex function

$$\boldsymbol{\beta} \mapsto \frac{\|\boldsymbol{L}_{I_m}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})\|_2^2}{2|I_m|} - \frac{\|\boldsymbol{L}_{I_m}\boldsymbol{X}(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}_m)\|^2}{2|I_m|} + g_m(\boldsymbol{\beta}),$$

thanks to the KKT (Karush-Kuhn-Tucker) condition for the original minimization problem, we have

$$\frac{\|\boldsymbol{L}_{I_m}\boldsymbol{X}(\check{\boldsymbol{b}}_m - \widehat{\boldsymbol{\beta}}_m)\|_2^2}{2|I_m|} + \frac{\|\boldsymbol{L}_{I_m}(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}_m)\|_2^2}{2|I_m|} + g_m(\widehat{\boldsymbol{\beta}}_m) \le \frac{\|\boldsymbol{L}_{I_m}(\boldsymbol{y} - \boldsymbol{X}\check{\boldsymbol{b}}_n)\|_2^2}{2|I_m|} + g_m(\check{\boldsymbol{b}}_m).$$

Summing the above two displays, most terms cancel, and we find

$$\frac{\|\boldsymbol{L}_{I_m}\boldsymbol{X}(\check{\boldsymbol{b}}_m - \widehat{\boldsymbol{\beta}}_m)\|_2^2}{2|I_m|} \le \mu_n\|\boldsymbol{\Sigma}^{1/2}(\widehat{\boldsymbol{\beta}}_m - \boldsymbol{\beta}_0)\|^2.$$

Since $p/k \le \gamma < 1$, the left-hand side is bounded from below by $\frac{1}{2}C(\gamma)\|\boldsymbol{\Sigma}^{1/2}(\check{\boldsymbol{b}}_m - \widehat{\boldsymbol{\beta}}_m)\|^2$ for $C(\gamma) = ((1 - \sqrt{\gamma})/2)^2$ with exponentially large probability. In this event, by the triangle inequality,

$$\begin{aligned}
&|((\|\boldsymbol{\Sigma}^{1/2}(\check{\boldsymbol{b}}_m - \boldsymbol{\beta}_0)\|^2 + \sigma^2)^{1/2} - ((\|\boldsymbol{\Sigma}^{1/2}(\boldsymbol{\beta}_0 - \widehat{\boldsymbol{\beta}}_m)\|^2 + \sigma^2)^{1/2}| \\
&\le \big|\|\boldsymbol{\Sigma}^{1/2}(\check{\boldsymbol{b}}_m - \boldsymbol{\beta}_0)\| - \|\|\boldsymbol{\Sigma}^{1/2}(\boldsymbol{\beta}_0 - \widehat{\boldsymbol{\beta}}_m)\|\|\big| \\
&\le \|\boldsymbol{\Sigma}^{1/2}(\check{\boldsymbol{b}}_m - \widehat{\boldsymbol{\beta}}_m)\| \\
&\le \sqrt{2\mu_n/C(\gamma)}\|\boldsymbol{\Sigma}^{1/2}(\widehat{\boldsymbol{\beta}}_m - \boldsymbol{\beta}_0)\|.
\end{aligned} \tag{58}$$

Dividing by $(\|\boldsymbol{\Sigma}^{1/2}(\widehat{\boldsymbol{\beta}}_m - \boldsymbol{\beta}_0)\|^2 + \sigma^2)^{1/2}$, we obtain the desired convergence

$$\frac{\sigma^2 + \|\boldsymbol{\Sigma}^{1/2}(\check{\boldsymbol{b}}_m - \boldsymbol{\beta}_0)\|^2}{\sigma^2 + \|\boldsymbol{\Sigma}^{1/2}(\widehat{\boldsymbol{\beta}}_m - \widehat{\boldsymbol{\beta}}_0)\|^2} \xrightarrow{\text{p}} 1,$$

thanks to $\mu_n \to 0$. For the convergence of the ratio of norms of residuals, we may bound from below the denominators by

$$\|\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}_m\| \geq \|\boldsymbol{L}_{I_m}(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}_m)\| \geq \sqrt{|I_m|C(\gamma)}(\|\boldsymbol{\Sigma}^{1/2}(\widehat{\boldsymbol{\beta}}_m - \boldsymbol{\beta}_0)\|^2 + \sigma^2)^{1/2}$$

with probability approaching one, since the eigenvalues of $|I_m|^{-1/2}\boldsymbol{L}_{I_m}[\boldsymbol{X} \mid \boldsymbol{\epsilon}/\sigma]$ are bounded away from 0. In the numerators, $\|\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}_m\| - \|\boldsymbol{y} - \boldsymbol{X}\check{\boldsymbol{b}}_m\|$ is bounded from above by $C'(\gamma)\|\boldsymbol{\Sigma}^{1/2}(\check{\boldsymbol{b}}_m - \widehat{\boldsymbol{\beta}}_m)\|$ and the ratio $\|\boldsymbol{\Sigma}^{1/2}(\check{\boldsymbol{b}}_m - \widehat{\boldsymbol{\beta}}_m)\|/\|\boldsymbol{\Sigma}^{1/2}(\widehat{\boldsymbol{\beta}}_m - \boldsymbol{\beta}_0)\|$ converges to 0 by (58). This proves that the ratios of norms of residuals converge to 1 in probability. The ratio $\frac{1-\widehat{\mathsf{df}}_m/k}{1-\check{\mathsf{df}}_m/k}$ converges to 1 in probability because GCV is consistent (Bellec, 2023, Section 3) for both $\check{\boldsymbol{b}}_m$ and $\widehat{\boldsymbol{\beta}}_m$, which gives

$$(1 - \widehat{\mathsf{df}}_m/k) \cdot \frac{\|\boldsymbol{L}_{I_m}(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}_m)\|}{\sqrt{|I_m|(\sigma^2 + \|\boldsymbol{\Sigma}^{1/2}(\widehat{\boldsymbol{\beta}}_m - \boldsymbol{\beta}_0)\|^2)}} \xrightarrow{\text{p}} 1,$$

and similarly for $\check{\boldsymbol{b}}_m$. Finally, (58) provides

$$\|\boldsymbol{\Sigma}^{1/2}(\widehat{\boldsymbol{\beta}}_m - \check{\boldsymbol{b}}_m)\|/\min\{\|\boldsymbol{\Sigma}^{1/2}(\widehat{\boldsymbol{\beta}}_m - \boldsymbol{\beta}_0)\|, \|\boldsymbol{\Sigma}^{1/2}(\check{\boldsymbol{b}}_m - \boldsymbol{\beta}_0)\|\} \xrightarrow{\text{p}} 0.$$

By (41), we further have, for $\check{\boldsymbol{b}}_1, \dots \check{\boldsymbol{b}}_M$ thanks to the strongly convex penalty, that

$$\mathbb{P}\Big(\sum_m (\sigma^2 + \|\boldsymbol{\Sigma}^{1/2}(\check{\boldsymbol{b}}_m - \boldsymbol{\beta}_0)\|^2) \leq 2(\sigma^2 + \|\sum_m \boldsymbol{\Sigma}^{1/2}(\check{\boldsymbol{b}}_m - \boldsymbol{\beta}_0)\|^2)\Big) \to 1.$$

Combining these inequalities, we get that all quantities involved in the definition of (17)-(18) and their targets are close to the corresponding quantities with $(\widehat{\boldsymbol{\beta}}_m, \widehat{\mathsf{df}}_m)$ replaced by $(\check{\boldsymbol{b}}_m, \check{\mathsf{df}}_m)$. Thus the consistency of the estimates for $(\check{\boldsymbol{b}}_m, \check{\mathsf{df}}_m)$ implies the consistency of the estimates for $(\widehat{\boldsymbol{\beta}}_m, \widehat{\mathsf{df}}_m)$. $\square$

B. Proofs in Section 5

B.1. Preparatory definitions

Fixed-point parameters. 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 $v_p(-\lambda; \theta)$ as the unique nonnegative solution to the fixed-point equation

$$v_p(-\lambda; \theta)^{-1} = \lambda + \theta \int r(1 + v_p(-\lambda; \theta)r)^{-1} \, \mathrm{d}H_p(r), \tag{59}$$

where $H_p(r) = p^{-1}\sum_{i=1}^p \mathbb{1}_{\{r_i \leq r\}}$ is the empirical spectral distribution of $\boldsymbol{\Sigma}$ and r_i's are the eigenvalues of $\boldsymbol{\Sigma}$. For $\lambda = 0$, define $v_p(0; \theta) := \lim_{\lambda \to 0^+} v_p(-\lambda; \theta)$ for $\theta > 1$ and ∞ otherwise.

Best linear projection. Since the ensemble ridge estimators are linear estimators, we evaluate their performance relative to the oracle parameter:

$$\boldsymbol{\beta}_0 = \mathbb{E}[\boldsymbol{x}\boldsymbol{x}^\top]^{-1}\mathbb{E}[\boldsymbol{x}y],$$

which is the best (population) linear projection of y onto $\boldsymbol{x}$ and minimizes the linear regression error. Note that we can decompose any response y into:

$$y = f_{\text{LI}}(\boldsymbol{x}) + f_{\text{NL}}(\boldsymbol{x}),$$

where $f_{\text{LI}}(\boldsymbol{x}) = \boldsymbol{\beta}_0^\top \boldsymbol{x}$ is the oracle linear predictor, and $f_{\text{NL}}(\boldsymbol{x}) = y - f_{\text{LI}}(\boldsymbol{x})$ is the nonlinear component that is not explained by $f_{\text{LI}}(\boldsymbol{x})$. The best linear projection has the useful property that $f_{\text{LI}}(\boldsymbol{x})$ is (linearly) uncorrelated with $f_{\text{NL}}(\boldsymbol{x})$, although they are generally dependent. It is worth mentioning that this does not imply that y and $\boldsymbol{x}$ follow a linear regression model. Indeed, our framework allows any nonlinear dependence structure between them and is model-free for the joint distribution of $(\boldsymbol{x}, y)$. For n i.i.d. samples from the same distribution as $(\boldsymbol{x}, y)$, we define analogously the vector decomposition:

$$\boldsymbol{y} = \boldsymbol{f}_{\text{LI}} + \boldsymbol{f}_{\text{NL}}, \tag{60}$$

where $\boldsymbol{f}_{\text{LI}} = \boldsymbol{X}\boldsymbol{\beta}_0$ and $\boldsymbol{f}_{\text{NL}} = [f_{\text{NL}}(\boldsymbol{x}_i)]_{i\in[n]}$.

Covariance and resolvents. For $j \in [M]$, let $\boldsymbol{L}_j$ be a diagonal matrix with i-th diagonal entry being 1 if $i \in I_m$ and 0 otherwise. Let $\widehat{\boldsymbol{\Sigma}}_j = \boldsymbol{X}^\top \boldsymbol{L}_j \boldsymbol{X}/k$ and $\boldsymbol{M}_j = (\widehat{\boldsymbol{\Sigma}}_j + \lambda \boldsymbol{I}_p)^{-1}$.

Risk and risk estimator. Recall that for $m, \ell \in [M]$, the risk component is defined as

$$R_{m,\ell} = \|f_{\text{NL}}\|_{L_2}^2 + (\widehat{\boldsymbol{\beta}}_m - \boldsymbol{\beta}_0)^\top \boldsymbol{\Sigma} (\widehat{\boldsymbol{\beta}}_\ell - \boldsymbol{\beta}_0), \tag{61}$$

while its two estimators are defined as

$$\widehat{R}_{m,\ell}^{\text{ovlp}} = \frac{N_{m,\ell}^{\text{ovlp}}}{D_{m,\ell}^{\text{ovlp}}}, \quad \text{and} \quad \widehat{R}_{m,\ell}^{\text{full}} = \frac{N_{m,\ell}^{\text{full}}}{D_{m,\ell}^{\text{full}}}, \tag{62}$$

where

$$\begin{aligned}
N_{m,\ell}^{\text{ovlp}} &= |I_m \cap I_\ell|^{-1} (\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}_m)^\top \boldsymbol{L}_{m\cap\ell} (\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}_\ell \\
D_{m,\ell}^{\text{ovlp}} &= 1 - \frac{\operatorname{tr}(\boldsymbol{S}_m)}{|I_m|} - \frac{\operatorname{tr}(\boldsymbol{S}_\ell)}{|I_\ell|} + \frac{\operatorname{tr}(\boldsymbol{S}_m)\operatorname{tr}(\boldsymbol{S}_\ell)}{|I_m||I_\ell|} \\
N_{m,\ell}^{\text{full}} &= n^{-1} (\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}_m)^\top (\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}_\ell) \\
D_{m,\ell}^{\text{full}} &= 1 - \frac{\operatorname{tr}(\boldsymbol{S}_m)}{n} - \frac{\operatorname{tr}(\boldsymbol{S}_\ell)}{n} + \frac{\operatorname{tr}(\boldsymbol{S}_m)\operatorname{tr}(\boldsymbol{S}_\ell)|I_m \cap I_\ell|}{n|I_m||I_\ell|},
\end{aligned}$$

and $\boldsymbol{L}_{m\cap\ell} = \boldsymbol{L}_{I_m \cap I_\ell}$.

Asymptotic equivalence. Let $\boldsymbol{A}_p$ and $\boldsymbol{B}_p$ be sequences of additively conformable random matrices with arbitrary dimensions (including vectors and scalars as special cases). We define $\boldsymbol{A}_p$ and $\boldsymbol{B}_p$ to be *asymptotically equivalent*, denoted as $\boldsymbol{A}_p \simeq \boldsymbol{B}_p$, 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 $\boldsymbol{C}_p$ with bounded trace norm that are multiplicatively conformable to $\boldsymbol{A}_p$ and $\boldsymbol{B}_p$ and are independent of $\boldsymbol{A}_p$ and $\boldsymbol{B}_p$. Observe that when dealing with sequences of scalar random variables, this definition simplifies the standard notion of almost sure convergence for the involved sequences.

B.2. Proofs of Theorems 4 and 5 (restated here for convenience)

Theorem 4 (Pointwise consistency of intermediate estimators in λ) *Under Assumptions A and E–G with $\sqrt{n}/k = \mathcal{O}(1)$ for the ovlp-estimator, for $\lambda \geq 0$, for any $M \in \mathbb{N}$ and $m, l \in [M]$, it holds that $|\widehat{R}_{m,\ell,\lambda}^{\text{ovlp}} - R_{m,\ell,\lambda}| \xrightarrow{\text{a.s.}} 0$ and $|\widehat{R}_{m,\ell,\lambda}^{\text{full}} - R_{m,\ell,\lambda}| \xrightarrow{\text{a.s.}} 0$. Consequently, it holds that $|\widetilde{R}_{M,\lambda}^{\text{ovlp}} - R_{M,\lambda}| \xrightarrow{\text{a.s.}} 0$ and $|\widetilde{R}_{M,\lambda}^{\text{full}} - R_{M,\lambda}| \xrightarrow{\text{a.s.}} 0$.*

Theorem 5 (Uniform consistency of intermediate estimators in λ) *Under same conditions in Theorem 4, when $\psi \neq 1$, it holds that $\sup_{\lambda\in\Lambda} |\widetilde{R}_{M,\lambda}^{\text{ovlp}} - R_{M,\lambda}| \xrightarrow{\text{a.s.}} 0$ and $\sup_{\lambda\in\Lambda} |\widetilde{R}_{M,\lambda}^{\text{full}} - R_{M,\lambda}| \xrightarrow{\text{a.s.}} 0$.*

Theorem 4 is a direct consequence of Theorem 5. Thus, below, we focus on proving Theorem 5. The proof of Theorem 5 builds on the following lemmas.

1. *Risk.* Lemma 24 establishes the asymptotics $\mathscr{R}_{m,\ell}$ of the prediction risk $R_{m,\ell}$.

2. *Ovlp-estimator.* Lemma 25 and Lemma 26 establishes the asymptotics $\mathscr{N}_{m,\ell}^{\mathrm{ovlp}}$ and $\mathscr{D}_{m,\ell}^{\mathrm{ovlp}}$ of the numerator $N_{m,\ell}^{\mathrm{ovlp}}$ and the denominator $D_{m,\ell}^{\mathrm{ovlp}}$, respectively, for the ovlp-estimator.
3. *Full-estimator.* Lemma 27 and Lemma 28 establishes the asymptotics $\mathscr{N}_{m,\ell}^{\mathrm{full}}$ and $\mathscr{D}_{m,\ell}^{\mathrm{full}}$ of the numerator $N_{m,\ell}^{\mathrm{full}}$ and the denominator $D_{m,\ell}^{\mathrm{full}}$, respectively, for the full-estimator.

For $\lambda > 0$ and $k \in \mathcal{K}_n$ such that $p/k \to [\phi, \infty)$, the consistency of the two estimators follows by showing that the ratio of the numerator asymptotics and the denominator asymptotics in (1) and (2) matches with the risk asymptotics in Lemma 24, respectively:

$$R_{m,\ell} \simeq \mathscr{R}_{m,\ell} = \frac{\mathscr{N}_{m,\ell}^{\mathrm{ovlp}}}{\mathscr{D}_{m,\ell}^{\mathrm{ovlp}}} \simeq \frac{N_{m,\ell}^{\mathrm{ovlp}}}{D_{m,\ell}^{\mathrm{ovlp}}} = \widehat{R}_{m,\ell}^{\mathrm{ovlp}},$$

and

$$R_{m,\ell} \simeq \mathscr{R}_{m,\ell} = \frac{\mathscr{N}_{m,\ell}^{\mathrm{full}}}{\mathscr{D}_{m,\ell}^{\mathrm{full}}} \simeq \frac{N_{m,\ell}^{\mathrm{full}}}{D_{m,\ell}^{\mathrm{full}}} = \widehat{R}_{m,\ell}^{\mathrm{full}}.$$

Next, Lemma 29 takes care of the boundary cases when $\psi = \infty$ and $\lambda = 0$.

Finally, from Lemma 29, we also have that the sequences of functions $\{R_{m,\ell} - \mathscr{R}_{m,\ell}\}_{p\in\mathbb{N}}$, $\{\widehat{R}_{m,\ell}^{\mathrm{ovlp}} - \mathscr{R}_{m,\ell}\}_{p\in\mathbb{N}}$, and $\{\widehat{R}_{m,\ell}^{\mathrm{full}} - \mathscr{R}_{m,\ell}\}_{p\in\mathbb{N}}$ are uniformly equicontinuous on $\lambda \in \Lambda = [0, \infty]$ when $\psi \neq 1$. This implies that the sequences of functions $\{R_{m,\ell} - \widehat{R}_{m,\ell}^{\mathrm{ovlp}}\}_{p\in\mathbb{N}}$ and $\{R_{m,\ell} - \widehat{R}_{m,\ell}^{\mathrm{full}}\}_{p\in\mathbb{N}}$ are also uniformly equicontinuous on $\lambda \in \Lambda = [0, \infty]$ almost surely. From Theorem 21.8 of Davidson (1994), it further follows that the sequences converge to zero uniformly over Λ almost surely.

The rest of the current subsection is committed to presenting and proving the important lemmas Lemmas 24–29.

Lemma 24 (Asymptotic equivalence for prediction risk)**.** Under Assumptions E–G, for $\psi \in [\phi, \infty)$ and $\lambda > 0$, it holds that

$$R_{m,\ell} \simeq \mathscr{R}_{m,\ell} := (\|f_{\mathrm{NL}}\|_{L_2}^2 + \widetilde{c}_p(-\lambda; \psi))(1 + \widetilde{v}_p(-\lambda; \varphi_{m\ell}, \psi)),$$

where $\varphi_{m\ell} = \psi\,\mathbb{1}_{\{m=\ell\}} + \phi\,\mathbb{1}_{\{m\neq\ell\}}$ and $\widetilde{v}_p(-\lambda; \phi, \psi), \widetilde{c}_p(-\lambda; \psi)$ are defined in Lemma 30.

Proof. From the definition of the ridge estimator (26) and the decomposition of the response (60), we have

$$\begin{aligned}\widehat{\boldsymbol{\beta}}_m - \boldsymbol{\beta}_0 &= \boldsymbol{M}_m \frac{\boldsymbol{X}^\top \boldsymbol{L}_m}{k}(\boldsymbol{X}\boldsymbol{\beta}_0 + \boldsymbol{f}_{\mathrm{NL}}) - \boldsymbol{\beta}_0 \\ &= -\lambda \boldsymbol{M}_m \boldsymbol{\beta}_0 + \boldsymbol{M}_m \frac{\boldsymbol{X}^\top \boldsymbol{L}_m}{k} \boldsymbol{f}_{\mathrm{NL}}.\end{aligned}$$

Then it follows that

$$\begin{aligned}&\|f_{\mathrm{NL}}\|_{L_2}^2 + (\widehat{\boldsymbol{\beta}}_m - \boldsymbol{\beta}_0)^\top \boldsymbol{\Sigma} (\widehat{\boldsymbol{\beta}}_\ell - \boldsymbol{\beta}_0) \\ &= \lambda^2 \boldsymbol{\beta}_0^\top \boldsymbol{M}_m \boldsymbol{\Sigma} \boldsymbol{M}_\ell \boldsymbol{\beta}_0 \\ &\quad + (\|f_{\mathrm{NL}}\|_{L_2}^2 + \boldsymbol{f}_{\mathrm{NL}}{}^\top \frac{\boldsymbol{L}_m \boldsymbol{X}}{k} \boldsymbol{M}_m \boldsymbol{\Sigma} \boldsymbol{M}_\ell \frac{\boldsymbol{X}^\top \boldsymbol{L}_\ell}{k} \boldsymbol{f}_{\mathrm{NL}}) \\ &\quad + (\lambda \boldsymbol{\beta}_0^\top \boldsymbol{M}_\ell \boldsymbol{\Sigma} \boldsymbol{M}_\ell \frac{\boldsymbol{X}^\top \boldsymbol{L}_m}{k} \boldsymbol{f}_{\mathrm{NL}} + \lambda \boldsymbol{\beta}_0^\top \boldsymbol{M}_\ell \boldsymbol{\Sigma} \boldsymbol{M}_m \frac{\boldsymbol{X}^\top \boldsymbol{L}_m}{k} \boldsymbol{f}_{\mathrm{NL}}) \\ &= T_B + T_V + T_C.\end{aligned}$$

Bias term. From Lemma 30, we have

$$T_B \simeq \widetilde{c}_p(-\lambda; \psi)(1 + \widetilde{v}_p(-\lambda; \varphi_{m\ell}, \psi)),$$

where $\varphi_{m\ell} = \psi\,\mathbb{1}_{\{m=\ell\}} + \phi\,\mathbb{1}_{\{m\neq\ell\}}$ and $\widetilde{v}_p(-\lambda;\phi,\psi), \widetilde{c}_p(-\lambda;\psi)$ are defined in Lemma 30.
Variance term. By conditional independence and Lemma D.2 of Patil and Du (2023), the variance term converges to the the quadratic term of $\boldsymbol{f}_0 = \boldsymbol{L}_{m\cap\ell}\boldsymbol{f}_{\text{NL}}$:

$$T_V \simeq \|f_{\text{NL}}\|_{L_2}^2 + \frac{1}{k^2}\boldsymbol{f}_0^\top \boldsymbol{X}\boldsymbol{M}_m\boldsymbol{\Sigma}\boldsymbol{M}_\ell\boldsymbol{X}^\top\boldsymbol{f}_0 \simeq \|f_{\text{NL}}\|_{L_2}^2(1+\widetilde{v}_p(-\lambda;\varphi_{m\ell},\psi)),$$

where the last equivalence is from Lemma 31.
Cross term. From Patil and Du (2023, Lemma A.3), the cross term vanishes, i.e., $T_C \xrightarrow{\text{a.s.}} 0$.

This completes the proof. $\square$

Lemma 25 (Asymptotic equivalence for numerator of the ovlp-estimator)**.** Under Assumptions E–G, for $\psi\in[\phi,\infty)$ and $\lambda>0$, and index sets $I_m, I_\ell \overset{\text{SRS}}{\sim} \mathcal{I}_k$, it holds that

$$N_{m,\ell}^{\text{ovlp}} \simeq \mathscr{N}_{m,\ell}^{\text{ovlp}} := \lambda^2 v_p(-\lambda;\psi)^2(\|f_{\text{NL}}\|_{L_2}^2 + \widetilde{c}_p(-\lambda;\psi))(1+\widetilde{v}_p(-\lambda;\varphi_{m\ell},\psi)),$$

where $\varphi_{m\ell} = \psi\,\mathbb{1}_{\{m=\ell\}} + \phi\,\mathbb{1}_{\{m\neq\ell\}}$ and $v_p(-\lambda;\psi)$, $\widetilde{v}_p(-\lambda;\phi,\psi)$ and $\widetilde{c}_p(-\lambda;\psi)$ are defined in Lemma 30.

Proof. From the definition of the ridge estimator (26) and the decomposition of the response (60), we have

$$\begin{aligned}\boldsymbol{L}_{m\cap\ell}(\boldsymbol{y}-\boldsymbol{X}\widehat{\boldsymbol{\beta}}_m) &= \boldsymbol{L}_{m\cap\ell}\left(\boldsymbol{I}_p - \boldsymbol{X}\boldsymbol{M}_m\frac{\boldsymbol{X}^\top\boldsymbol{L}_m}{k}\right)\boldsymbol{y}\\ &= \lambda\boldsymbol{L}_{m\cap\ell}\boldsymbol{X}\boldsymbol{M}_m\boldsymbol{\beta}_0 + \boldsymbol{L}_{m\cap\ell}\left(\boldsymbol{I}_p - \boldsymbol{X}\boldsymbol{M}_m\frac{\boldsymbol{X}^\top\boldsymbol{L}_m}{k}\right)\boldsymbol{f}_{\text{NL}}.\end{aligned}$$

Then it follows that

$$\begin{aligned}&\frac{1}{|I_m\cap I_\ell|}(\boldsymbol{y}-\boldsymbol{X}\widehat{\boldsymbol{\beta}}_m)^\top\boldsymbol{L}_{m\cap\ell}(\boldsymbol{y}-\boldsymbol{X}\widehat{\boldsymbol{\beta}}_\ell)\\ &= \lambda^2\boldsymbol{\beta}_0^\top\boldsymbol{M}_m\widehat{\boldsymbol{\Sigma}}_{m\cap\ell}\boldsymbol{M}_\ell\boldsymbol{\beta}_0\\ &\quad+\frac{1}{|I_m\cap I_\ell|}(\boldsymbol{L}_m\boldsymbol{f}_{\text{NL}})^\top\left(\boldsymbol{I}_p - \frac{\boldsymbol{X}\boldsymbol{M}_m\boldsymbol{X}^\top}{k}\right)\boldsymbol{L}_{m\cap\ell}\left(\boldsymbol{I}_p - \frac{\boldsymbol{X}\boldsymbol{M}_\ell\boldsymbol{X}^\top}{k}\right)(\boldsymbol{L}_\ell\boldsymbol{f}_{\text{NL}})\\ &\quad+\frac{\lambda}{|I_m\cap I_\ell|}\left(\boldsymbol{\beta}_0^\top\boldsymbol{M}_\ell\boldsymbol{X}^\top\boldsymbol{L}_{m\cap\ell}\left(\boldsymbol{I}_p - \boldsymbol{X}\boldsymbol{M}_m\frac{\boldsymbol{X}^\top\boldsymbol{L}_m}{k}\right)\boldsymbol{f}_{\text{NL}} + \boldsymbol{\beta}_0^\top\boldsymbol{M}_\ell\boldsymbol{X}^\top\boldsymbol{L}_{m\cap\ell}\left(\boldsymbol{I}_p - \boldsymbol{X}\boldsymbol{M}_\ell\frac{\boldsymbol{X}^\top\boldsymbol{L}_\ell}{k}\right)\boldsymbol{f}_{\text{NL}}\right)\\ &= T_B + T_V + T_C.\end{aligned}$$

Next, we analyze the three terms separately.
Bias term. From Lemma 30,

$$T_B \simeq \lambda^2 v_p(-\lambda;\psi)^2\widetilde{c}_p(-\lambda;\psi)(1+\widetilde{v}_p(-\lambda;\varphi_{m\ell},\psi)),$$

where $\varphi_{m\ell} = \psi\,\mathbb{1}_{\{m=\ell\}} + \phi\,\mathbb{1}_{\{m\neq\ell\}}$ and $v_p(-\lambda;\psi)$, $\widetilde{v}_p(-\lambda;\phi,\psi)$ and $\widetilde{c}_p(-\lambda;\psi)$ are defined in Lemma 30.
Variance term. By conditional independence and Lemma D.2 of (Patil and Du, 2023), the variance term converges to the the quadratic term of $\boldsymbol{f}_0 = \boldsymbol{L}_{m\cap\ell}\boldsymbol{f}_{\text{NL}}$:

$$\begin{aligned}T_V &\simeq \frac{1}{|I_m\cap I_\ell|}(\boldsymbol{L}_{m\cap\ell}\boldsymbol{f}_{\text{NL}})^\top\left(\boldsymbol{I}_p - \frac{\boldsymbol{X}\boldsymbol{M}_m\boldsymbol{X}^\top}{k}\right)\boldsymbol{L}_{m\cap\ell}\left(\boldsymbol{I}_p - \frac{\boldsymbol{X}\boldsymbol{M}_\ell\boldsymbol{X}^\top}{k}\right)(\boldsymbol{L}_{m\cap\ell}\boldsymbol{f}_{\text{NL}})\\ &= \frac{1}{|I_m\cap I_\ell|}\|\boldsymbol{f}_0\|_2^2 - \frac{1}{k}\sum_{j\in\{m,\ell\}}\boldsymbol{f}_0^\top\frac{\boldsymbol{X}\boldsymbol{M}_j\boldsymbol{X}^\top}{|I_m\cap I_\ell|}\boldsymbol{f}_0 + \frac{|I_m\cap I_\ell|}{k^2}\boldsymbol{f}_0^\top\frac{\boldsymbol{X}\boldsymbol{M}_m\widehat{\boldsymbol{\Sigma}}_{m\cap\ell}\boldsymbol{M}_\ell\boldsymbol{X}^\top}{|I_m\cap I_\ell|}\boldsymbol{f}_0\\ &\simeq \lambda^2 v_p(-\lambda;\psi)^2(1+\widetilde{v}_p(-\lambda;\varphi_{m\ell},\psi)),\end{aligned}$$

where the last equivalence is from Lemma 31.
Cross term. From Patil and Du (2023, Lemma A.3), the cross term vanishes, i.e., $T_C \xrightarrow{\text{a.s.}} 0$. □

Lemma 26 (Asymptotic equivalence for denominator of the ovlp-estimator)**.** Under Assumptions F and G, for $\psi \in [\phi, \infty)$ and $\lambda > 0$, and index sets $I_m, I_\ell \overset{\mathsf{SRS}}{\sim} \mathcal{I}_k$, it holds that

$$D^{\text{ovlp}}_{m,\ell} \simeq \mathscr{D}^{\text{ovlp}}_{m,\ell} := \lambda^2 v_p(-\lambda; \psi)^2,$$

where $v_p(-\lambda; \psi)$ is the unique nonnegative solution to fixed-point equation (59).

Proof. Since $I_m, I_\ell \overset{\mathsf{SRS}}{\sim} \mathcal{I}_k$, we have that $|I_m| = |I_\ell| = k$. From Du et al. (2023, Lemma C.1), we have that

$$\frac{\operatorname{tr}(\boldsymbol{S}_j)}{|I_j|} = \frac{1}{k^2}\operatorname{tr}(\boldsymbol{X}\boldsymbol{M}_j\boldsymbol{X}^\top \boldsymbol{L}_j) = \frac{1}{k}\operatorname{tr}(\boldsymbol{M}_j\boldsymbol{\Sigma}_j) \simeq 1 - \lambda v_p(-\lambda; \psi).$$

By continuous mapping theorem,

$$1 - \frac{\operatorname{tr}(\boldsymbol{S}_m)}{|I_m|} - \frac{\operatorname{tr}(\boldsymbol{S}_\ell)}{|I_\ell|} + \frac{\operatorname{tr}(\boldsymbol{S}_m)\operatorname{tr}(\boldsymbol{S}_\ell)}{|I_m||I_\ell|} \simeq 1 - 2(1 - \lambda v_p(-\lambda; \psi)) + (1 - \lambda v_p(-\lambda; \psi)^2 = \lambda^2 v_p(-\lambda; \psi)^2,$$

which completes the proof. □

Lemma 27 (Asymptotic equivalence for the numerator of the full-estimator)**.** Under Assumptions E–G, for $\psi \in [\phi, \infty)$ and $\lambda > 0$, it holds that

$$N^{\text{full}}_{m,\ell} \simeq \mathscr{N}^{\text{full}}_{m,\ell} := d_p(-\lambda; \varphi_{m\ell}, \psi)(\|f_{\text{NL}}\|^2_{L_2} + \widetilde{c}_p(-\lambda; \psi))(1 + \widetilde{v}_p(-\lambda; \varphi_{m\ell}, \psi)),$$

where

$$d_p(-\lambda; \varphi_{m\ell}, \psi) = \begin{cases} \dfrac{\psi - \phi}{\psi} + \dfrac{\phi}{\psi}\lambda^2 v_p(-\lambda; \psi)^2, & \text{if } \varphi_{m\ell} = \psi, \\ \left(\dfrac{\psi - \phi}{\psi} + \dfrac{\phi}{\psi}\lambda v_p(-\lambda; \psi)\right)^2, & \text{if } \varphi_{m\ell} = \phi, \end{cases}$$

$\varphi_{m\ell} = \psi \mathbb{1}_{\{m=\ell\}} + \phi \mathbb{1}_{\{m\neq\ell\}}$, and $\widetilde{v}_p(-\lambda; \phi, \psi), \widetilde{c}_p(-\lambda; \psi)$ are defined in Lemma 30.

Proof. Analogous to the proof of Lemma 25, the numerator splits into

$$\begin{aligned}
&\frac{1}{n}(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}_m)^\top(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}_\ell) \\
&= \frac{1}{n}\lambda^2 \boldsymbol{\beta}_0^\top \boldsymbol{M}_m \widehat{\boldsymbol{\Sigma}} \boldsymbol{M}_\ell \boldsymbol{\beta}_0 \\
&\quad + \frac{1}{n}\boldsymbol{f}_{\text{NL}}{}^\top \left(\boldsymbol{I}_p - \frac{\boldsymbol{L}_m \boldsymbol{X}\boldsymbol{M}_m\boldsymbol{X}^\top}{k}\right)\left(\boldsymbol{I}_p - \frac{\boldsymbol{X}\boldsymbol{M}_\ell\boldsymbol{X}^\top\boldsymbol{L}_\ell}{k}\right)\boldsymbol{f}_{\text{NL}} \\
&\qquad + \frac{\lambda}{n}\left(\boldsymbol{\beta}_0^\top \boldsymbol{M}_\ell \boldsymbol{X}^\top\left(\boldsymbol{I}_p - \boldsymbol{X}\boldsymbol{M}_m\frac{\boldsymbol{X}^\top\boldsymbol{L}_m}{k}\right)\boldsymbol{f}_{\text{NL}} + \boldsymbol{\beta}_0^\top \boldsymbol{M}_\ell \boldsymbol{X}^\top\left(\boldsymbol{I}_p - \boldsymbol{X}\boldsymbol{M}_\ell\frac{\boldsymbol{X}^\top\boldsymbol{L}_\ell}{k}\right)\boldsymbol{f}_{\text{NL}}\right) \\
&= T_B + T_V + T_C.
\end{aligned}$$

Next, we analyze the three terms separately.

The bias and the cross term are analyzed as in the proof of Lemma 25, by involving Lemma 30 and Patil and Du (2023, Lemma A.3), respectively:

$$T_B \simeq d_p(-\lambda; \varphi_{m\ell}, \psi)\widetilde{c}_p(-\lambda; \psi)(1 + \widetilde{v}_p(-\lambda; \varphi_{m\ell}, \psi)),$$

$$T_C \simeq 0.$$

For the variance term, by conditional independence and Lemma D.2 of (Patil and Du, 2023), the variance term converges to the quadratic terms of $\boldsymbol{f}_1 = \boldsymbol{L}_{m\cup\ell}\boldsymbol{f}_{\text{NL}}$ and $\boldsymbol{f}_0 = \boldsymbol{L}_{m\cap\ell}\boldsymbol{f}_{\text{NL}}$, respectively:

$$\begin{aligned}
&\frac{1}{n}\boldsymbol{f}_{\text{NL}}{}^\top\left(\boldsymbol{I}_p - \frac{\boldsymbol{L}_m\boldsymbol{X}\boldsymbol{M}_m\boldsymbol{X}^\top}{k}\right)\left(\boldsymbol{I}_p - \frac{\boldsymbol{X}\boldsymbol{M}_\ell\boldsymbol{X}^\top\boldsymbol{L}_\ell}{k}\right)\boldsymbol{f}_{\text{NL}}\\
&= \frac{|I_m\cup I_\ell|}{n}\boldsymbol{f}_{\text{NL}}{}^\top\left(\boldsymbol{I}_p - \frac{\boldsymbol{L}_m\boldsymbol{X}\boldsymbol{M}_m\boldsymbol{X}^\top}{k}\right)\boldsymbol{L}_{m\cup\ell}\left(\boldsymbol{I}_p - \frac{\boldsymbol{X}\boldsymbol{M}_\ell\boldsymbol{X}^\top\boldsymbol{L}_\ell}{k}\right)\boldsymbol{f}_{\text{NL}}+\\
&\quad + \frac{|(I_m\cup I_\ell)^c|}{n}\boldsymbol{f}_{\text{NL}}{}^\top\left(\boldsymbol{I}_p - \frac{\boldsymbol{L}_m\boldsymbol{X}\boldsymbol{M}_m\boldsymbol{X}^\top}{k}\right)\boldsymbol{L}_{(m\cup\ell)^c}\left(\boldsymbol{I}_p - \frac{\boldsymbol{X}\boldsymbol{M}_\ell\boldsymbol{X}^\top\boldsymbol{L}_\ell}{k}\right)\boldsymbol{f}_{\text{NL}}\\
&= \frac{|I_m\cup I_\ell|}{n}\boldsymbol{f}_1^\top\left(\boldsymbol{I}_p - \frac{\boldsymbol{X}\boldsymbol{M}_m\boldsymbol{X}^\top}{k}\right)\boldsymbol{L}_{m\cup\ell}\left(\boldsymbol{I}_p - \frac{\boldsymbol{X}\boldsymbol{M}_\ell\boldsymbol{X}^\top}{k}\right)\boldsymbol{f}_1\\
&\quad + \frac{|(I_m\cup I_\ell)^c|}{n}\|\boldsymbol{f}_0\|_2^2 + \frac{|(I_m\cup I_\ell)^c|^2}{nk^2}\boldsymbol{f}_0^\top\boldsymbol{X}\boldsymbol{M}_m\widehat{\boldsymbol{\Sigma}}_{(m\cup\ell)^c}\boldsymbol{M}_\ell\boldsymbol{X}^\top\boldsymbol{f}_0.
\end{aligned}$$

From Lemma 31, it follows that

$$T_V \simeq d_p(-\lambda;\varphi_{m\ell},\psi)\|f_{\text{NL}}\|_{L_2}^2(1+\widetilde{v}_p(-\lambda;\varphi_{m\ell},\psi)).$$

Combining the above results completes the proof. □

Lemma 28 (Asymptotic equivalence for denominator of the full-estimator)**.** Under Assumptions F and G, for $\psi\in[\phi,\infty)$ and $\lambda>0$, and index sets $I_m, I_\ell \overset{\text{SRS}}{\sim} \mathcal{I}_k$, it holds that

$$D^{\text{full}}_{m,\ell} \simeq \mathscr{D}^{\text{full}}_{m,\ell} := \begin{cases} \dfrac{\psi-\phi}{\psi} + \dfrac{\phi}{\psi}\lambda^2 v_p(-\lambda;\psi)^2, & m=\ell,\\ \left(\dfrac{\psi-\phi}{\psi} + \dfrac{\phi}{\psi}\lambda v_p(-\lambda;\psi)\right)^2, & m\neq\ell,\end{cases}$$

where $v_p(-\lambda;\psi)$ is the unique nonnegative solution to fixed-point equation (59).

Proof. From the proof of Lemma 26, we have $\operatorname{tr}(\boldsymbol{S}_j)/|I_j| \xrightarrow{\text{a.s.}} 1-\lambda v_p(-\lambda;\psi)$, for $j\in\{m,\ell\}$. From Du et al. (2023, Lemma G.2), we also have $k/n \xrightarrow{\text{a.s.}} \phi/\psi$.

For $m=\ell$, we have

$$\begin{aligned}
1 - \frac{\operatorname{tr}(\boldsymbol{S}_m)}{n} - \frac{\operatorname{tr}(\boldsymbol{S}_\ell)}{n} + \frac{\operatorname{tr}(\boldsymbol{S}_m)\operatorname{tr}(\boldsymbol{S}_\ell)}{nk} &\simeq 1 - \frac{2\phi}{\psi}(1-\lambda v_p(-\lambda;\psi)) + \frac{\phi}{\psi}(1-\lambda v_p(-\lambda;\psi))^2\\
&= \frac{\psi-\phi}{\psi} + \frac{\phi}{\psi}(\lambda v_p(-\lambda;\psi))^2
\end{aligned}$$

For $m\neq\ell$, from Du et al. (2023, Lemma G.2), we also have $|I_m\cap I_\ell|/k \xrightarrow{\text{a.s.}} \phi/\psi$. By continuous mapping theorem, it follows that

$$\begin{aligned}
1 - \frac{\operatorname{tr}(\boldsymbol{S}_m)}{n} - \frac{\operatorname{tr}(\boldsymbol{S}_\ell)}{n} + \frac{\operatorname{tr}(\boldsymbol{S}_m)\operatorname{tr}(\boldsymbol{S}_\ell)|I_m\cap I_\ell|}{n|I_m||I_\ell|} &\simeq 1 - \frac{2\phi}{\psi}(1-\lambda v_p(-\lambda;\psi)) + \frac{\phi^2}{\psi^2}(1-\lambda v_p(-\lambda;\psi))^2\\
&= \left(\frac{\psi-\phi}{\psi} + \frac{\phi}{\psi}\lambda v_p(-\lambda;\psi)\right)^2,
\end{aligned}$$

which finishes the proof. □

Lemma 29 (Boundary cases: diverging subsample aspect ratio and ridgeless)**.** Under Assumptions E–G, the conclusions in Theorem 4 hold for $\psi = \infty$ or $\lambda = 0$, if $k = \omega(\sqrt{n})$ for the first estimator and no restriction on k is needed for the second estimator.

Furthermore, for $\mathscr{R}_{m,\ell}, \mathscr{N}^{\mathrm{ovlp}}_{m,\ell}, \mathscr{D}^{\mathrm{ovlp}}_{m,\ell}, \mathscr{N}^{\mathrm{full}}_{m,\ell}, \mathscr{D}^{\mathrm{full}}_{m,\ell}$ defined in Lemmas 24 and 26–28, the sequences of functions $\{R_{m,\ell} - \mathscr{R}_{m,\ell}\}_{p\in\mathbb{N}}$, $\{N^{\mathrm{ovlp}}_{m,\ell}/D^{\mathrm{ovlp}}_{m,\ell} - \mathscr{N}^{\mathrm{ovlp}}_{m,\ell}/\mathscr{D}^{\mathrm{ovlp}}_{m,\ell}\}_{p\in\mathbb{N}}$, and $\{N^{\mathrm{full}}_{m,\ell}/D^{\mathrm{full}}_{m,\ell} - \mathscr{N}^{\mathrm{full}}_{m,\ell}/\mathscr{D}^{\mathrm{full}}_{m,\ell}\}_{p\in\mathbb{N}}$ are uniformly bounded, equicontinuous and approaching zero on $\lambda \in [0,\infty]$ almost surely.

Proof. We split the proof into two parts.

Part (1) Diverging subsample aspect ratio for $\lambda > 0$. Recall that

$$\begin{aligned}
&\frac{1}{|I_m\cap I_\ell|}(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}_m)^\top \boldsymbol{L}_{m\cap\ell}(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}_\ell)\\
&= (\boldsymbol{\beta}_0 - \widehat{\boldsymbol{\beta}}_m)^\top \widehat{\boldsymbol{\Sigma}}_{m\cap\ell}(\boldsymbol{\beta}_0 - \widehat{\boldsymbol{\beta}}_m) + \frac{1}{|I_m\cap I_\ell|}\|\boldsymbol{L}_{m\cap\ell}\boldsymbol{f}_{\mathrm{NL}}\|_2^2 + \frac{1}{|I_m\cap I_\ell|}\sum_{j\in\{m,\ell\}} \boldsymbol{f}_{\mathrm{NL}}{}^\top \boldsymbol{L}_{m\cap\ell}\boldsymbol{X}(\boldsymbol{\beta}_0 - \widehat{\boldsymbol{\beta}}_m).
\end{aligned}$$

From law of large numbers, $\|\boldsymbol{L}_{m\cap\ell}\boldsymbol{f}_{\mathrm{NL}}\|_2^2/|I_m\cap I_\ell| \xrightarrow{\text{a.s.}} \|f_{\mathrm{NL}}\|_{L_2}^2$ as $|I_m\cap I_\ell|$ tends to infinity, which is guaranteed when $k = \omega(\sqrt{n})$. From Patil and Du (2023, Lemma A.3), $\boldsymbol{f}_{\mathrm{NL}}{}^\top \boldsymbol{L}_{m\cap\ell}\boldsymbol{X}\boldsymbol{\beta}_0/|I_m\cap I_\ell| \xrightarrow{\text{a.s.}} 0$.

For the other term, note that,

$$\begin{aligned}
\|\widehat{\boldsymbol{\beta}}_m\|_2 &\leq \|(\boldsymbol{X}^\top\boldsymbol{L}_m\boldsymbol{X}/k + \lambda\boldsymbol{I}_p)^{-1}(\boldsymbol{X}^\top\boldsymbol{L}_m\boldsymbol{y}/k)\|_2\\
&\leq \|(\boldsymbol{X}^\top\boldsymbol{L}_m\boldsymbol{X}/k + \lambda\boldsymbol{I}_p)^{-1}\boldsymbol{X}^\top\boldsymbol{L}_m/\sqrt{k}\| \cdot \|\boldsymbol{L}_m\boldsymbol{y}/\sqrt{k}\|_2\\
&\leq C\sqrt{\mathbb{E}[y_1^2]} \cdot \|(\boldsymbol{X}^\top\boldsymbol{L}_m\boldsymbol{X}/k + \lambda\boldsymbol{I}_p)^{-1}\boldsymbol{X}^\top\boldsymbol{L}_m/\sqrt{k}\|_{\mathrm{op}},
\end{aligned}$$

where the last inequality holds eventually almost surely since Assumption E implies 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 $C\sqrt{\mathbb{E}[y_1^2]}$ for some constant C. On the other hand, the 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 $\max_i s_i/(s_i^2+\lambda) \leq 1/s_{\min}$ 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, $\|\widehat{\boldsymbol{\beta}}_m\|_2 \to 0$ almost surely. Because $\|\widehat{\boldsymbol{\Sigma}}\|_{\mathrm{op}}$ is upper bounded almost surely, we further have $\boldsymbol{f}_{\mathrm{NL}}{}^\top \boldsymbol{L}_{m\cap\ell}\boldsymbol{X}\widehat{\boldsymbol{\beta}}_m/|I_m\cap I_\ell| \xrightarrow{\text{a.s.}} 0$. Consequently we have $\boldsymbol{f}_{\mathrm{NL}}{}^\top \boldsymbol{L}_{m\cap\ell}\boldsymbol{X}(\boldsymbol{\beta}_0 - \widehat{\boldsymbol{\beta}}_m)/|I_m\cap I_\ell| \xrightarrow{\text{a.s.}} 0$ and

$$\frac{1}{|I_m\cap I_\ell|}(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}_m)^\top \boldsymbol{L}_{m\cap\ell}(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}_\ell) \xrightarrow{\text{a.s.}} \boldsymbol{\beta}_0^\top\boldsymbol{\Sigma}\boldsymbol{\beta}_0 + \|f_{\mathrm{NL}}\|_{L_2}^2.$$

Since $\boldsymbol{S}_j = \boldsymbol{X}_{I_j}\widehat{\boldsymbol{\beta}}_j$ for $j \in \{m,\ell\}$, we have that $\mathrm{tr}(\boldsymbol{S}_j)/k \xrightarrow{\text{a.s.}} 0$. So the denominator converges to 1, almost surely.

On the other hand, the asymptotic formula for $\lambda > 0$ satisfies that

$$\lim_{\psi\to\infty} \frac{\lambda^2 v_p(-\lambda;\psi)^2(\|f_{\mathrm{NL}}\|_{L_2}^2 + \widetilde{c}_p(-\lambda;\psi))(1 + \widetilde{v}_p(-\lambda;\varphi_{m\ell},\psi))}{\lambda^2 v_p(-\lambda;\psi)^2} = \|f_{\mathrm{NL}}\|_{L_2}^2 + \boldsymbol{\beta}_0^\top\boldsymbol{\Sigma}\boldsymbol{\beta}_0,$$

where we use the property that $\lim_{\psi\to\infty} v_p(-\lambda;\psi) = 0$. Thus, the asymptotic formula is also well-defined and right continuous at $\psi = \infty$.

Part (2) Ridgeless predictor when $\lambda = 0$. Below, we analyze the numerator and the denominator separately for the first estimator. To indicate the dependency on p and λ, we denote the denominator and its asymptotic equivalent by

$$P_{p,\lambda} := D^{\mathrm{ovlp}}_{m,\ell}, \qquad Q_{p,\lambda} := \mathscr{D}^{\mathrm{ovlp}}_{m,\ell},$$

where we view k and n as sequences $\{k_p\}$ and $\{n_p\}$ that are indexed by p. For $j \in \{m, \ell\}$, since $\boldsymbol{S}_j \succeq \boldsymbol{0}_{n\times n}$ and $\|\boldsymbol{S}_j\|_{\rm op} = \|\boldsymbol{L}_j\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{L}_j\boldsymbol{X}/k+\lambda\boldsymbol{I}_p)^+\boldsymbol{X}^\top\boldsymbol{L}_j/|I_j|\|_{\rm op} \leq 1$ is upper bounded for $\lambda \geq 0$, with equality holds only if $\lambda = 0$. When $\lambda = 0$ and $\psi < 1$, we have $\operatorname{tr}(\boldsymbol{S}_j)/|I_j| = \operatorname{tr}(\boldsymbol{L}_j\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{L}_j\boldsymbol{X}/k)^{-1}\boldsymbol{X}^\top\boldsymbol{L}_j/|I_j|)/|I_j| = p/|I_j| \leq \psi$ almost surely. When $\lambda = 0$ and $\psi > 1$, we have that $\operatorname{tr}(\boldsymbol{S}_j)/|I_j| \geq r_{\min}r_{\max}^{-1}p/|I_j| \xrightarrow{\text{a.s.}} r_{\min}r_{\max}^{-1}\psi > \psi$. Thus, we have that $0 < P_{p,\lambda} < (1-\psi)^2$ for $\lambda > 0$ and $\lambda = 0$, $\psi \neq 1$.

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

$$\frac{\partial}{\partial\lambda}P_{p,\lambda} = -\frac{\operatorname{tr}(\frac{\partial}{\partial\lambda}\boldsymbol{S}_m)}{|I_m|} - \frac{\operatorname{tr}(\frac{\partial}{\partial\lambda}\boldsymbol{S}_\ell)}{|I_\ell|} + \frac{\operatorname{tr}(\frac{\partial}{\partial\lambda}\boldsymbol{S}_m)\operatorname{tr}(\boldsymbol{S}_\ell) + \operatorname{tr}(\boldsymbol{S}_m)\operatorname{tr}(\frac{\partial}{\partial\lambda}\boldsymbol{S}_\ell)}{|I_m||I_\ell|}.$$

For $j \in \{m, \ell\}$, note that

$$\frac{\partial}{\partial\lambda}\boldsymbol{S}_j = \boldsymbol{L}_j\boldsymbol{X}\left(\frac{\boldsymbol{X}^\top\boldsymbol{L}_j\boldsymbol{X}}{k} + \lambda\boldsymbol{I}\right)^{-2}\frac{\boldsymbol{X}^\top\boldsymbol{L}_j}{|I_j|}.$$

We also have $\|\partial\boldsymbol{S}_j/\partial\lambda\|_{\rm op}$ is upper bounded almost surely on Λ, and thus so does $|\partial P_{p,\lambda}/\partial\lambda|$. We know that $P_{p,\lambda} - Q_{p,\lambda} \xrightarrow{\text{a.s.}} 0$ for $\lambda > 0$. Define $Q_{p,0} := \lim_{\lambda\to0^+} Q_{p,\lambda} = (1-\psi)^2$, which is well-defined according to Du et al. (2023, Proposition E.2). The equicontinuity property, together with the Moore-Osgood theorem, we have that almost surely,

$$\lim_{p\to\infty}|P_{p,0} - Q_{p,0}| = \lim_{\lambda\to0^+}\lim_{p\to\infty}|P_{p,\lambda} - Q_{p,\lambda}| = \lim_{\lambda\to0^+} 0 = 0,$$

which proves that the denominator formula $Q_{p,0}$ is valid for $\lambda = 0$.

For the numerator, similarly, we define

$$P'_{p,\lambda} := N^{\rm ovlp}_{m,\ell} = \frac{1}{|I_m \cap I_\ell|}\boldsymbol{y}^\top L_{m\cap\ell}\boldsymbol{y} - \sum_{j\in\{m,\ell\}}\boldsymbol{y}^\top\boldsymbol{S}_j L_{m\cap\ell}\boldsymbol{y} + \boldsymbol{y}^\top\boldsymbol{S}_m\boldsymbol{S}_\ell\boldsymbol{y},$$

and $Q'_{p,\lambda} = \mathscr{N}^{\rm ovlp}_{m,\ell}$. By the similar argument above, the sequence of function $P'_{p,\lambda} - Q'_{p,\lambda}$ is equicontinuous on Λ and thus $Q'_{p,0} = \lim_{\lambda\to0^+} Q'_{p,\lambda}$ is well-defined. Finally, since the proof for the second estimator and the risk are similar, we omit it here for simplicity.

Part (3) Equicontinuity (in λ) of the ratio. Again, we focus on proof for the first estimator. Define $P''_{p,\lambda} := N^{\rm ovlp}_{m,\ell}/D^{\rm ovlp}_{m,\ell}$ and $Q''_{p,\lambda} := \mathscr{N}^{\rm ovlp}_{m,\ell}/\mathscr{D}^{\rm ovlp}_{m,\ell}$. From the proof of Part (1), we know that $D^{\rm ovlp}_{m,\ell}$ is positive and upper bounded almost surely, and $N^{\rm ovlp}_{m,\ell}$ is nonnegative and upper bounded almost surely. Thus, we have that $P''_{p,\lambda}$ is upper bounded almost surely over Λ.

On the other hand, from the monotonicity and boundedness of fixed-point quantities (Du et al., 2023, Lemma F.12), it follows that when $\lambda = 0$ and $\psi \neq 1$, $\mathscr{D}^{\rm ovlp}_{m,\ell} = (1-\psi)^2$; when $\lambda > 0$, $\mathscr{D}^{\rm ovlp}_{m,\ell}$ is a continuous function, with left limit at $\lambda = 0$ bounded away from zero and infinity and right limit $\lim_{\lambda\to\infty}\lambda^2 v_p(-\lambda;\psi) = 1$, and is therefore also bounded away from zero and infinity over Λ. This implies that $Q''_{p,\lambda}$ is upper bounded over Λ.

Next, we examine the derivative of $P''_{p,\lambda}$ and $Q''_{p,\lambda}$:

$$\frac{\partial}{\partial\lambda}P''_{p,\lambda} = \frac{\frac{\partial N^{\rm ovlp}_{m,\ell}}{\partial\lambda}D^{\rm ovlp}_{m,\ell} - N^{\rm ovlp}_{m,\ell}\frac{\partial D^{\rm ovlp}_{m,\ell}}{\partial\lambda}}{(D^{\rm ovlp}_{m,\ell})^2}, \quad \text{and} \quad \frac{\partial}{\partial\lambda}Q''_{p,\lambda} = \frac{\frac{\partial\mathscr{N}^{\rm ovlp}_{m,\ell}}{\partial\lambda}\mathscr{D}^{\rm ovlp}_{m,\ell} - \mathscr{N}^{\rm ovlp}_{m,\ell}\frac{\partial\mathscr{D}^{\rm ovlp}_{m,\ell}}{\partial\lambda}}{(\mathscr{D}^{\rm ovlp}_{m,\ell})^2}.$$

By a similar argument as above, we can also show that $|\partial P''_{p,\lambda}/\partial\lambda|$ and $|\partial Q''_{p,\lambda}/\partial\lambda|$ are upper bounded over Λ. Applying the Moore-Osgood theorem, the conclusion follows. $\square$

B.3. Proof of Theorem 6 (restated here for convenience)

Theorem 6 (Uniform consistency of corrected GCV in λ) *Under the same conditions in Theorem 4, it holds that* $|\widehat{R}^{\rm cgcv,ovlp}_{M,\lambda} - R_{M,\lambda}| \xrightarrow{\text{a.s.}} 0$ *and* $|\widehat{R}^{\rm cgcv,full}_{M,\lambda} - R_{M,\lambda}| \xrightarrow{\text{a.s.}} 0$. *Moreover, when* $\psi \neq 1$, *it holds that* $\sup_{\lambda\in\Lambda}|\widehat{R}^{\rm cgcv,ovlp}_{M,\lambda} - R_{M,\lambda}| \xrightarrow{\text{a.s.}} 0$ *and* $\sup_{\lambda\in\Lambda}|\widehat{R}^{\rm cgcv,full}_{M,\lambda} - R_{M,\lambda}| \xrightarrow{\text{a.s.}} 0$.

Pointwise consistency

From Lemma 24 and Lemma 27, we have that

$$\begin{aligned} R_M &= \|f_{\mathrm{NL}}\|_{L_2}^2 + \frac{1}{M^2}\sum_{m=1}^M \|\widehat{\boldsymbol{\beta}}(\mathcal{D}_{I_m}) - \boldsymbol{\beta}_0\|_{\boldsymbol{\Sigma}}^2 \\ &\qquad + \frac{1}{M^2}\sum_{m,\ell=1}^M (\widehat{\boldsymbol{\beta}}(\mathcal{D}_{I_m}) - \boldsymbol{\beta}_0)^\top \boldsymbol{\Sigma}(\widehat{\boldsymbol{\beta}}_\ell - \boldsymbol{\beta}_0)) \\ &\simeq \bigg(\frac{1}{M}(1+\widetilde{v}_p(-\lambda;\psi,\psi)) \\ &\qquad + \frac{M-1}{M}(1+\widetilde{v}_p(-\lambda;\phi,\psi))\bigg)(\|f_{\mathrm{NL}}\|_{L_2}^2 + \widetilde{c}_p(-\lambda;\psi)), \end{aligned} \tag{63}$$

$$\frac{1}{n}\|\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}_m)\|_2^2 \simeq d_p(-\lambda;\psi,\psi)(1+\widetilde{v}_p(-\lambda;\psi,\psi))(\|f_{\mathrm{NL}}\|_{L_2}^2 + \widetilde{c}_p(-\lambda;\psi)), \tag{64}$$

and

$$\begin{aligned} &\frac{1}{n}\|\boldsymbol{y} - \boldsymbol{X}\widetilde{\boldsymbol{\beta}}_M\|_2^2 \\ &= \frac{1}{M^2}\sum_{m=1}^M \frac{1}{n}\|\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}_m)\|_2^2 + \frac{1}{M^2}\sum_{m,\ell=1}^M \frac{1}{n}(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}_m)^\top(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}_\ell) \\ &\simeq \bigg(\frac{1}{M}d_p(-\lambda;\psi,\psi)(1+\widetilde{v}_p(-\lambda;\psi,\psi)) \\ &\qquad + \frac{M-1}{M}d_p(-\lambda;\phi,\psi)(1+\widetilde{v}_p(-\lambda;\phi,\psi))\bigg)(\|f_{\mathrm{NL}}\|_{L_2}^2 + \widetilde{c}_p(-\lambda;\psi)). \end{aligned} \tag{65}$$

On the other hand, from Du et al. (2023, Lemma 3.4.), the average degrees of freedom $\widetilde{\mathsf{df}}$ and the denominator $(1-\widetilde{\mathsf{df}}/n)^2$ of the naive GCV estimator satisfy that

$$\frac{1}{n}\widetilde{\mathsf{df}} \simeq \frac{\phi}{\psi}(1-\lambda v_p(-\lambda;\psi)), \tag{66}$$

and

$$(1-\widetilde{\mathsf{df}}/n)^2 \simeq \left(\frac{\psi-\phi}{\psi} + \frac{\phi}{\psi}\lambda v_p(-\lambda;\psi)\right)^2 = d_p(-\lambda;\phi,\psi). \tag{67}$$

Thus, from (63), (65) and (67), the difference between the prediction risk and the naive GCV estimator admits the following asymptotic representations:

$$\begin{aligned} &R_M - \frac{\|\boldsymbol{y} - \boldsymbol{X}\widetilde{\boldsymbol{\beta}}_M\|_2^2/n}{(1-\widetilde{\mathsf{df}}/n)^2} \\ &\simeq \frac{1}{M}\left(1 - \frac{d_p(-\lambda;\psi,\psi)}{d_p(-\lambda;\phi,\psi)}\right)(1+\widetilde{v}_p(-\lambda;\psi,\psi))(\|f_{\mathrm{NL}}\|_{L_2}^2 + \widetilde{c}_p(-\lambda;\psi)). \end{aligned} \tag{68}$$

On the other hand, from (64), (66) and (67), we also have that for all $m \in [M]$,

$$\begin{aligned} &\frac{1}{(1-\widetilde{\mathsf{df}}/n)^2}\cdot\frac{1}{M}\cdot\frac{(\psi-\phi)(\widetilde{\mathsf{df}}/n)^2}{\phi(1-\widetilde{\mathsf{df}}/n)^2 + (\psi-\phi)(\widetilde{\mathsf{df}}/n)^2}\cdot\frac{\|\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}_m\|_2^2}{n} \\ &\simeq \frac{1}{M}\frac{d_p(-\lambda;\psi,\psi)}{d_p(-\lambda;\phi,\psi)}\frac{(\psi-\phi)\frac{\phi^2}{\psi^2}(1-\lambda v_p(-\lambda;\psi))^2}{\phi d_p(-\lambda;\phi,\psi) + (\psi-\phi)\frac{\phi^2}{\psi^2}(1-\lambda v_p(-\lambda;\psi))^2}(1+\widetilde{v}_p(-\lambda;\psi,\psi))(\|f_{\mathrm{NL}}\|_{L_2}^2 + \widetilde{c}_p(-\lambda;\psi)) \end{aligned}$$

$$
= \frac{1}{M} \frac{\frac{\phi(\psi-\phi)}{\psi^2}(1-\lambda v_p(-\lambda;\psi))^2}{d_p(-\lambda;\phi,\psi)}(1+\widetilde{v}_p(-\lambda;\psi,\psi))(\|f_{\rm NL}\|_{L_2}^2+\widetilde{c}_p(-\lambda;\psi))
$$

$$
= -\frac{1}{M}\left(1-\frac{d_p(-\lambda;\psi,\psi)}{d_p(-\lambda;\phi,\psi)}\right)(1+\widetilde{v}_p(-\lambda;\psi,\psi))(\|f_{\rm NL}\|_{L_2}^2+\widetilde{c}_p(-\lambda;\psi)), \tag{69}
$$

by noting that $d_p(-\lambda;\psi,\psi)-d_p(-\lambda;\phi,\psi)=\phi(\psi-\phi)\psi^{-2}(1-\lambda v_p(-\lambda;\psi))^2$. Matching (68) and (69) finishes the proof when $\widehat{R}_{m,m}$ is the full-estimator. The proof when $\widehat{R}_{m,m}$ is the ovlp-estimator follows similarly.

Uniform consistency

From the proof of Lemma 28, we have that

$$
(1-\widetilde{\mathsf{df}}/n)^2 \simeq \left(1-\frac{\phi}{\psi}(1-\lambda v_p(-\lambda;\psi)\right)^2 =: \mathscr{D}_{m,\ell}^{\rm full}, \qquad \text{for all} \quad m\neq\ell,
$$

and

$$
\begin{aligned}
\widetilde{R}_M^{\rm gcv} &\simeq \frac{1}{M^2}\sum_{m,\ell\in[M]}\frac{\mathscr{N}_{m,\ell}^{\rm full}}{\mathscr{D}_{1,2}^{\rm full}} \simeq \frac{1}{M^2}\sum_{m,\ell\in[M]}R_{m,\ell}-\frac{1}{M^2}\sum_{m\in[M]}\frac{\mathscr{D}_{m,m}^{\rm full}}{\mathscr{D}_{1,2}^{\rm full}}R_m\\
\widehat{R}_{m,m}^{\#} &\simeq R_{m,m}.
\end{aligned}
$$

Then we have

$$
\begin{aligned}
\widehat{R}_M^{{\rm cgcv},\#} &= \widetilde{R}_M^{\rm gcv} - \frac{1}{M}\frac{(\widetilde{\mathsf{df}}/n)^2}{(1-\widetilde{\mathsf{df}}/n)^2}\frac{\psi-\phi}{\phi}\frac{1}{M}\sum_{m=1}^M \widehat{R}_{m,m}^{\#}\\
&\simeq \widetilde{R}_M^{\rm gcv} - \frac{1}{M^2}\sum_{m=1}^M\frac{\mathscr{D}_{1,1}^{\rm full}-\mathscr{D}_{1,2}^{\rm full}}{\mathscr{D}_{1,2}^{\rm full}}R_{m,m}\\
&\simeq \frac{1}{M^2}\sum_{m,\ell\in[M]}R_{m,\ell}\\
&= R_M,
\end{aligned}
$$

which establishes the point-wise consistency.

Similar to the proof of Theorem 5, the uniform equicontinuity of $\widetilde{R}_M^{\rm gcv}$ and $R_{m,\ell}$ to their asymptotic limits follows from Lemma 29. And the uniformity for $|\widehat{R}_M^{{\rm cgcv},\#}-R_M|$ follows similarly.

B.4. Technical lemmas and their proofs

Lemma 30 (Bias term of risk)**.** Suppose the same assumptions in Theorem 4 hold and let $\varphi_{m\ell}=\psi\,\mathbb{1}_{\{m=\ell\}}+\phi\,\mathbb{1}_{\{m\neq\ell\}}$, then it holds that:

1. $\lambda^2\boldsymbol{\beta}_0^\top\boldsymbol{M}_m\boldsymbol{\Sigma}\boldsymbol{M}_\ell\boldsymbol{\beta}_0 \simeq \widetilde{c}_p(-\lambda;\psi)(1+\widetilde{v}_p(-\lambda;\varphi_{m\ell},\psi))$.
2. $\lambda^2\boldsymbol{\beta}_0^\top\boldsymbol{M}_m\widehat{\boldsymbol{\Sigma}}_{m\cap\ell}\boldsymbol{M}_\ell\boldsymbol{\beta}_0 \simeq \lambda^2 v_p(-\lambda;\psi)^2\widetilde{c}_p(-\lambda;\psi)(1+\widetilde{v}_p(-\lambda;\varphi_{m\ell},\psi))$
3. $\lambda^2\boldsymbol{\beta}_0^\top\boldsymbol{M}_m\widehat{\boldsymbol{\Sigma}}\boldsymbol{M}_\ell\boldsymbol{\beta}_0 \simeq \begin{cases}\left(\dfrac{\psi-\phi}{\psi}+\dfrac{\phi}{\psi}\lambda^2 v_p(-\lambda;\psi)^2\right)\widetilde{c}_p(-\lambda;\psi)(1+\widetilde{v}_p(-\lambda;\varphi_{m\ell},\psi)), & m=\ell,\\ \left(\dfrac{\psi-\phi}{\psi}+\dfrac{\phi}{\psi}\lambda v_p(-\lambda;\psi)\right)^2\widetilde{c}_p(-\lambda;\psi)(1+\widetilde{v}_p(-\lambda;\varphi_{m\ell},\psi)), & m\neq\ell.\end{cases}$

Here the nonnegative constants $v_p(-\lambda;\psi)$, $\widetilde{v}_p(-\lambda;\phi,\psi)$ and $\widetilde{c}_p(-\lambda;\psi)$ are defined through the following equations:

$$
\frac{1}{v_p(-\lambda;\psi)} = \lambda+\psi\int\frac{r}{1+v_p(-\lambda;\psi)r}\,\mathrm{d}H_p(r),
$$

$$\widetilde{v}_p(-\lambda;\phi,\psi) = \frac{\phi\displaystyle\int \frac{r^2}{(1+v_p(-\lambda;\psi)r)^2}\,\mathrm{d}H_p(r)}{v_p(-\lambda;\psi)^{-2} - \phi\displaystyle\int \frac{r^2}{(1+v_p(-\lambda;\psi)r)^2}\,\mathrm{d}H_p(r)},$$
$$\widetilde{c}_p(-\lambda;\psi) = \boldsymbol{\beta}_0^\top(v_p(-\lambda;\psi)\boldsymbol{\Sigma}+\boldsymbol{I}_p)^{-1}\boldsymbol{\Sigma}(v_p(-\lambda;\psi)\boldsymbol{\Sigma}+\boldsymbol{I}_p)^{-1}\boldsymbol{\beta}_0.$$

Proof of Lemma 30. Note that $\boldsymbol{\beta}_0$ is independent of $\boldsymbol{M}_m\boldsymbol{\Sigma}\boldsymbol{M}_\ell$, $\boldsymbol{M}_m\widehat{\boldsymbol{\Sigma}}_{m\cap\ell}\boldsymbol{M}_\ell$ and $\boldsymbol{M}_m\widehat{\boldsymbol{\Sigma}}\boldsymbol{M}_\ell$. We analyze the deterministic equivalents of the latter for the three cases.

Part (1) From Patil et al. (2023, Lemma S.2.4), we have that

$$\lambda^2\boldsymbol{M}_m\boldsymbol{\Sigma}\boldsymbol{M}_\ell \simeq (v_p(-\lambda;\psi)\boldsymbol{\Sigma}+\boldsymbol{I}_p)^{-1}\,(1+\widetilde{v}_p(-\lambda;\varphi_{m\ell},\psi))\boldsymbol{\Sigma}\,(v_p(-\lambda;\psi)\boldsymbol{\Sigma}+\boldsymbol{I}_p)^{-1}.$$

Since $\|\boldsymbol{\beta}_0\|_2$ is almost surely bounded from Patil and Du (2023, Lemma D.5.), by the trace property of deterministic equivalents in Patil and Du (2023, Lemma E.3 (4)), we have

$$\begin{aligned}\boldsymbol{\beta}_0^\top\boldsymbol{M}_m\boldsymbol{\Sigma}\boldsymbol{M}_\ell\boldsymbol{\beta}_0 &\xlongequal{\text{a.s.}} \boldsymbol{\beta}_0^\top\,(v_p(-\lambda;\psi)\boldsymbol{\Sigma}+\boldsymbol{I}_p)^{-1}\,(1+\widetilde{v}_p(-\lambda;\varphi_{m\ell},\psi))\boldsymbol{\Sigma}\,(v_p(-\lambda;\psi)\boldsymbol{\Sigma}+\boldsymbol{I}_p)^{-1}\,\boldsymbol{\beta}_0\\ &= \widetilde{c}_p(-\lambda;\psi)(1+\widetilde{v}_p(-\lambda;\varphi_{m\ell},\psi)).\end{aligned}$$

Part (2) From Du et al. (2023, Lemma D.6 (1) and Lemma F.8 (3)), we have that

$$\lambda^2\boldsymbol{M}_m\widehat{\boldsymbol{\Sigma}}_{m\cap\ell}\boldsymbol{M}_\ell \simeq \lambda^2 v_p(-\lambda;\psi)^2\,(v_p(-\lambda;\psi)\boldsymbol{\Sigma}+\boldsymbol{I}_p)^{-1}\,(1+\widetilde{v}_p(-\lambda;\varphi_{m\ell},\psi))\boldsymbol{\Sigma}\,(v_p(-\lambda;\psi)\boldsymbol{\Sigma}+\boldsymbol{I}_p)^{-1}. \tag{70}$$

Then, by the trace property of deterministic equivalents in Patil and Du (2023, Lemma E.3 (4)), we have

$$\begin{aligned}\boldsymbol{\beta}_0^\top\boldsymbol{M}_m\widehat{\boldsymbol{\Sigma}}_{m\cap\ell}\boldsymbol{M}_\ell\boldsymbol{\beta}_0 &\xlongequal{\text{a.s.}} \lambda^2 v_p(-\lambda;\psi)^2\boldsymbol{\beta}_0^\top\,(v_p(-\lambda;\psi)\boldsymbol{\Sigma}+\boldsymbol{I}_p)^{-1}\,(1+\widetilde{v}_p(-\lambda;\varphi_{m\ell},\psi))\boldsymbol{\Sigma}\,(v_p(-\lambda;\psi)\boldsymbol{\Sigma}+\boldsymbol{I}_p)^{-1}\,\boldsymbol{\beta}_0\\ &= \lambda^2 v_p(-\lambda;\psi)^2\widetilde{c}_p(-\lambda;\psi)(1+\widetilde{v}_p(-\lambda;\varphi_{m\ell},\psi)).\end{aligned}$$

Part (3) Note that $\widehat{\boldsymbol{\Sigma}} = \frac{|I_m\cup I_\ell|}{n}\widehat{\boldsymbol{\Sigma}}_{m\cup\ell} + \frac{|(I_m\cup I_\ell)^c|}{n}\widehat{\boldsymbol{\Sigma}}_{(m\cup\ell)^c}$, we have

$$\boldsymbol{M}_m\widehat{\boldsymbol{\Sigma}}\boldsymbol{M}_\ell = \frac{|I_m\cup I_\ell|}{n}\boldsymbol{M}_m\widehat{\boldsymbol{\Sigma}}_{m\cup\ell}\boldsymbol{M}_\ell + \frac{|(I_m\cup I_\ell)^c|}{n}\boldsymbol{M}_m\widehat{\boldsymbol{\Sigma}}_{(m\cup\ell)^c}\boldsymbol{M}_\ell. \tag{71}$$

For the first term in (71), from Du et al. (2023, Equation (54)), when $m\neq\ell$, it holds that

$$\lambda^2\boldsymbol{M}_m\widehat{\boldsymbol{\Sigma}}_{m\cup\ell}\boldsymbol{M}_\ell \simeq \lambda^2 v_p(-\lambda;\psi)^2(1+\widetilde{v}_p(-\lambda;\phi,\psi))\left(\frac{2(\psi-\phi)}{2\psi-\phi}\frac{1}{\lambda v_p(-\lambda;\psi)} + \frac{\phi}{2\psi-\phi}\right)(v_p(-\lambda;\psi)\boldsymbol{\Sigma}+\boldsymbol{I}_p)^{-2}\boldsymbol{\Sigma},$$

From Part (2), when $m=\ell$, it holds that

$$\lambda^2\boldsymbol{M}_m\widehat{\boldsymbol{\Sigma}}_{m\cup\ell}\boldsymbol{M}_\ell \simeq \lambda^2 v_p(-\lambda;\psi)^2\,(v_p(-\lambda;\psi)\boldsymbol{\Sigma}+\boldsymbol{I}_p)^{-1}\,(1+\widetilde{v}_p(-\lambda;\varphi_{m\ell},\psi))\boldsymbol{\Sigma}\,(v_p(-\lambda;\psi)\boldsymbol{\Sigma}+\boldsymbol{I}_p)^{-1},$$

For the second term in (71), from Du et al. (2023, Lemma F.8 (1)), we have

$$\lambda^2\boldsymbol{M}_m\widehat{\boldsymbol{\Sigma}}_{(m\cup\ell)^c}\boldsymbol{M}_\ell \simeq \lambda^2\boldsymbol{M}_1\boldsymbol{\Sigma}\boldsymbol{M}_2 \simeq (v_p(-\lambda;\psi)\boldsymbol{\Sigma}+\boldsymbol{I}_p)^{-1}\,(1+\widetilde{v}_p(-\lambda;\varphi_{m\ell},\psi))\boldsymbol{\Sigma}\,(v_p(-\lambda;\psi)\boldsymbol{\Sigma}+\boldsymbol{I}_p)^{-1},$$

where the last equivalence is from Part (1).

When $m=\ell$, the coefficients in (71) concentrate $|I_m\cup I_\ell|/n \xrightarrow{\text{a.s.}} \phi/\psi$ and $|(I_m\cup I_\ell)^c|/n \xrightarrow{\text{a.s.}} (\psi-\phi)/\psi$ from Du et al. (2023, Lemma G.6). Then (71) implies that

$$\lambda^2\boldsymbol{M}_m\widehat{\boldsymbol{\Sigma}}\boldsymbol{M}_\ell \simeq \left(\frac{\psi-\phi}{\psi} + \frac{\phi}{\psi}\lambda^2 v_p(-\lambda;\psi)^2\right)(v_p(-\lambda;\psi)\boldsymbol{\Sigma}+\boldsymbol{I}_p)^{-1}\,(1+\widetilde{v}_p(-\lambda;\varphi_{m\ell},\psi))\boldsymbol{\Sigma}\,(v_p(-\lambda;\psi)\boldsymbol{\Sigma}+\boldsymbol{I}_p)^{-1}.$$

When $m \neq \ell$, the coefficients in (71) concentrate $|I_m \cup I_\ell|/n \xrightarrow{\text{a.s.}} \phi(2\psi-\phi)/\psi^2$ and $|(I_m \cup I_\ell)^c|/n \xrightarrow{\text{a.s.}} (\psi-\phi)^2/\psi^2$ from Du et al. (2023, Lemma G.6). Then (71) implies that

$$\lambda^2 \boldsymbol{M}_m \widehat{\boldsymbol{\Sigma}} \boldsymbol{M}_\ell \simeq \left(\frac{\psi-\phi}{\psi} + \frac{\phi}{\psi}\lambda v_p(-\lambda;\psi)\right)^2 (v_p(-\lambda;\psi)\boldsymbol{\Sigma} + \boldsymbol{I}_p)^{-1} (1+\widetilde{v}_p(-\lambda;\varphi_{m\ell},\psi))\boldsymbol{\Sigma} (v_p(-\lambda;\psi)\boldsymbol{\Sigma} + \boldsymbol{I}_p)^{-1}.$$

Finally, applying the trace property of deterministic equivalents in Patil and Du (2023, Lemma E.3 (4)) as in the previous parts completes the proof. □

Lemma 31 (Variance term of risk)**.** Suppose the same assumptions in Theorem 4 hold and let $\varphi_{m\ell} = \psi\,\mathbb{1}_{\{m=\ell\}} + \phi\,\mathbb{1}_{\{m\neq\ell\}}$, then it holds that:

1. $k^{-2}\boldsymbol{f}_0^\top \boldsymbol{X}\boldsymbol{M}_m\boldsymbol{\Sigma}\boldsymbol{M}_\ell\boldsymbol{X}^\top \boldsymbol{f}_0 \simeq \|f_{\rm NL}\|_{L_2}^2 \widetilde{v}_p(-\lambda;\varphi_{m\ell},\psi)$.
2. $\frac{1}{|I_m\cap I_\ell|}\boldsymbol{f}_0^\top \left(\boldsymbol{I}_p - \frac{\boldsymbol{X}\boldsymbol{M}_m\boldsymbol{X}^\top}{k}\right)\boldsymbol{L}_{m\cap\ell}\left(\boldsymbol{I}_p - \frac{\boldsymbol{X}\boldsymbol{M}_\ell\boldsymbol{X}^\top}{k}\right)\boldsymbol{f}_0 \simeq \|f_{\rm NL}\|_{L_2}^2 \lambda^2 v_p(-\lambda;\psi))^2(1+\widetilde{v}_p(-\lambda;\varphi_{m\ell},\psi))$.
3. $\frac{|I_m\cup I_\ell|}{n}\boldsymbol{f}_1^\top \left(\boldsymbol{I}_p - \frac{\boldsymbol{X}\boldsymbol{M}_m\boldsymbol{X}^\top}{k}\right)\boldsymbol{L}_{m\cup\ell}\left(\boldsymbol{I}_p - \frac{\boldsymbol{X}\boldsymbol{M}_\ell\boldsymbol{X}^\top}{k}\right)\boldsymbol{f}_1$
$+ \frac{|(I_m\cup I_\ell)^c|}{n}\|\boldsymbol{f}_0\|_2^2 + \frac{|(I_m\cup I_\ell)^c|^2}{nk^2}\boldsymbol{f}_0^\top\boldsymbol{X}\boldsymbol{M}_m\widehat{\boldsymbol{\Sigma}}_{(m\cup\ell)^c}\boldsymbol{M}_\ell\boldsymbol{X}^\top\boldsymbol{f}_0$
$$\simeq \begin{cases} \|f_{\rm NL}\|_{L_2}^2 \left(\dfrac{\psi-\phi}{\psi} + \dfrac{\phi}{\psi}\lambda^2 v_p(-\lambda;\psi)^2\right)(1+\widetilde{v}_p(-\lambda;\varphi_{m\ell},\psi)), & m=\ell,\\ \|f_{\rm NL}\|_{L_2}^2 \left(\dfrac{\psi-\phi}{\psi} + \dfrac{\phi}{\psi}\lambda v_p(-\lambda;\psi)\right)^2(1+\widetilde{v}_p(-\lambda;\varphi_{m\ell},\psi)), & m\neq\ell. \end{cases}$$

Here $\boldsymbol{f}_0 = \boldsymbol{L}_{m\cap\ell}\boldsymbol{f}_{\rm NL}$, $\boldsymbol{f}_1 = \boldsymbol{L}_{m\cup\ell}\boldsymbol{f}_{\rm NL}$ and the nonnegative constants $v_p(-\lambda;\psi)$ and $\widetilde{v}_p(-\lambda;\phi,\psi)$ are defined in Lemma 30.

Proof. Since $\|f_{\rm NL}\|_{4+\delta} < \infty$ from Patil and Du (2023, Lemma D.5), we have that $\|\boldsymbol{f}_0\|_2^2/|I_m\cap I_\ell| \xrightarrow{\text{a.s.}} \|f_{\rm NL}\|_{L_2}^2$ by strong law of large number. Then, from Lemma 32, we have that the quadratic term

$$\frac{1}{|I_m\cap I_\ell|}\boldsymbol{f}_0^\top\boldsymbol{X}\boldsymbol{A}\boldsymbol{X}^\top\boldsymbol{f}_0 \simeq \frac{1}{|I_m\cap I_\ell|}\|f_{\rm NL}\|_{L_2}^2 \operatorname{tr}(\boldsymbol{L}_{m\cap\ell}\boldsymbol{X}\boldsymbol{A}\boldsymbol{X}^\top\boldsymbol{L}_{m\cap\ell}) = \|f_{\rm NL}\|_{L_2}^2 \operatorname{tr}(\boldsymbol{A}\widehat{\boldsymbol{\Sigma}}_{m\cap\ell}) \tag{72}$$

for any symmetric matrix $\boldsymbol{A}$ with bounded operator norm. We next apply this result for different values of $\boldsymbol{A}$.

Part (1) Let $\boldsymbol{A} = \boldsymbol{M}_m\boldsymbol{\Sigma}\boldsymbol{M}_\ell$. Since from (70) and the product rule (Patil et al., 2023, Lemma S.7.4 (3)), we have that

$$\boldsymbol{M}_\ell\widehat{\boldsymbol{\Sigma}}_{m\cap\ell}\boldsymbol{M}_m\boldsymbol{\Sigma} \simeq v_p(-\lambda;\psi)^2 (v_p(-\lambda;\psi)\boldsymbol{\Sigma}+\boldsymbol{I}_p)^{-1}(1+\widetilde{v}_p(-\lambda;\varphi_{m\ell},\psi))\boldsymbol{\Sigma}(v_p(-\lambda;\psi)\boldsymbol{\Sigma}+\boldsymbol{I}_p)^{-1}\boldsymbol{\Sigma}.$$

Then by the trace property of deterministic equivalents in Patil and Du (2023, Lemma E.3 (4)), we have

$$\begin{aligned}\frac{p}{k\,\mathbb{1}_{\{m=\ell\}} + n\,\mathbb{1}_{\{m\neq\ell\}}}\cdot\frac{1}{p}\operatorname{tr}(\boldsymbol{A}\widehat{\boldsymbol{\Sigma}}_{m\cap\ell}) &\simeq \varphi_{m\ell}(1+\widetilde{v}_p(-\lambda;\varphi_{m\ell},\psi))\int\left(\frac{v_p(-\lambda;\psi)r}{1+v_p(-\lambda;\psi)r}\right)^2 \mathrm{d}H_p(r)\\ &= \widetilde{v}_p(-\lambda;\varphi_{m\ell},\psi)).\end{aligned}$$

Finally, note that $|I_m\cap I_\ell|(k\,\mathbb{1}_{\{m=\ell\}} + n\,\mathbb{1}_{\{m\neq\ell\}}) \simeq k^2$. This implies that

$$k^{-2}\boldsymbol{f}_0^\top\boldsymbol{X}\boldsymbol{M}_m\boldsymbol{\Sigma}\boldsymbol{M}_\ell\boldsymbol{X}^\top\boldsymbol{f}_0 \simeq \|f_{\rm NL}\|_{L_2}^2\widetilde{v}_p(-\lambda;\varphi_{m\ell},\psi).$$

Part (2) Note that

$$\frac{1}{|I_m\cap I_\ell|}\boldsymbol{f}_0^\top\left(\boldsymbol{I}_p - \frac{\boldsymbol{X}\boldsymbol{M}_m\boldsymbol{X}^\top}{k}\right)\boldsymbol{L}_{m\cap\ell}\left(\boldsymbol{I}_p - \frac{\boldsymbol{X}\boldsymbol{M}_\ell\boldsymbol{X}^\top}{k}\right)\boldsymbol{f}_0$$

$$= \frac{1}{|I_m \cap I_\ell|} \boldsymbol{f}_0^\top \boldsymbol{f}_0 - \frac{1}{|I_m \cap I_\ell|} \sum_{j\in\{m,\ell\}} \boldsymbol{f}_0^\top \frac{\boldsymbol{X}\boldsymbol{M}_j\boldsymbol{X}^\top}{k} \boldsymbol{f}_0 + \frac{1}{k^2} \boldsymbol{f}_0^\top \boldsymbol{X}\boldsymbol{M}_m \widehat{\boldsymbol{\Sigma}}_{m\cap\ell} \boldsymbol{M}_\ell \boldsymbol{X}^\top \boldsymbol{f}_0.$$

We next analyze the three terms separately for $m \neq \ell$. By the law of large numbers, the first term converges as $\boldsymbol{f}_0^\top \boldsymbol{f}_0/|I_m \cap I_\ell| \xrightarrow{\text{a.s.}} \|f_{\text{NL}}\|_{L_2}^2$. For the second term, let $\boldsymbol{A} = \boldsymbol{M}_m$ and $\boldsymbol{M}_\ell$. From Patil and Du (2023, Corollary F.5 and Lemma F.8 (4)), it follows that for $j \in \{m, \ell\}$,

$$\boldsymbol{M}_j \widehat{\boldsymbol{\Sigma}}_{m\cap\ell} \simeq \boldsymbol{I}_p - (v_p(-\lambda;\psi)\boldsymbol{\Sigma} + \boldsymbol{I}_p)^{-1},$$

and thus, we have

$$\frac{1}{p} \operatorname{tr}(\boldsymbol{M}_j \widehat{\boldsymbol{\Sigma}}_{m\cap\ell}) \simeq \int \frac{v_p(-\lambda;\psi)r}{1+v_p(-\lambda;\psi)r} \,\mathrm{d}H_p(r) = \psi^{-1}(1-\lambda v_p(-\lambda;\psi)).$$

It then follows that

$$\frac{1}{|I_m \cap I_\ell|} \sum_{j\in\{m,\ell\}} \boldsymbol{f}_0^\top \frac{\boldsymbol{X}\boldsymbol{M}_j\boldsymbol{X}^\top}{k} \boldsymbol{f}_0 \simeq 2\|f_{\text{NL}}\|_{L_2}^2 (1-\lambda v_p(-\lambda;\psi)).$$

For the third term, let $\boldsymbol{A} = \boldsymbol{M}_m \widehat{\boldsymbol{\Sigma}}_{m\cap\ell} \boldsymbol{M}_\ell$. When $m \neq \ell$, from Patil and Du (2023, Lemma D.7 (1) and Lemma F.8 (5)), we have that

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

where $\widetilde{v}_v(-\lambda;\psi) = v_p(-\lambda;\psi)^2(1+\widetilde{v}_p(-\lambda;\phi,\psi))$. Then from (72), we have

$$\begin{aligned}&\frac{1}{|I_m \cap I_\ell|} \boldsymbol{f}_0^\top \left(\boldsymbol{I}_p - \frac{\boldsymbol{X}\boldsymbol{M}_m\boldsymbol{X}^\top}{k} \right) \boldsymbol{L}_{m\cap\ell} \left(\boldsymbol{I}_p - \frac{\boldsymbol{X}\boldsymbol{M}_\ell\boldsymbol{X}^\top}{k} \right) \boldsymbol{f}_0 \\ &\simeq \|f_{\text{NL}}\|_{L_2}^2 \Bigg(1 - 2(1-\lambda v_p(-\lambda;\psi)) + \psi \left(v_p(-\lambda;\psi) - \frac{\psi-\phi}{\psi} \lambda \widetilde{v}_v(-\lambda;\phi,\psi) \right) \int \frac{r}{1+v_p(-\lambda;\psi)r} \,\mathrm{d}H_p(r) \\ &\qquad -\phi\lambda\widetilde{v}_v(-\lambda;\phi,\psi) \int \frac{r}{(1+v_p(-\lambda;\psi)r)^2} \,\mathrm{d}H_p(r) \Bigg) \\ &= \|f_{\text{NL}}\|_{L_2}^2 \lambda \widetilde{v}_v(-\lambda;\phi,\psi) \Bigg(\frac{1}{v_p(-\lambda;\psi)} + \phi \int \left(\frac{r}{1+v_p(-\lambda;\psi)r} \right)^2 \mathrm{d}H_p(r) - (\psi-\phi) \int \frac{r}{1+v_p(-\lambda;\psi)r} \,\mathrm{d}H_p(r) \\ &\qquad -\phi \int \frac{r}{(1+v_p(-\lambda;\psi)r)^2} \,\mathrm{d}H_p(r) \Bigg) \\ &= \|f_{\text{NL}}\|_{L_2}^2 \lambda \widetilde{v}_v(-\lambda;\phi,\psi) \left(\frac{1}{v_p(-\lambda;\psi)} - \psi \int \frac{r}{1+v_p(-\lambda;\psi)r} \,\mathrm{d}H_p(r) \right) \\ &= \|f_{\text{NL}}\|_{L_2}^2 \lambda^2 \widetilde{v}_v(-\lambda;\phi,\psi) \\ &= \|f_{\text{NL}}\|_{L_2}^2 \lambda^2 v_p(-\lambda;\psi)^2 (1+\widetilde{v}_p(-\lambda;\phi,\psi)),\end{aligned}$$

when $m \neq \ell$.

When $m = \ell$, from (72) and Du et al. (2023, Lemma D.7 (1)), we have

$$\begin{aligned}&\frac{1}{|I_m \cap I_\ell|} \boldsymbol{f}_0^\top \left(\boldsymbol{I}_p - \frac{\boldsymbol{X}\boldsymbol{M}_m\boldsymbol{X}^\top}{k} \right) \boldsymbol{L}_{m\cap\ell} \left(\boldsymbol{I}_p - \frac{\boldsymbol{X}\boldsymbol{M}_\ell\boldsymbol{X}^\top}{k} \right) \boldsymbol{f}_0 \\ &= \frac{1}{k} \boldsymbol{f}_0^\top \left(\boldsymbol{I}_p - \frac{\boldsymbol{X}\boldsymbol{M}_m\boldsymbol{X}^\top}{k} \right) \boldsymbol{L}_m \left(\boldsymbol{I}_p - \frac{\boldsymbol{X}\boldsymbol{M}_m\boldsymbol{X}^\top}{k} \right) \boldsymbol{f}_0\end{aligned}$$

$$\simeq \|f_{\rm NL}\|_{L_2}^2(1 - \frac{2}{k}\operatorname{tr}(\boldsymbol{M}\widehat{\boldsymbol{\Sigma}}_m) + \frac{1}{k}\operatorname{tr}(\boldsymbol{M}_m\widehat{\boldsymbol{\Sigma}}_m\boldsymbol{M}_m\widehat{\boldsymbol{\Sigma}}_m))$$
$$\simeq \|f_{\rm NL}\|_{L_2}^2\lambda^2 v_p(-\lambda;\psi)^2(1+\widetilde{v}_p(-\lambda;\psi,\psi)).$$

Combining the above results finish the proof for Part (2).

Part (3) We analyze the three terms separately for $m \neq \ell$. From Du et al. (2023, Lemma D.7 (3)), we have

$$\begin{aligned}&\frac{1}{|I_m\cup I_\ell|}\operatorname{tr}\left(\left(\boldsymbol{I}_p - \frac{\boldsymbol{X}\boldsymbol{M}_m\boldsymbol{X}^\top}{k}\right)\boldsymbol{L}_{m\cup\ell}\left(\boldsymbol{I}_p - \frac{\boldsymbol{X}\boldsymbol{M}_\ell\boldsymbol{X}^\top}{k}\right)\right)\\ &\simeq \lambda^2 v_p(-\lambda;\psi)^2\left(\frac{2(\psi-\phi)}{2\psi-\phi}\frac{1}{\lambda v_p(-\lambda;\psi)} + \frac{\phi}{2\psi-\phi}\right)(1+\widetilde{v}_p(-\lambda;\psi,\psi)).\end{aligned}$$

From Lemma 32, it then follows that

$$\begin{aligned}&\frac{|I_m\cup I_\ell|}{n}\boldsymbol{f}_1^\top\left(\boldsymbol{I}_p - \frac{\boldsymbol{X}\boldsymbol{M}_m\boldsymbol{X}^\top}{k}\right)\boldsymbol{L}_{m\cup\ell}\left(\boldsymbol{I}_p - \frac{\boldsymbol{X}\boldsymbol{M}_\ell\boldsymbol{X}^\top}{k}\right)\boldsymbol{f}_1\\ &\simeq \|f_{\rm NL}\|_{L^2}^2\frac{\phi(2\psi-\phi)}{\psi^2}\lambda^2 v_p(-\lambda;\psi)^2\left(\frac{2(\psi-\phi)}{2\psi-\phi}\frac{1}{\lambda v_p(-\lambda;\psi)} + \frac{\phi}{2\psi-\phi}\right)(1+\widetilde{v}_p(-\lambda;\psi,\psi)).\end{aligned}$$

From strong law of large numbers, the second term converges $\frac{|(I_m\cup I_\ell)^c|}{n}\|\boldsymbol{f}_0\|_2^2 \xrightarrow{\text{a.s.}} \|f_{\rm NL}\|_{L^2}^2$. For the third term, let $\boldsymbol{A} = \boldsymbol{M}_m\widehat{\boldsymbol{\Sigma}}_{(m\cup\ell)^c}\boldsymbol{M}_\ell$, then we have

$$\frac{1}{p}\operatorname{tr}(\boldsymbol{M}_m\widehat{\boldsymbol{\Sigma}}_{(m\cup\ell)^c}\boldsymbol{M}_\ell\widehat{\boldsymbol{\Sigma}}_{m\cap\ell} \simeq \frac{1}{p}\operatorname{tr}(\boldsymbol{M}_m\boldsymbol{\Sigma}\boldsymbol{M}_\ell\widehat{\boldsymbol{\Sigma}}_{m\cap\ell} \simeq \psi^{-1}\widetilde{v}_p(-\lambda;\phi,\psi),$$

where the first equality is from the conditional independence property and the second is from Du et al. (2023, Lemma F.8 (3)). Again, from Lemma 32, it follows that

$$\frac{|(I_m\cup I_\ell)^c|^2}{nk^2}\boldsymbol{f}_0^\top\boldsymbol{X}\boldsymbol{M}_m\widehat{\boldsymbol{\Sigma}}_{(m\cup\ell)^c}\boldsymbol{M}_\ell\boldsymbol{X}^\top\boldsymbol{f}_0 \simeq \|f_{\rm NL}\|_{L_2}^2\left(\frac{\psi-\phi}{\psi}\right)^2\widetilde{v}_p(-\lambda;\phi,\psi).$$

Combining the above results finishes the proof of Part (3) for $m \neq \ell$.

When $m = \ell$, the formula simplifies to

$$\frac{1}{n}\boldsymbol{f}_0^\top\left(\boldsymbol{I}_p - \frac{\boldsymbol{X}\boldsymbol{M}_m\boldsymbol{X}^\top}{k}\right)\left(\boldsymbol{I}_p - \frac{\boldsymbol{X}\boldsymbol{M}_\ell\boldsymbol{X}^\top}{k}\right)\boldsymbol{f}_0.$$

From (72), Part (2) and Patil et al. (2023, Lemma S.2.5 (1)), we have

$$\begin{aligned}&\frac{1}{n}\boldsymbol{f}_0^\top\left(\boldsymbol{I}_p - \frac{\boldsymbol{X}\boldsymbol{M}_m\boldsymbol{X}^\top}{k}\right)\left(\boldsymbol{I}_p - \frac{\boldsymbol{X}\boldsymbol{M}_\ell\boldsymbol{X}^\top}{k}\right)\boldsymbol{f}_0\\ &= \frac{1}{n}\boldsymbol{f}_0^\top\left(\boldsymbol{I}_p - \frac{\boldsymbol{X}\boldsymbol{M}_m\boldsymbol{X}^\top}{k}\right)\left(\boldsymbol{I}_p - \frac{\boldsymbol{X}\boldsymbol{M}_m\boldsymbol{X}^\top}{k}\right)\boldsymbol{f}_0\\ &\simeq \|f_{\rm NL}\|_{L_2}^2\frac{k}{n}\left(1 - \frac{2}{k}\operatorname{tr}(\boldsymbol{M}\widehat{\boldsymbol{\Sigma}}_m) + \frac{1}{k}\operatorname{tr}(\boldsymbol{M}_m\widehat{\boldsymbol{\Sigma}}_m\boldsymbol{M}_m\widehat{\boldsymbol{\Sigma}}_m)\right)\\ &\qquad + \|f_{\rm NL}\|_{L_2}^2\frac{n-k}{n}\left(1 + \frac{1}{n}\operatorname{tr}(\boldsymbol{M}_m\widehat{\boldsymbol{\Sigma}}_m\boldsymbol{M}_m\widehat{\boldsymbol{\Sigma}}_{m^c})\right)\\ &\simeq \|f_{\rm NL}\|_{L_2}^2\frac{\phi}{\psi}\lambda^2 v_p(-\lambda;\psi)^2(1+\widetilde{v}_p(-\lambda;\varphi_{m\ell},\psi)) + \frac{\psi-\phi}{\psi}\left(1 + \frac{1}{k}\operatorname{tr}(\boldsymbol{M}_m\widehat{\boldsymbol{\Sigma}}_m\boldsymbol{M}_m\boldsymbol{\Sigma})\right)\\ &\simeq \|f_{\rm NL}\|_{L_2}^2\frac{\phi}{\psi}\lambda^2 v_p(-\lambda;\psi)^2(1+\widetilde{v}_p(-\lambda;\varphi_{m\ell},\psi)) + \frac{\psi-\phi}{\psi}(1+\widetilde{v}_p(-\lambda;\psi,\psi))\\ &\simeq \|f_{\rm NL}\|_{L_2}^2\left(\frac{\psi-\phi}{\psi} + \frac{\phi}{\psi}\lambda^2 v_p(-\lambda;\psi)^2\right)(1+\widetilde{v}_p(-\lambda;\varphi_{m\ell},\psi)).\end{aligned}$$

Combining the above results finish the proof for Part (3). $\square$

Lemma 32 (Quadratic concentration with uncorrelated components)**.** Let $\boldsymbol{X} \in \mathbb{R}^{n\times p}$ be the design matrix whose rows $\boldsymbol{x}_i$'s are independent samples drawn according to Assumption F. Let $\boldsymbol{f} \in \mathbb{R}^p$ be a random vector with i.i.d. entries f_i's, where f_i has bounded L_2 norm and is uncorrelated with $\boldsymbol{x}_i$. Let $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ be a symmetric matrix such that $\limsup \|\boldsymbol{A}\|_{\mathrm{op}} \leq M_0$ almost surely as $p \to \infty$ for some constant $M_0 < \infty$. Then as $n, p \to \infty$ such that $p/n \to \phi \in (0, \infty)$, it holds that

$$\frac{1}{n}\boldsymbol{f}^\top \boldsymbol{X}\boldsymbol{A}\boldsymbol{X}^\top \boldsymbol{f} \simeq \frac{1}{n}\|f_{\text{NL}}\|_{L_2}^2 \operatorname{tr}(\boldsymbol{X}\boldsymbol{A}\boldsymbol{X}^\top). \tag{73}$$

Proof. Let $\widehat{\boldsymbol{\Sigma}} = \boldsymbol{X}^\top\boldsymbol{X}/n$ and $\boldsymbol{M} = (\widehat{\boldsymbol{\Sigma}} + \lambda\boldsymbol{I}_p)^{-1}$ be the resolvent. Note that

$$\frac{1}{n}\boldsymbol{f}^\top \boldsymbol{X}\boldsymbol{A}\boldsymbol{X}^\top \boldsymbol{f} = \frac{1}{n}\boldsymbol{f}^\top \boldsymbol{X}\boldsymbol{M}(\boldsymbol{M}^{-1}\boldsymbol{A}\boldsymbol{M}^{-1})\boldsymbol{M}\boldsymbol{X}^\top \boldsymbol{f} \tag{74}$$

Since $\boldsymbol{X} = \boldsymbol{Z}\boldsymbol{\Sigma}^{1/2}$, we have $\boldsymbol{M}\boldsymbol{X}^\top = \boldsymbol{\Sigma}^{-1/2}(\boldsymbol{Z}^\top\boldsymbol{Z}/n + \lambda\boldsymbol{\Sigma}^{-1})^{-1}\boldsymbol{Z}^\top = \boldsymbol{\Sigma}^{1/2}\boldsymbol{Z}^\top(\boldsymbol{Z}\boldsymbol{\Sigma}\boldsymbol{Z}^\top/n + \lambda\boldsymbol{I}_p)^{-1}$. Let $\boldsymbol{B}_1 = (\boldsymbol{Z}\boldsymbol{\Sigma}\boldsymbol{Z}^\top/n + \lambda\boldsymbol{I}_p)^{-1}$ and $\boldsymbol{B}_2 = \boldsymbol{Z}\boldsymbol{\Sigma}^{1/2}\boldsymbol{M}^{-1}\boldsymbol{A}\boldsymbol{M}^{-1}\boldsymbol{\Sigma}^{1/2}\boldsymbol{Z}^\top/n$, then (74) becomes

$$\frac{1}{n}\boldsymbol{f}^\top \boldsymbol{X}\boldsymbol{A}\boldsymbol{X}^\top \boldsymbol{f} = \boldsymbol{f}^\top\boldsymbol{B}_1\boldsymbol{B}_2\boldsymbol{B}_1\boldsymbol{f}. \tag{75}$$

Next, we adapt the idea of Bartlett et al. (2021, Lemma A.16), to show the diagonal concentration and trace concentration successively.

Diagonal concentration. From a matrix identity in Patil and Du (2023, Lemma D.4), we have that, for any $t > 0$,

$$\boldsymbol{B}_1^{-1}\boldsymbol{B}_2\boldsymbol{B}_1^{-1} = \frac{1}{t}(\boldsymbol{B}_1^{-1} - (\boldsymbol{B}_1 + t\boldsymbol{B}_2)^{-1}) + t\boldsymbol{B}_1^{-1}\boldsymbol{B}_2(\boldsymbol{B}_1 + t\boldsymbol{B}_2)^{-1}\boldsymbol{B}_2\boldsymbol{B}_1^{-1}.$$

Let $\boldsymbol{U} \in \mathbb{R}^{n\times n}$ with $U_{ij} = [\boldsymbol{f}]_i[\boldsymbol{f}]_j\,\mathbb{1}\{i \neq j\}$. We then have

$$\begin{aligned}
&\left|\sum_{1\leq i\neq j\leq n}[\boldsymbol{B}_1^{-1}\boldsymbol{B}_2\boldsymbol{B}_1^{-1}]_{ij}[\boldsymbol{f}]_i[\boldsymbol{f}]_j\right| \\
&= |\langle \boldsymbol{B}_1^{-1}\boldsymbol{B}_2\boldsymbol{B}_1^{-1}, \boldsymbol{U}\rangle| \\
&\leq \frac{1}{t}|\langle \boldsymbol{B}_1^{-1}, \boldsymbol{U}\rangle| - \frac{1}{t}|\langle (\boldsymbol{B}_1 + t\boldsymbol{B}_2)^{-1}, \boldsymbol{U}\rangle| + t\|\boldsymbol{B}_1^{-1}\|_{\mathrm{op}}^2\|\boldsymbol{B}_2\|_{\mathrm{op}}^2\|(\boldsymbol{B}_1 + t\boldsymbol{B}_2)^{-1}\|_{\mathrm{op}}\|\boldsymbol{U}\|_{\mathrm{tr}}.
\end{aligned} \tag{76}$$

For the first two terms, Patil and Du (2023, Lemma D.3) implies that

$$\frac{1}{n}|\langle \boldsymbol{B}_1^{-1}, \boldsymbol{U}\rangle| \xrightarrow{\text{a.s.}} 0, \quad \text{and} \quad \frac{1}{n}|\langle (\boldsymbol{B}_1 + t\boldsymbol{B}_2)^{-1}, \boldsymbol{U}\rangle| \xrightarrow{\text{a.s.}} 0,$$

where the second convergence is due to $|\langle (\boldsymbol{B}_1 + t\boldsymbol{B}_2)^{-1}, \boldsymbol{U}\rangle| \lesssim t|\langle \boldsymbol{B}_1^{-1}, \boldsymbol{U}\rangle|$ because of Woodbury matrix identity and the bounded spectrums of $\boldsymbol{B}_1$ and $\boldsymbol{B}_2$. For the last term, note that $\|\boldsymbol{B}_1\|_{\mathrm{op}} \leq \lambda^{-1}$ and $\|\boldsymbol{B}_2\|_{\mathrm{op}} \leq \|\boldsymbol{A}\|_{\mathrm{op}}\|\widehat{\boldsymbol{\Sigma}}\|_{\mathrm{op}}$, where $\|\boldsymbol{A}\|_{\mathrm{op}}$ is almost surely bounded as assumed, and $\|\boldsymbol{Z}\boldsymbol{\Sigma}^{1/2}\|_{\mathrm{op}}^2/n = \|\widehat{\boldsymbol{\Sigma}}\|_{\mathrm{op}} \leq r_{\max}(1 + \sqrt{\phi})^2$ almost surely as $n, p \to \infty$ and $p/n \to \phi \in (0, \infty)$ (see, e.g., Bai and Silverstein (2010)). Also, $\|\boldsymbol{U}\|_{\mathrm{tr}}/n \leq 2\|\boldsymbol{f}\|_2^2/n \xrightarrow{\text{a.s.}} 2\|f_{\text{NL}}\|_{L_2}^2 < \infty$ from the strong law of large numbers. Thus, the last term is almost surely bounded. By choosing $t = \sqrt{|\langle \boldsymbol{B}_1^{-1}, \boldsymbol{U}\rangle|/\|\boldsymbol{U}\|_{\mathrm{tr}}}$, it then follows that $n^{-1}|\langle \boldsymbol{B}_1^{-1}\boldsymbol{B}_2\boldsymbol{B}_1^{-1}, \boldsymbol{U}\rangle| \xrightarrow{\text{a.s.}} 0$. Therefore,

$$\left|\frac{1}{n}\boldsymbol{f}^\top\boldsymbol{B}_1^{-1}\boldsymbol{B}_2\boldsymbol{B}_1^{-1}\boldsymbol{f} - \frac{1}{n}\sum_{i=1}^n[\boldsymbol{B}_1^{-1}\boldsymbol{B}_2\boldsymbol{B}_1^{-1}]_{ii}[\boldsymbol{f}]_i^2\right| \xrightarrow{\text{a.s.}} 0.$$

Trace concentration. From the results in Knowles and Yin (2017), it holds that

$$\max_{1\le i\le n}\left|[\boldsymbol{B}_1^{-1}\boldsymbol{B}_2\boldsymbol{B}_1^{-1}]_{ii} - \frac{1}{n}\operatorname{tr}[\boldsymbol{B}_1^{-1}\boldsymbol{B}_2\boldsymbol{B}_1^{-1}]\right| \xrightarrow{\text{a.s.}} 0.$$

Further, since $n^{-1}\|\boldsymbol{f}\|_2^2 \xrightarrow{\text{a.s.}} \|f_{\text{NL}}\|_{L^2}^2$, we have

$$\frac{1}{n}|\boldsymbol{f}^\top\boldsymbol{B}_1^{-1}\boldsymbol{B}_2\boldsymbol{B}_1^{-1}\boldsymbol{f} - \operatorname{tr}[\boldsymbol{B}_1^{-1}\boldsymbol{B}_2\boldsymbol{B}_1^{-1}]\|f_{\text{NL}}\|_{L^2}^2| \xrightarrow{\text{a.s.}} 0,$$

which finishes the proof. $\square$

C. Additional numerical illustrations in Section 4

C.1. Comparison of intermediate ovlp- and full-estimators for elastic net and lasso

<u>Elastic net:</u>

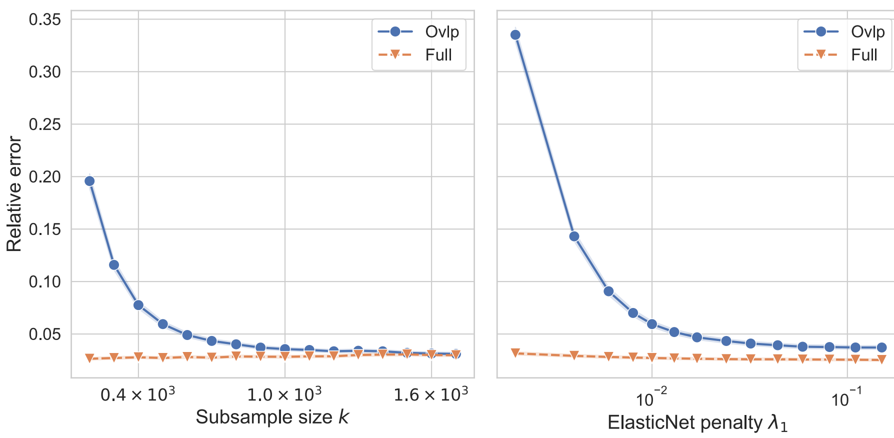

Fig. 9: **Full-estimator performs better than ovlp-estimator across different subsample sizes.** The relative errors of the "ovlp" and "full" estimators for ridge ensemble with ensemble size $M = 10$. The left panel shows the results with ridge penalty $\lambda = 1$ and varying subsample size k. The right panel shows the results with subsample size $k = 300$ and varying ridge penalty λ. The data generating process is the same as in Figure 1.

<u>Lasso:</u>

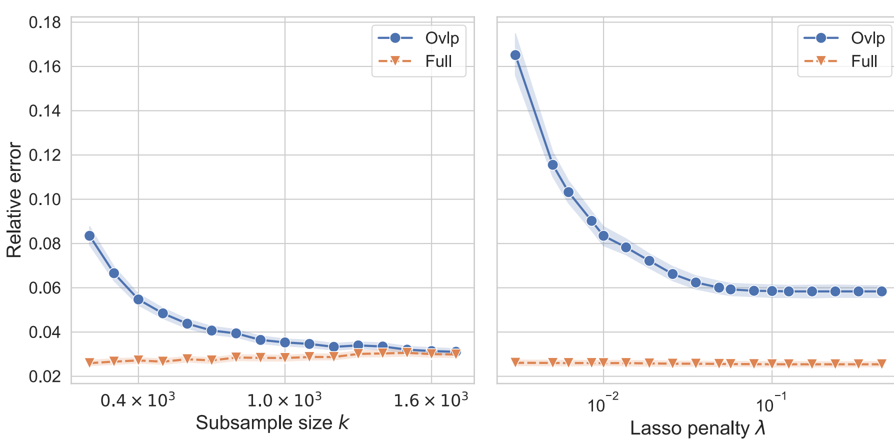

Fig. 10: **Full-estimator performs better than ovlp-estimator across different subsample sizes.** The relative errors of the "ovlp" and "full" estimators for ridge ensemble with ensemble size $M = 10$. The left panel shows the results with ridge penalty $\lambda = 1$ and varying subsample size k. The right panel shows the results with subsample size $k = 300$ and varying ridge penalty λ. The data generating process is the same as in Figure 1.

C.2. Comparison of ovlp- and full-CGCV for ridge, elastic net, and lasso

Ridge:

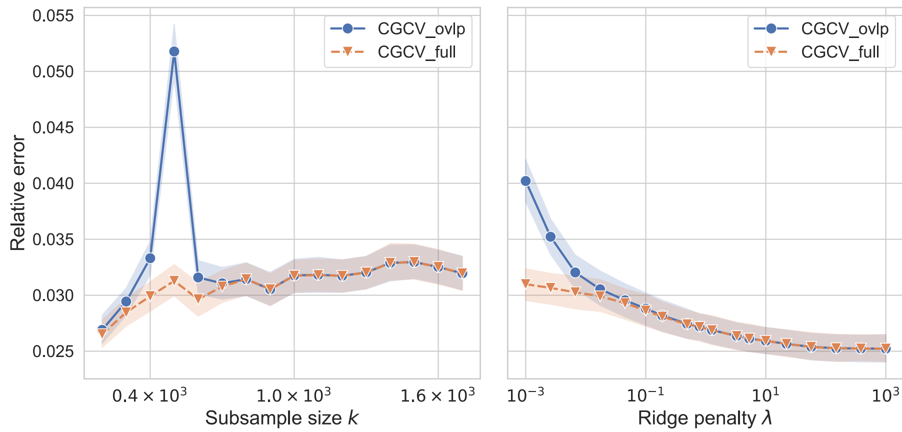

Fig. 11: **The CGCV_full estimator ($\widetilde{R}_M^{\mathbf{cgcv,full}}$) performs better than CGCV_ovlp estimator ($\widetilde{R}_M^{\mathbf{cgcv,ovlp}}$).** Plots of relative errors of the $\widetilde{R}_M^{\mathrm{cgcv,full}}$ and $\widetilde{R}_M^{\mathrm{cgcv,ovlp}}$ for ridge ensemble with ensemble size $M = 10$. The left panel shows the results with ridge penalty $\lambda = 0.001$ and varying subsample size k. The right panel shows the results with subsample size $k = 500$ and varying ridge penalty λ. The data generating process is the same as in Figure 1.

Elastic net:

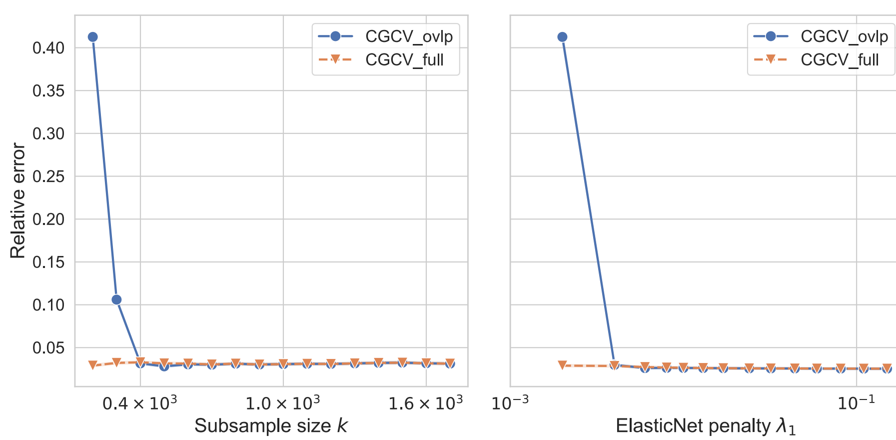

Fig. 12: **The CGCV_full estimator ($\widetilde{R}_M^{\mathbf{cgcv,full}}$) performs better than CGCV_ovlp estimator ($\widetilde{R}_M^{\mathbf{cgcv,ovlp}}$).** Plots of relative errors of the $\widetilde{R}_M^{\mathrm{cgcv,full}}$ and $\widetilde{R}_M^{\mathrm{cgcv,ovlp}}$ for elastic net ensemble with ensemble size $M = 10$ and $\lambda_2 = 0.01$. The left panel shows the results with elastic net penalty $\lambda_1 = 0.002$ and varying subsample size k. The right panel shows the results with subsample size $k = 200$ and varying elastic net penalty λ_1. The data generating process is the same as in Figure 1.

Lasso:

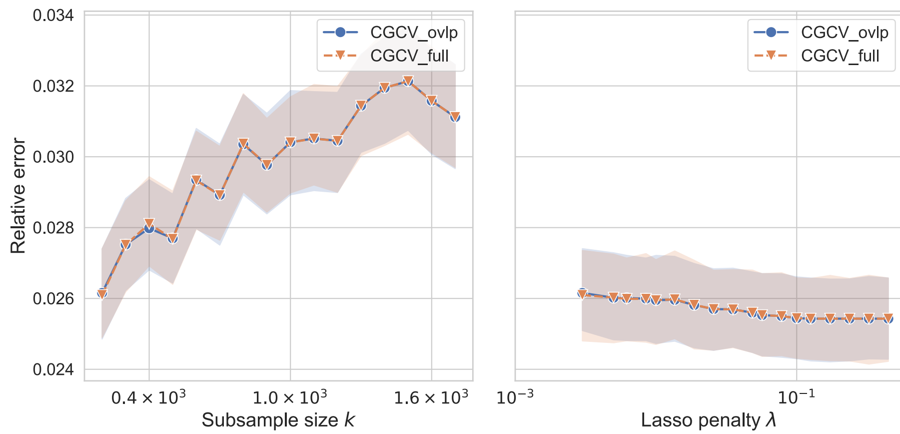

Fig. 13: **The CGCV_full estimator ($\widetilde{R}_M^{\mathbf{cgcv,full}}$) performs similar to CGCV_ovlp estimator ($\widetilde{R}_M^{\mathbf{cgcv,ovlp}}$).** Plots of relative errors of the $\widetilde{R}_M^{\mathrm{cgcv,full}}$ and $\widetilde{R}_M^{\mathrm{cgcv,ovlp}}$ for lasso ensemble with ensemble size $M = 10$. The left panel shows the results with lasso penalty $\lambda = 0.003$ and varying subsample size k. The right panel shows the results with subsample size $k = 200$ and varying lasso penalty λ. The data generating process is the same as in Figure 1.

C.3. Comparison of CGCV and GCV in k for ridge, elastic net, and lasso

Ridge:

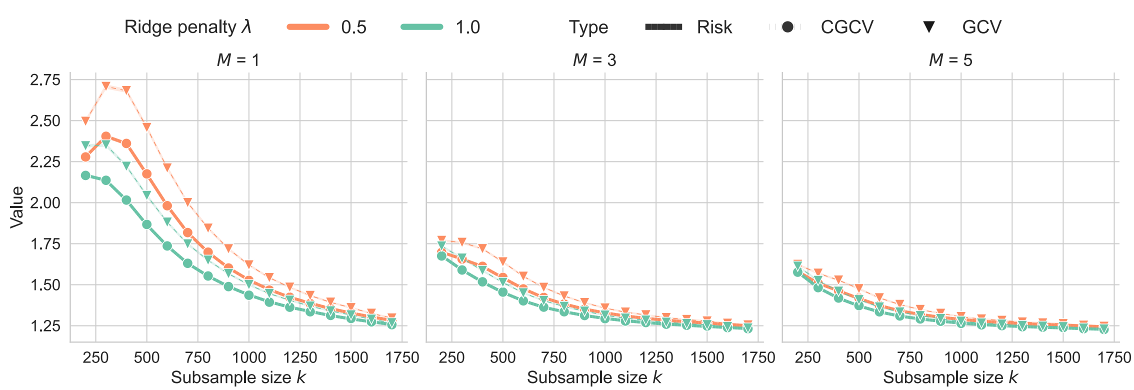

Fig. 14: **CGCV is consistent for finite ensembles across different subsample sizes.** Plot of risks versus the subsample size k for the ridge ensemble with different λ and M. The data generating process is the same as in Figure 1.

Elastic Net:

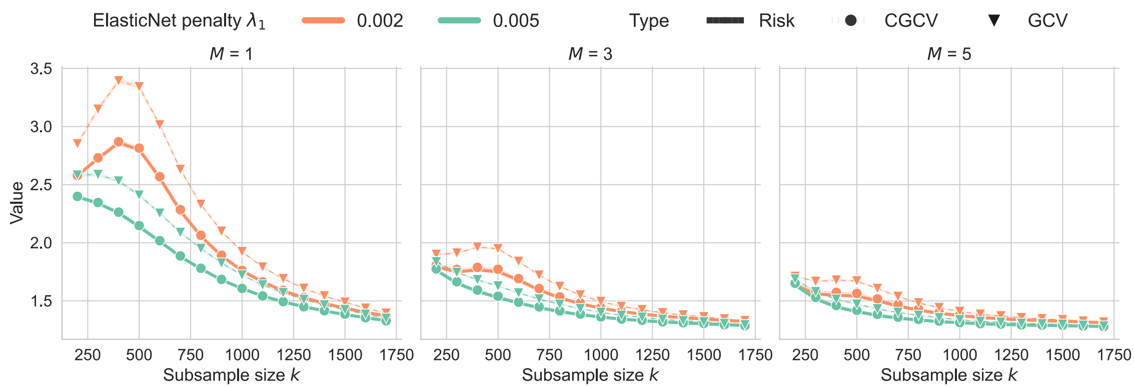

Fig. 15: **CGCV is consistent for finite ensembles across different subsample sizes.** Plot of risks versus the subsample size k for elastic net ensemble with different λ and M.

Lasso:

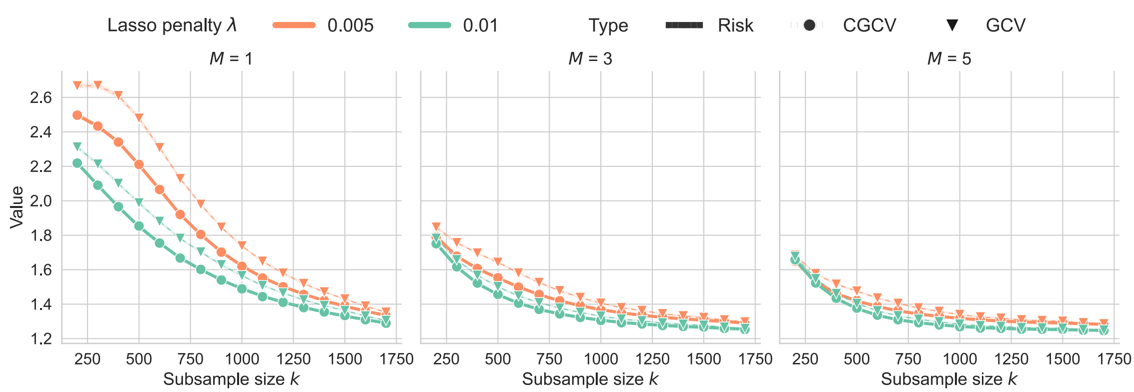

Fig. 16: **CGCV is consistent for finite ensembles across different subsample sizes.** Plot of risks versus the subsample size k for elastic net ensemble with different λ and M.

C.4. Comparison of GCV and GCV in M for ridge, elastic net, and lasso

Ridge:

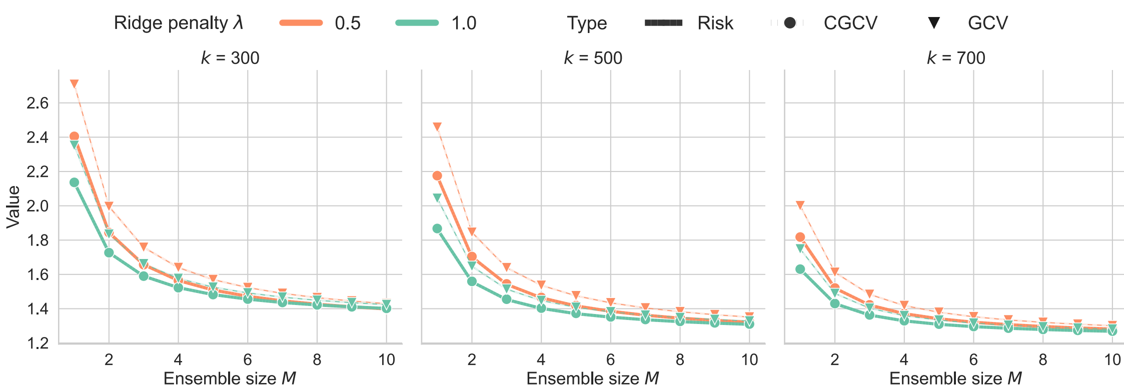

Fig. 17: **GCV gets closer to the risk as ensemble size increases across different subsample sizes.** Plot of risks versus the subsample size k for the lasso ensemble with different λ and M. The data generating process is the same as in Figure 1.

Elastic Net:

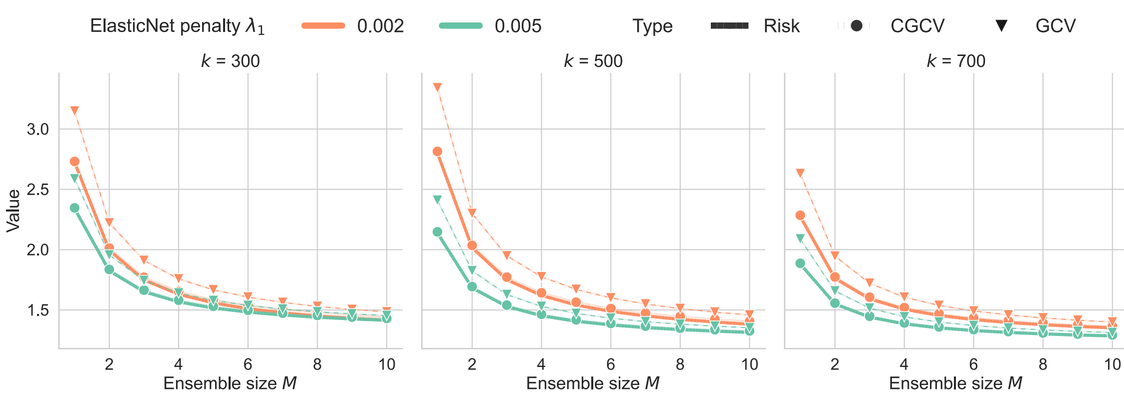

Fig. 18: **GCV gets closer to the risk as ensemble size increases across different subsample sizes.** Plot of risks versus the ensemble size M for ridge ensemble with different λ and k.

Lasso:

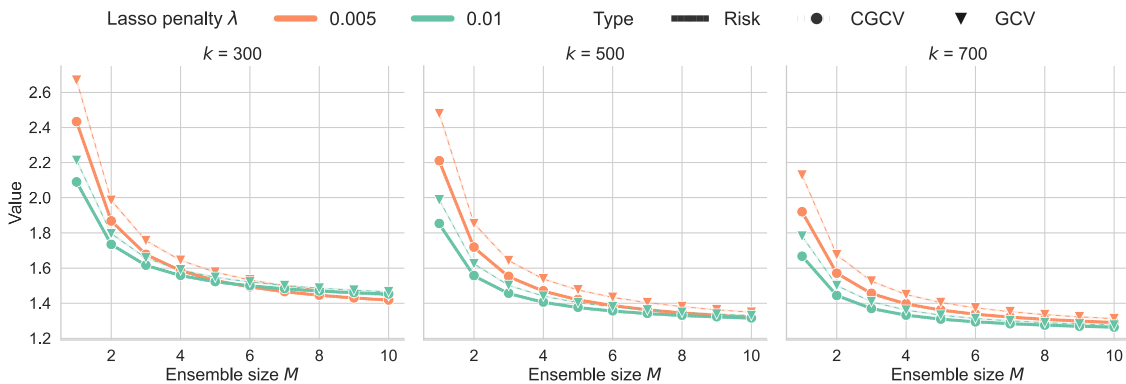

Fig. 19: **GCV gets closer to the risk as ensemble size increases across different subsample sizes.** Plot of risks versus the ensemble size M for ridge ensemble with different λ and k.

C.5. Comparison of CGCV and GCV in λ for ridge, elastic net and lasso

Ridge:

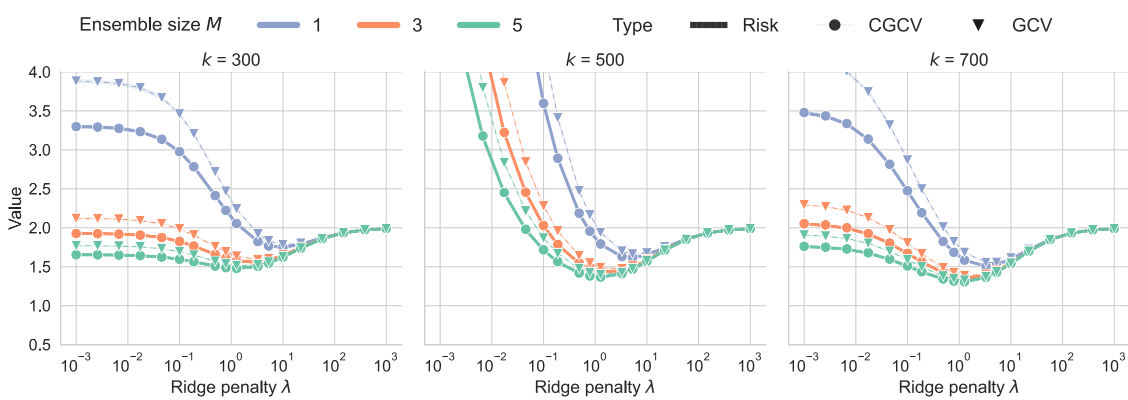

Fig. 20: **CGCV is consistent for finite ensembles across different regularization levels.** Plot of risks versus the ensemble size M for ridge ensemble with different λ and k.

Elastic Net:

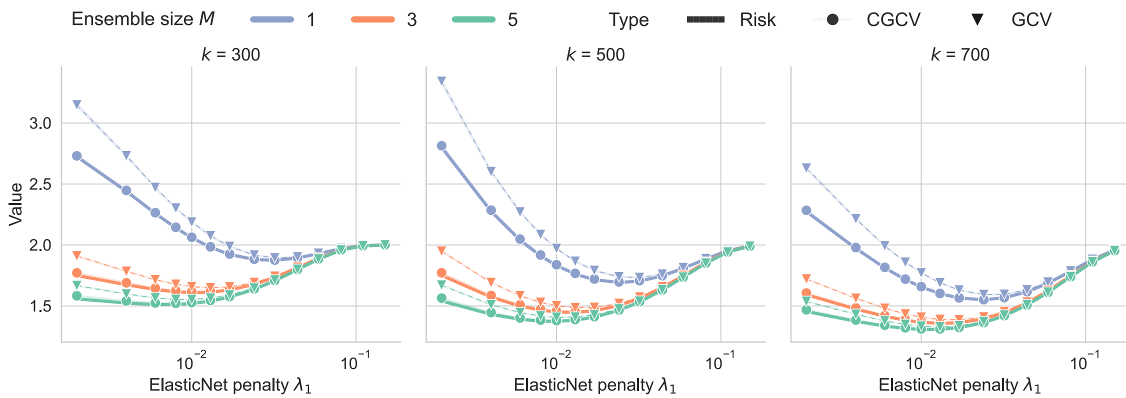

Fig. 21: **GCV gets closer to the risk as ensemble size increases across different subsample sizes.** Plot of risks versus the ensemble size M for elastic net ensemble with different λ and k.

Lasso:

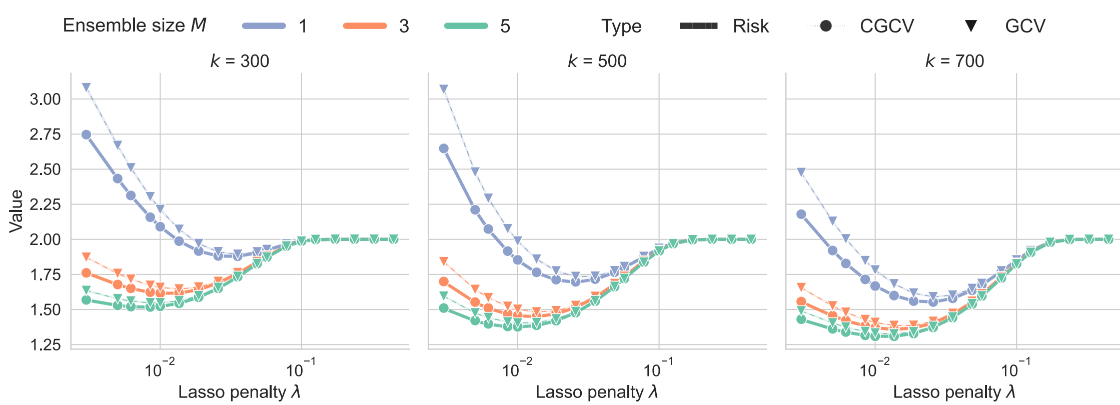

Fig. 22: **GCV gets closer to the risk as ensemble size increases across different subsample sizes.** Plot of risks versus the ensemble size M for lasso ensemble with different λ and k.

C.6. Numerical simulations with large ensemble size M

Recall that Section 4.4 presents simulation results when M is at most 10. In this section, we provide simulation results in the same setting as described in Section 4.4, but with M as large as 100. The goal is to illustrate that our proposed CGCV estimator is efficient not only for small M but also for large M. Since the bias of the GCV is decreasing with M, we expect the GCV works well for large M. It is evident from Figure 23 that the CGCV accurately estimates the actual risk for all values of M, while the bias of GCV diminishes only when M is large.

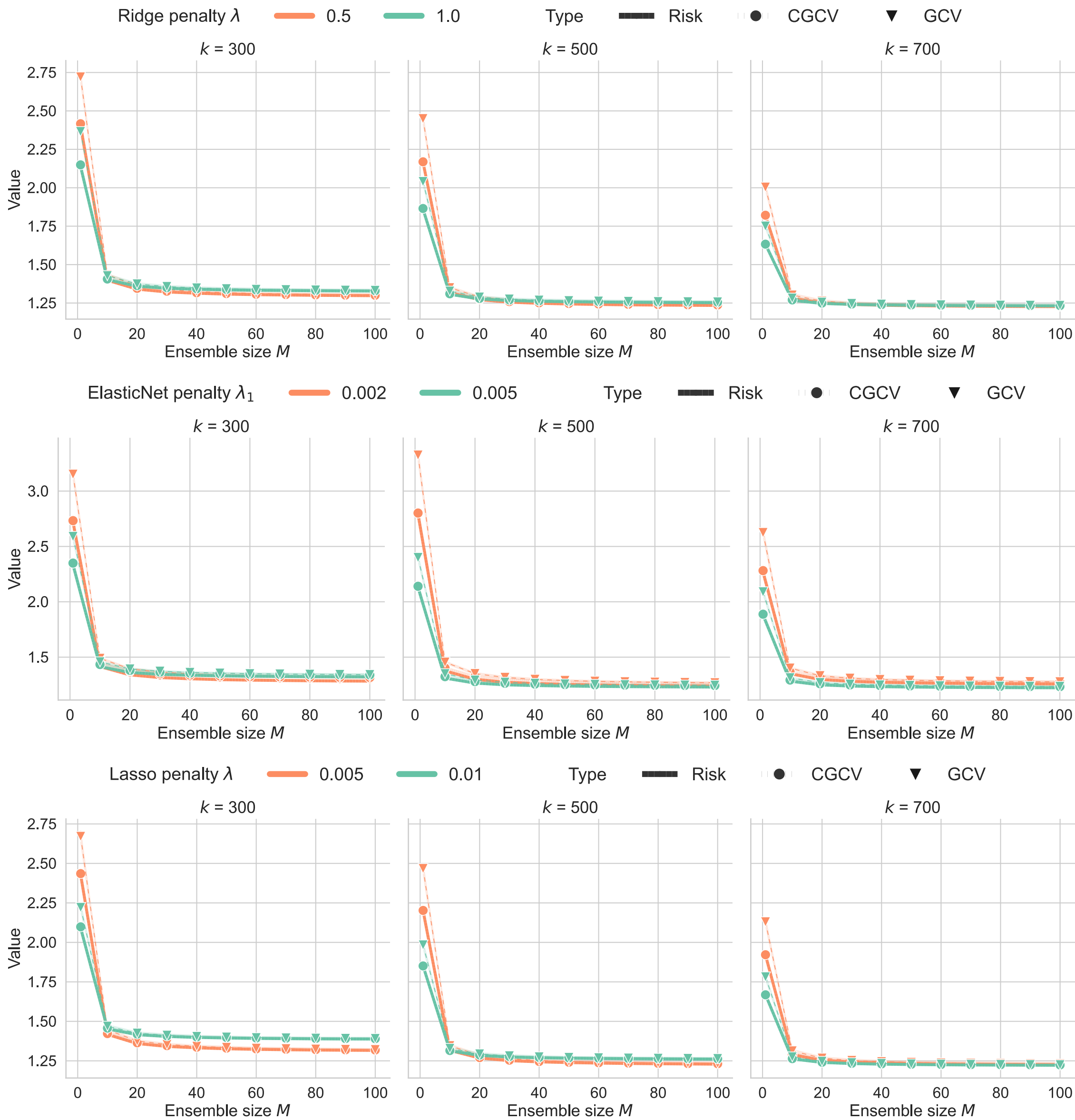

Fig. 23: **CGCV efficiently estimates the actual risk for all values of M. GCV gets closer to the actual risk as M increases.** Top row: Ridge; Middle row: Elastic Net; Bottom row: Lasso. The data generating process is the same as in Figure 1.

C.7. Illustration of inconsistency for GCV variant

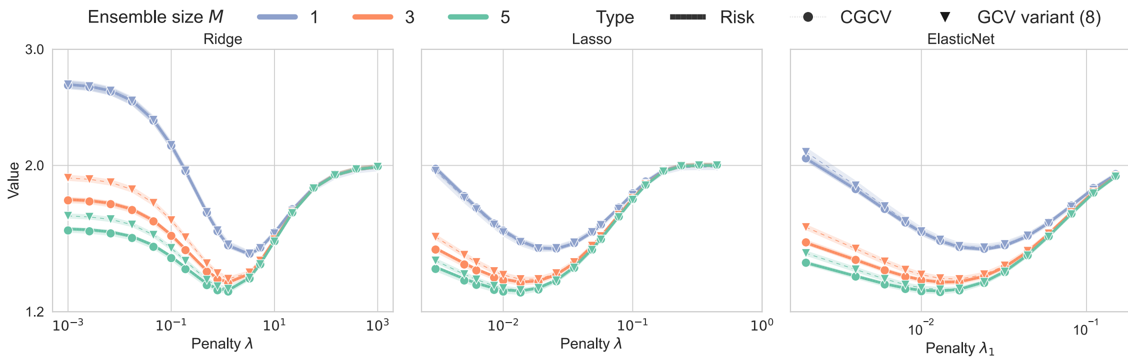

Fig. 24: **The GCV variant** (8) **is not consistent for finite** $M > 1$**.** Plot of risks versus the regularization parameter for ridge, lasso, and elastic net ensembles, under different M and fixed $k = 800$. The data generating process is the same as in Figure 1.

D. Additional numerical illustrations for Section 4 with non-Gaussian data

In this section, we conduct the same numerical experiments as in Section 4 except that here the design matrix is not Gaussian distributed. We consider two non-Gaussian distributions: the Rademacher distribution and the uniform distribution in $[-\sqrt{3}, \sqrt{3}]$. The design matrix $\boldsymbol{X}$ is then generated by a scale transformation to ensure the covariance matrix is the same as $\boldsymbol{\Sigma}$ in Section 4.4. In the following subsections, we present counterpart plots to Figures 1 and 3–5 in the same setting, except here $\boldsymbol{X}$ is non-Gaussian distributed.

D.1. Results for Rademacher distribution

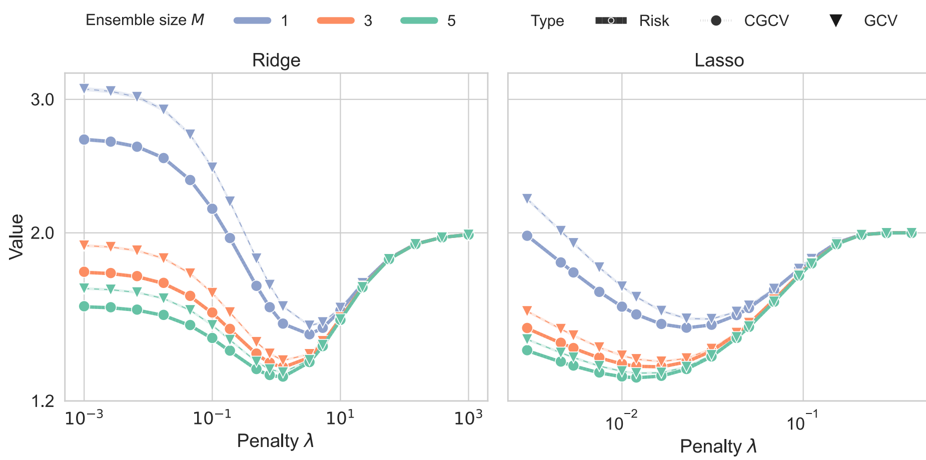

Fig. 25: **CGCV is consistent for finite ensembles of penalized estimators while GCV is not.** The simulation setting is the same as Figure 1, except that here $\boldsymbol{X}$ is Rademacher distributed.

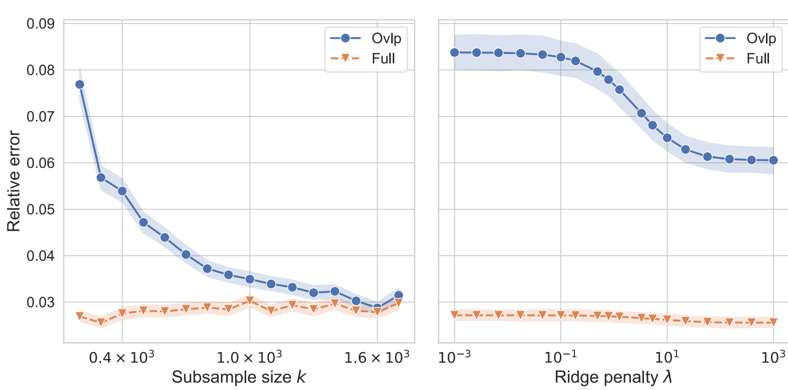

Fig. 26: **Full-estimator $\widetilde{R}_M^{\mathbf{full}}$ performs better than ovlp-estimator $\widetilde{R}_M^{\mathbf{ovlp}}$ across different regularization and for small subsample size** k**.** The simulation setting is the same as Figure 3, except that here $\boldsymbol{X}$ is Rademacher distributed.

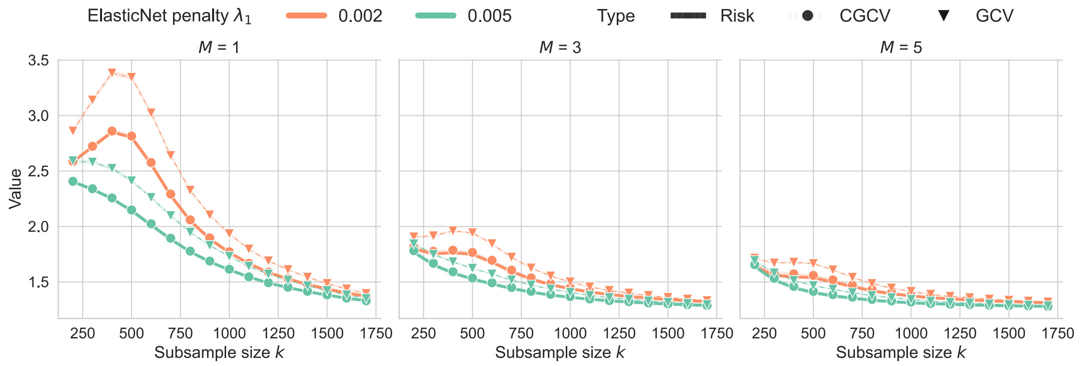

Fig. 27: **CGCV is consistent for finite ensembles across different subsample sizes.** The plot of risks versus the subsample size k for Elastic Net ensemble with different λ and M and fixed $\lambda_2 = 0.01$. The simulation setting is the same as Figure 4, except that here $\boldsymbol{X}$ is Rademacher distributed.

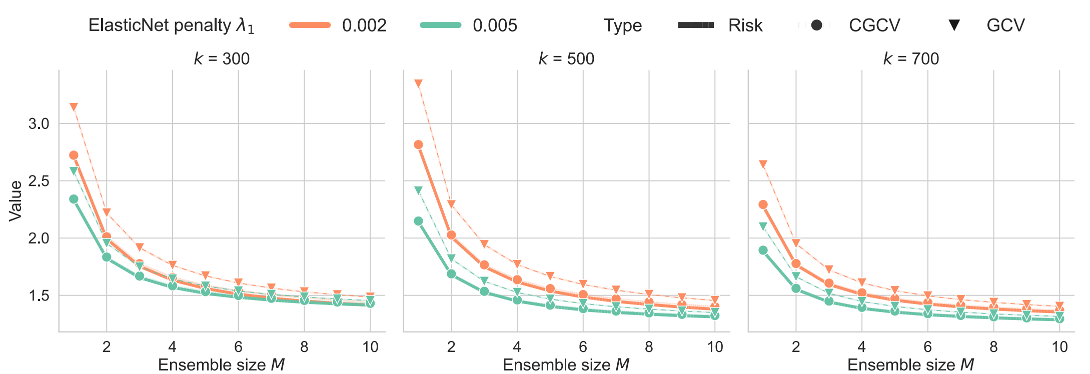

Fig. 28: **GCV gets closer to the risk as ensemble size increases across different subsample sizes.** The plot of risks versus the Elastic Net ensemble size M for ridge ensemble with different λ and k and fixed $\lambda_2 = 0.01$. The simulation setting is the same as Figure 5, except that here $\boldsymbol{X}$ is Rademacher distributed.

D.2. Results for uniform distribution

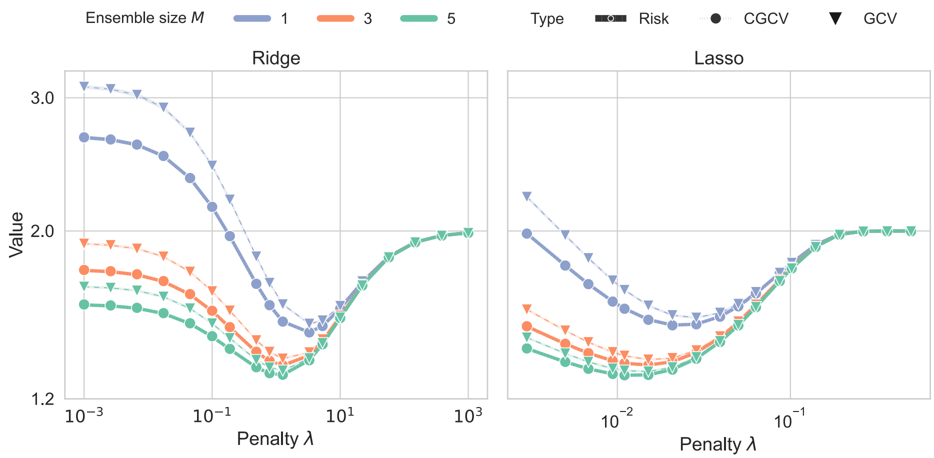

Fig. 29: **CGCV is consistent for finite ensembles of penalized estimators while GCV is not.** The simulation setting is the same as Figure 1, except that here $\boldsymbol{X}$ is uniformly distributed.

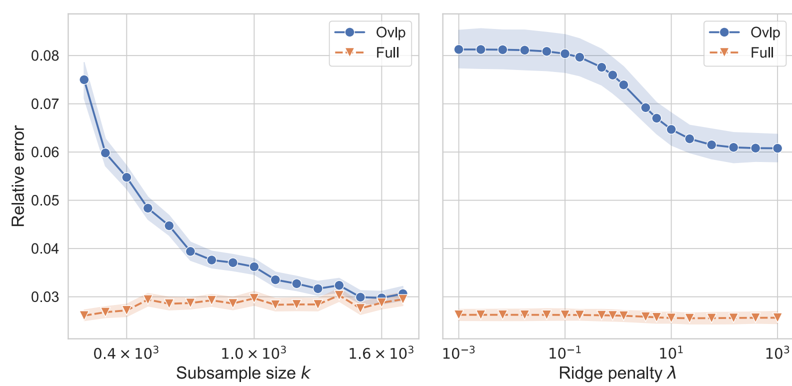

Fig. 30: **Full-estimator $\widetilde{R}_M^{\mathbf{full}}$ performs better than ovlp-estimator $\widetilde{R}_M^{\mathbf{ovlp}}$ across different regularization and for small subsample size** k**.** The simulation setting is the same as Figure 3, except that here $\boldsymbol{X}$ is uniformly distributed.

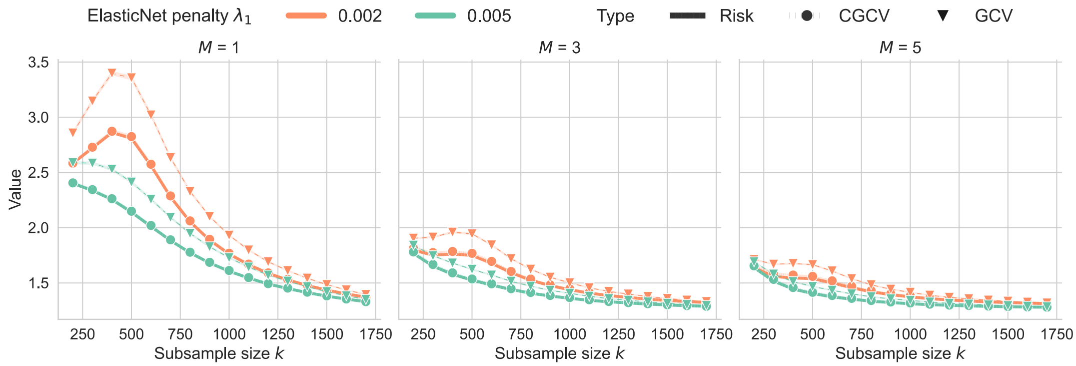

Fig. 31: **CGCV is consistent for finite ensembles across different subsample sizes.** The plot of risks versus the subsample size k for Elastic Net ensemble with different λ and M and fixed $\lambda_2 = 0.01$. The simulation setting is the same as Figure 4, except that here $\boldsymbol{X}$ is uniformly distributed.

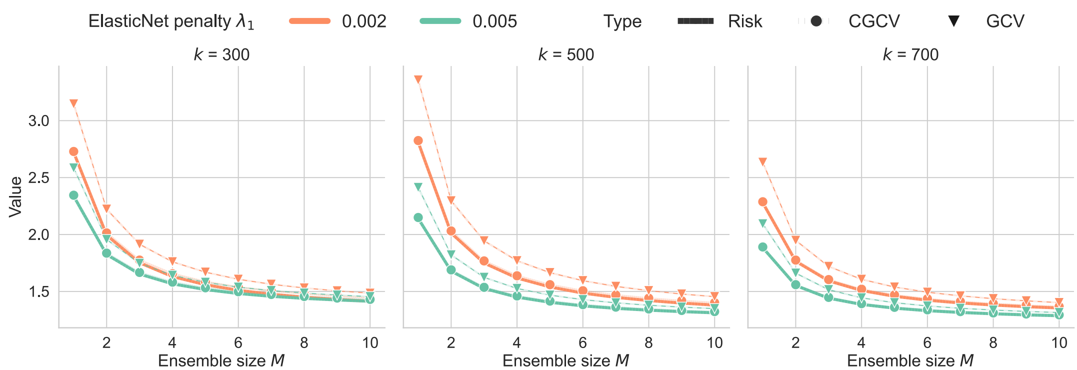

Fig. 32: **GCV gets closer to the risk as ensemble size increases across different subsample sizes.** The plot of risks versus the Elastic Net ensemble size M for ridge ensemble with different λ and k and fixed $\lambda_2 = 0.01$. The simulation setting is the same as Figure 5, except that here $\boldsymbol{X}$ is uniformly distributed.

E. Additional numerical illustrations for Section 5 with non-Gaussian data

E.1. Comparison of CGCV and GCV in λ for elastic net and lasso

Elastic net:

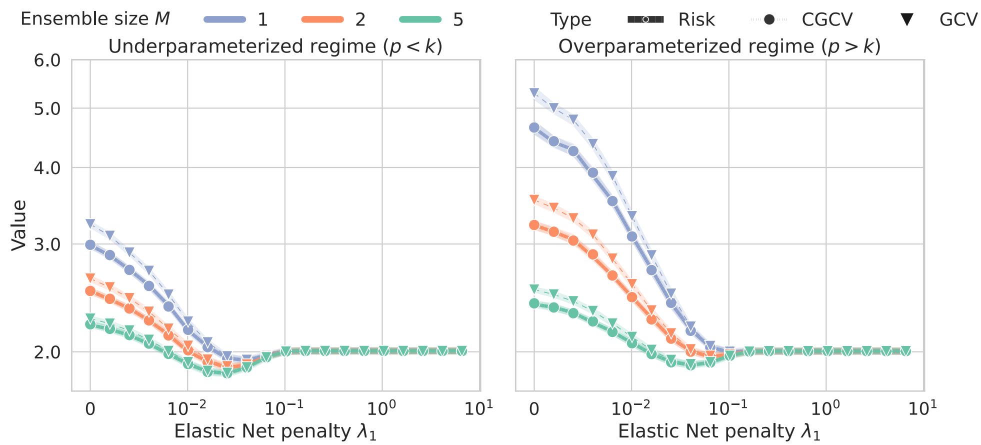

Fig. 33: **CGCV is consistent for finite ensembles across different regularization levels.** The GCV estimates for the Elastic Net ensemble with varying lasso penalty λ_1 and fixed $\lambda_2 = 0.01$ in a problem setup of Section 5.4 over 50 repetitions of the datasets. Here the feature dimension is $p = 1200$. The left and the right panels show the cases when the subsample sizes are $k = 2400$ and $k = 800$, respectively.

Lasso:

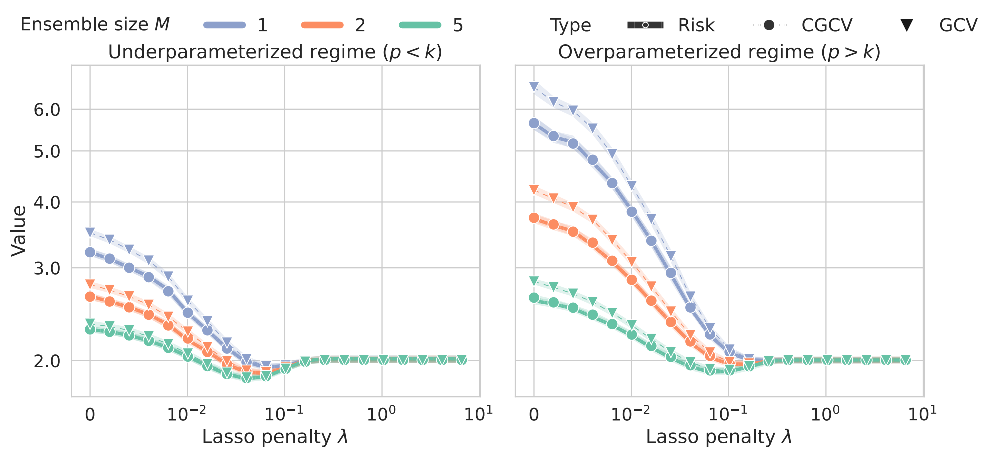

Fig. 34: **CGCV is consistent for finite ensembles across different regularization levels.** The GCV estimates for the lasso ensemble with varying lasso penalty λ in a problem setup of Section 5.4 over 50 repetitions of the datasets. Here the feature dimension is $p = 1200$. The left and the right panels show the cases when the subsample sizes are $k = 2400$ and $k = 800$, respectively.