Corrected generalized cross-validation
for finite ensembles of penalized estimators

Pierre C. Bellec[†]
pierre.bellec@rutgers.edu

Jin-Hong Du[‡§]
jinhongd@andrew.cmu.edu

Takuya Koriyama[*†]
takuya.koriyama@rutgers.edu

Pratik Patil[*¶]
pratikpatil@berkeley.edu

Kai Tan[†]
kai.tan@rutgers.edu

Abstract

Generalized cross-validation (GCV) is a widely-used method for estimating the squared out-of-sample prediction risk that employs a scalar degrees of freedom adjustment (in a multiplicative sense) to the squared training error. In this paper, we examine the consistency of GCV for estimating the prediction risk of arbitrary ensembles of penalized least-squares estimators. We show that GCV is inconsistent for any finite ensemble of size greater than one. Towards repairing this shortcoming, we identify a correction that involves an additional scalar correction (in an additive sense) based on degrees of freedom adjusted training errors from each ensemble component. The proposed estimator (termed CGCV) maintains the computational advantages of GCV and requires neither sample splitting, model refitting, or out-of-bag risk estimation. The estimator stems from a finer inspection of the ensemble risk decomposition and two intermediate risk estimators for the components in this decomposition. We provide a non-asymptotic analysis of the CGCV and the two intermediate risk estimators for ensembles of convex penalized estimators under Gaussian features and a linear response model. Furthermore, in the special case of ridge regression, we extend the analysis to general feature and response distributions using random matrix theory, which establishes model-free uniform consistency of CGCV.

1 Introduction

Ensemble methods are an important part of the toolkit in current machine learning practice. These methods combine multiple models and improve predictive accuracy and stability compared to any single component model (Hastie et al., 2009; Dietterich, 1998, 2000). Bagging and its variants form an important class of ensemble methods that average predictors fitted on different subsamples of the full dataset. The classical literature on bagging includes work by Breiman (1996); Bühlmann and Yu (2002); Friedman and Hall (2007), among others. There has been a flurry of recent work on characterizing the risk behavior of such subsampling-based ensemble methods in high dimensions; see, e.g., Loureiro et al. (2022); Adlam and Pennington (2020b); Mücke et al. (2022); Patil et al. (2023); Du et al. (2023). These works demonstrate that the risk can be significantly reduced by properly choosing the size of the ensemble. Complementary to characterizing and understanding

---

Author names sorted alphabetically. [*]Corresponding authors.

[†]Department of Statistics, Rutgers University, New Brunswick, NJ 08854, USA.

[‡]Department of Statistics and Data Science, Carnegie Mellon University, Pittsburgh, PA 15213, USA.

[§]Machine Learning Department, Carnegie Mellon University, Pittsburgh, PA 15213, USA.

[¶]Department of Statistics, University of California, Berkeley, CA 94720, USA.

the risk behavior, it is important to have a reliable method for estimating these risks when implementing ensemble methods. Accurate risk estimation provides a benchmark for comparing the performance of different ensemble configurations and provides insights into the expected performance of the combined model on unseen data. In particular, Du et al. (2023); Patil and Du (2023) demonstrate that the implicit regularization provided by ensembling amounts to ridge regularization, and optimizing over both the ensemble and subsample size helps to mitigate double descent behaviors. In this paper, our primary focus is on risk estimation techniques suitable for ensemble methods, when these ensembles are composed of penalized least-squares estimators. Several tuning parameters define an ensemble even for a single penalty function, including the regularization parameter and the subsample size. Developing consistent risk estimators informs the selection of the optimal ensemble (regularization parameters and subsample size) and leads to better generalization performance.

One widely recognized technique for risk estimation is Generalized Cross-Validation (GCV). GCV adjusts the training error multiplicatively based on a factor known as degrees of freedom correction. GCV is conceptually motivated as an approximation to leave-one-out cross-validation and provides consistent risk estimation without sample splitting and refitting; see, e.g., Chapter 7 of Hastie et al. (2009) and Chapter 5 of Wasserman (2006). The consistency of GCV has been extensively studied, initially for fixed-X design settings and linear smoothers (Golub et al., 1979; Craven and Wahba, 1979). Subsequent work has extended this to the random-X settings, which are arguably more relevant in modern machine learning applications where observations are drawn from some underlying distribution rather than being predetermined. In particular, the consistency of GCV for ridge regression has been shown in Patil et al. (2021, 2022); Wei et al. (2022) in various settings. For ridge regression, the consistency holds for arbitrary response models with only the assumption of bounded moments. Beyond ridge regression, the consistency of GCV is shown for Lasso in (Miolane and Montanari, 2021; Celentano et al., 2023) and for convex regularized estimates in (Bellec, 2023, Section 3). For linear base estimators (e.g., least squares estimator, ridge, etc.), the degrees of freedom of the ensemble estimator is simply the average of degrees of freedom for all base predictors. The first question that this paper asks is the following:

<div align="center">Is GCV still consistent for ensembles of general penalized estimators?</div>

Perhaps the simplest base predictor to study this question is ridge regression, which has emerged as a simple but powerful tool, particularly when dealing with high-dimensional data without any special sparsity structure. Recent work by Du et al. (2023) considers the problem of risk estimation for ensembles of ridge predictors. The authors analyze the use of GCV for the full ensemble ridge estimator when the number of ensemble components goes to infinity, demonstrating that GCV is consistent for the prediction risk. However, an intriguing surprise from their work is the inconsistency of GCV for any finite ensemble: The intricate correlation caused by overlapping samples in different subsampled datasets complicates risk estimation and necessitates a correction. This finding suggests a limitation of GCV and prompts the question of whether this inconsistency is special to ridge regression or is a broader issue with GCV's applicability to other models as well. And, in particular, this prompts the second key question of our paper:

<div align="center">Is there a GCV correction that is consistent for all possible ensemble sizes?</div>

Understanding these two questions forms the basis of our work. We propose a novel risk estimator that addresses the limitations of naive GCV for ensembles of penalized estimators. Focusing on convex penalized predictors, we show that for a finite ensemble size larger than one, GCV is always inconsistent and deviates from the true risk by an additive error that we characterize ex-
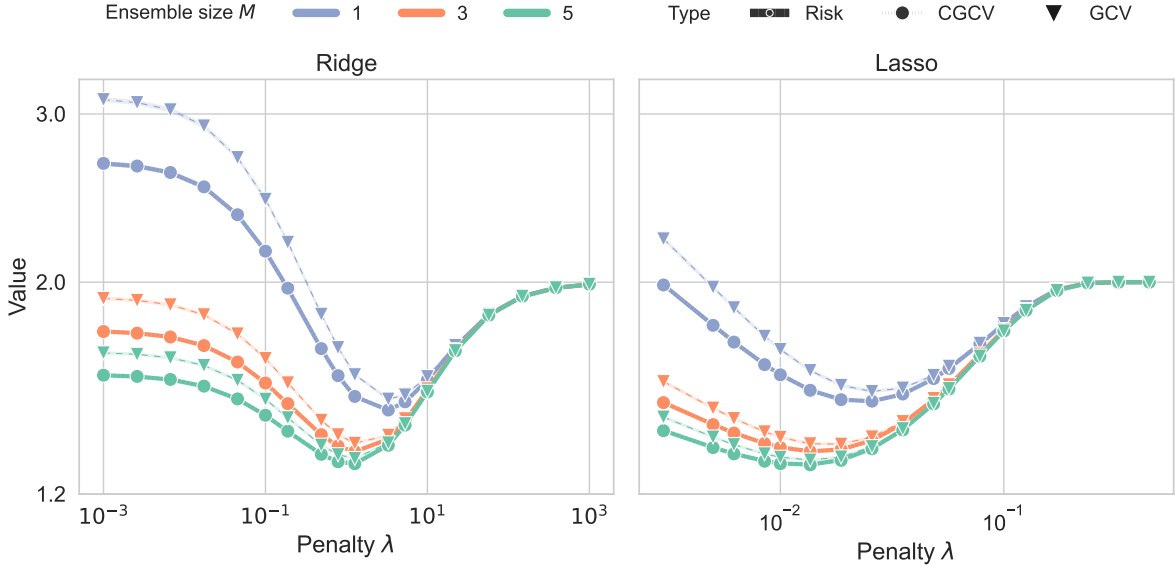
Figure 1: **CGCV is consistent for finite ensembles of penalized estimators while GCV is not.** We show a numerical comparison between the true squared risk (Risk), GCV error (GCV), and corrected GCV (CGCV) for ensembles across different tuning parameters λ and ensemble sizes M, over 1000 repetitions of the datasets. The left panel shows ridge ensembles with different λ and M. Data is generated from a linear model with Gaussian design, $(n, p) = (2000, 500)$, and a signal-to-noise ratio of 1. The subsample size is fixed at $k = 800$. The right panel shows the plot of lasso ensembles under the same setting as the left panel except for a different range of λ. Further details of the experimental setup are given in Section 4.4.

plicitly. Understanding this additive error leads to a novel correction to GCV, and we provide both non-asymptotic and asymptotic analyses of this corrected GCV under two different sets of data-generating assumptions. Our non-asymptotic analysis is restricted to Gaussian features and well-specified linear models, allowing general strongly convex penalty functions. Our asymptotic analysis is more general in accommodating random matrix theory features beyond Gaussianity and arbitrary data models while being restricted to ridge predictors. Our asymptotic analysis also enables the analysis of the "ridgeless" predictors (minimum-norm interpolators) in the overparameterized regime. For the case of ridge regression, this also allows us to show the uniform consistency in the regularization parameter. Before diving into the precise details of our contributions, we present our results in Figure 1, which demonstrates the clear gap between GCV and true risk. Our proposed corrected GCV (CGCV), on the other hand, accurately estimates the risk across varying settings for ensembles of both ridge and lasso predictors.

1.1 Summary of results

In the classical low-dimensional regime, where the feature dimension p is fixed, the excess prediction risk goes to zero as the sample size n diverges. Thus, the average training error itself already serves as a consistent estimator of the risk. In the proportional asymptotic regime, where the feature size p scales proportionally with the sample size n, the prediction risk typically converges to a non-vanishing constant (Bayati and Montanari, 2011; El Karoui, 2013; Thrampoulidis et al., 2018, among others). Therefore, the regime of primary interest for our results is the proportional asymptotic regime. More formally, we implicitly consider a sequence of regression problems, say, indexed by n. The dimension $p(n)$, the distribution of the observations, the estimators and penalty under consideration, and the subsample size $k(n)$ all implicitly depend on n with $p(n), k(n)$ growing

3

with n. We omit the dependence on n to lighten the notation and write simply, e.g., p for the dimension and k for the subsample size. Throughout the paper, the ratio p/n is bounded from above by a constant independent of n, so that by extracting a subsequence if necessary, we may assume that p/n has a finite limit. While several of our results are non-asymptotic with explicit dependence on certain problem parameters, we also state consistency results asymptotically using $\xrightarrow{\text{p}}/\xrightarrow{\text{a.s.}}$ (convergence in probability/almost sure convergence), $o_{\mathbb{P}}(\cdot)$ and $\mathcal{O}_{\mathbb{P}}(\cdot)$ notation, which are defined with respect to the aforementioned sequence of regression problems. Our contributions are four-fold, as summarized below.

1. **Inconsistency of GCV for finite ensembles.** We establish that ordinary GCV is inconsistent in estimating the squared out-of-sample prediction risk of finite ensembles with more than one penalized estimator (Theorem 3). Here, we assume strongly convex penalties that include ridge regression and elastic net, among others; and the inconsistency result of Theorem 3 applies to base estimators that have normalized degrees of freedom bounded away from 0 (see Remark 8 for details).

2. **Corrected GCV for finite ensembles.** We introduce a novel risk estimator, termed corrected GCV (CGCV), designed to estimate the prediction risk of the arbitrary ensemble of penalized estimators (Definition 1). The proposed estimator employs a scalar data-dependent correction term (in an additive sense) that incorporates degrees of freedom adjusted training errors of individual component estimators (Figure 2). Importantly, CGCV preserves all the computational advantages inherent to GCV, since it requires neither sample splitting nor model refitting nor out-of-bag risk estimation.

3. **Non-asymptotic analysis under Gaussian designs.** The structure of CGCV stems from inspecting the ensemble risk decomposition and two intermediate risk estimators that we construct for the components in this decomposition (Equations (11) and (12)). For Gaussian features and a well-specified linear model, we provide non-asymptotic bounds for these component estimators (Theorem 1). Building on these, we provide a non-asymptotic bound for the proposed corrected GCV estimator (Theorem 2). Our error bounds decrease at the $n^{-1/2}$ rate and provide explicit dependence on problem parameters such as the ensemble size M, the subsample size k, and the strong convexity parameter of the penalty (Equations (21) and (22)). The derived rates are pertinent within the scope of the proportional asymptotic regime.

4. **Asymptotic analysis under general designs.** For general feature structure (multiplicatively) composed of an arbitrary covariance matrix and independent components with bounded $4+\eta$ moments and response vector with bounded $4+\eta$ moments for some $\eta > 0$, we show that the intermediate risk estimators are asymptotically consistent for ridge ensemble (Theorem 4). We further show uniform consistency in the regularization level λ that includes the case of ridgeless regression for both intermediate estimators and, consequently, CGCV (Theorem 5 and Theorem 6). The asymptotic analysis allows us to relax the data and model assumptions while obtaining model-free uniform consistency of CGCV in the regularization parameter.

1.2 Related work

In this section, we describe related work on ensemble risk analysis and risk estimation, as well as a detailed comparison to other baseline risk estimation approaches based on sample splitting and out-of-bag risk estimation.

4

GCV for linear smoothers. Risk estimation is a crucial aspect of statistical learning that plays an important role in model evaluation and selection. Over the years, a myriad of methods have been proposed, each with its own strengths and limitations. Among these, GCV has emerged as a widely adopted technique for risk estimation. GCV is conceptually motivated as an approximation to leave-one-out cross-validation, a method that provides unbiased estimates of the true prediction error but can be computationally expensive (Hastie et al., 2009). The origins of GCV can be traced back to the work of Golub et al. (1979) and Craven and Wahba (1979), where it was studied in the context of fixed-X design settings for linear smoothers. These settings, where the predictors are considered fixed and non-random, are common in experimental designs. The consistency of GCV was subsequently investigated in a series of papers by Li (1985, 1986, 1987). More recently, the focus has shifted toward the random-X and high-dimensional setting. This change in focus is motivated by modern machine learning applications, which often deal with large datasets and where predictors are typically considered to be random variables drawn from some distribution rather than fixed values. In this context, GCV has been the subject of considerable research attention. In particular, Leeb (2008) establishes the consistency of GCV for the ordinary least-squares, and the consistency of GCV for ridge regression has been established in a number of recent studies in various data settings by Adlam and Pennington (2020a); Patil et al. (2021, 2022); Wei et al. (2022); Han and Xu (2023); Misiakiewicz and Saeed (2024), which provide different flavors of results (both asymptotic and non-asymptotic).

GCV beyond linear smoothers. Generalized cross-validation was initially defined for linear smoothers, where the sum of training errors is adjusted multiplicatively by the trace of the smoothing matrix. There is a general way of understanding this estimator, and the definition of GCV for estimators that are not linear smoothers can be extended by using degrees of freedom adjustments. The notion of degrees of freedom of linear estimate $\widehat{\boldsymbol{\beta}}$ dates back to the pioneering paper of Stein (1981). This literature established the multivariate version of Stein's formula and proposed an unbiased estimator for estimating the mean square error of a Gaussian multivariate mean, which is often called Stein's Unbiased Risk Estimate (SURE). For example, under the linear model $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta}_0 + \boldsymbol{\varepsilon}$ with Gaussian covariates $\boldsymbol{X} \in \mathbb{R}^{n \times p}$ and Gaussian noise $\boldsymbol{\varepsilon} \in \mathbb{R}^n$, the Stein's Unbiased Risk Estimate has expression $\widehat{\mathsf{SURE}} = \|\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}\|_2^2 + 2\sigma^2 \widehat{\mathsf{df}} - n\sigma^2$, where σ^2 is the noise variance, and

$$\widehat{\mathsf{df}} = \mathrm{tr}[(\partial/\partial\boldsymbol{y})\boldsymbol{X}\widehat{\boldsymbol{\beta}}] \tag{1}$$

is the degrees of freedom of the estimate $\widehat{\boldsymbol{\beta}}$. The SURE is unbiased in the sense that $\mathbb{E}[\widehat{\mathsf{SURE}}] = \mathbb{E}[\|\boldsymbol{X}\widehat{\boldsymbol{\beta}} - \boldsymbol{X}\boldsymbol{\beta}_0\|_2^2]$. This definition of degrees of freedom as the trace of the Jacobian yields a natural generalization of the GCV formula, where the trace of the smoothing matrix (for linear smoothers) is replaced by $\widehat{\mathsf{df}}$, i.e., the trace of the Jacobian $(\partial/\partial\boldsymbol{y})\boldsymbol{X}\widehat{\boldsymbol{\beta}}$. Moving beyond the estimation of the in-sample error $\|\boldsymbol{X}\widehat{\boldsymbol{\beta}} - \boldsymbol{X}\boldsymbol{\beta}_0\|_2^2$, this generalization of GCV beyond linear smoothers is known to be consistent in the sense:

$$\frac{\|\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}\|_2^2/n}{(1 - \widehat{\mathsf{df}}/n)^2} \bigg/ \left(\|\boldsymbol{\Sigma}^{1/2}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)\|_2^2 + \sigma^2 \right) \xrightarrow{\mathrm{p}} 1,$$

under linear models with jointly normal observations $(\boldsymbol{x}_i, y_i)$ and $\mathbb{E}[\boldsymbol{x}_i \boldsymbol{x}_i^\top] = \boldsymbol{\Sigma}$. This consistency is proven for the lasso (Bayati and Montanari, 2011; Bayati et al., 2013; Miolane and Montanari, 2021; Celentano et al., 2023), and for regularized least-squares with general convex penalty (Bellec, 2023, Section 3). Bellec and Shen (2022) generalized the above GCV estimator for the risk $\|\boldsymbol{\Sigma}^{1/2}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)\|_2^2 + \|\boldsymbol{\varepsilon}\|_2^2/n$ in the context of regularized M-estimation where the noise has possibly infinite

variance. Tan et al. (2022) studied a generalization of GCV to estimate the out-of-sample error in a multi-task regression model where each response y_i is vector-valued.

Closely related to GCV, Rad and Maleki (2020) introduced the Approximate Leave-One Out (ALO) estimator for the out-of-sample error of regularized M-estimators and showed the consistency of ALO under smoothness assumptions on the data-fitting loss and the regularizer. Xu et al. (2019) used Approximate Message Passing to develop risk estimates for estimators constructed with separable loss and separable regularizer, and showed that these estimates, including ALO, enjoy finite error bounds. ALO has been extended to non-differentiable regularizers (Rad and Maleki, 2020, Section 2.2) as well as non-differentiable losses and non-differentiable regularizers (Wang et al., 2018). Its theoretical accuracy has recently been demonstrated for LASSO and elastic net (Auddy et al., 2023).

Ensemble risk characterization. Ensemble methods combine weak predictors to produce strong predictors and are a common part of the statistical machine learning toolkit (Hastie et al., 2009). Several variants exist such as bagging (Breiman, 1996; Bühlmann and Yu, 2002), random forests (Breiman, 2001), stacking (Wolpert, 1992), among others. These methods are widely used because of their ability to improve prediction accuracy and robustness. The risk behavior of ensemble methods has been a topic of recent interest. Patil et al. (2023) provide a strategy for analyzing the risk of an ensemble of general predictors and present exact risk asymptotics for an ensemble of ridge predictors. Other related work on ensemble risk analysis include LeJeune et al. (2020); Ando and Komaki (2023); Loureiro et al. (2022); Adlam and Pennington (2020b), among others, which provide insights into the fluctuations in risk of ensemble methods, demonstrating that the risk can be significantly reduced by properly choosing the ensemble size. They also highlight the importance of the correlation structure among the predictors in determining the risk of the ensemble. Recently, Du et al. (2023) show that the naive extension of the GCV estimator for the ridge ensemble is generally inconsistent for a finite ensemble size $M > 1$, but it becomes consistent again when M approaches infinity. The follow-up work of Patil and Du (2023) provides certain structural and risk equivalences between subsampling and ridge regularization. This work identifies an inflation factor of Δ/M, for some finite bias term Δ (independent of M), between the risk of the ensemble at a given regularization level and full ridge at a larger "induced" regularization level. This inflation vanishes as $M \to \infty$. Even though GCV is consistent for ridge regression ($M = 1$) or the infinite-ensemble ridge ($M = \infty$), these two works suggest that the naive extension of GCV for finite-ensemble ridge may be inconsistent for $1 < M < \infty$.

Comparison with sample splitting and out-of-bag risk estimation. In the context of ensemble learning, other types of cross-validation methods include the sample-split CV and out-of-bag CV, which utilize test samples to estimate the out-of-sample risk. The sample-split CV is the most common strategy for risk estimation (Patil et al., 2023). However, this approach presents several drawbacks. The estimator can suffer from finite-sample efficiency due to the inherent sample splitting: V-fold CV estimates the risk of an estimator trained on $(V - 1)n$ observations only instead of the risk of the estimator trained on the full dataset of size n. Figure 1 of Rad and Maleki (2020) provides a clear illustration of this drawback. Furthermore, the accuracy of the estimator can suffer when dealing with large subsample sizes. On the other hand, Lopes (2019); Lopes et al. (2020); Du et al. (2024) use out-of-bag risk estimates to extrapolate the risk estimation. Their estimators provide a more efficient risk estimator without sample splitting but do not perform well with large subsample sizes, as the out-of-bag sample size becomes small. In light of the aforementioned limitations of existing methods, our aim is to develop a risk estimator that fulfills certain desiderata.

First, we would like an estimator that does not require sample splitting. This would address the issue of finite-sample efficiency faced by split cross-validation. Second, we aim to extend the applicability of GCV to accommodate any ensemble size. This would overcome the limitation of GCV being inconsistent for finite ensembles, as identified in Du et al. (2023). In this paper, we present two intermediate risk estimators that fulfill both these desiderata, which ultimately lead to our final corrected GCV.

1.3 Organization

The rest of the paper is organized as follows. In Section 2, we set our notation and define the ensemble estimators and their prediction risks. In Section 3, we define our risk estimators and provide a qualitative comparison to other related risk estimators. In Section 4, we provide a finite-sample analysis under stronger assumptions on the feature and response distributions. In Section 5, we show that the risk estimators are consistent under mild assumptions on the feature and response distributions for ridge ensembles. Section 6 concludes the paper and offers several follow-up directions. Proofs of all the theoretical results are included in the supplement, which also contains a summary of the notational conventions used throughout the paper. The source code for generating all the experimental figures in this paper can be accessed at: [code repository].

2 Background

Consider the standard supervised regression setting where we observe n i.i.d. samples $\{(\boldsymbol{x}_i, y_i) : i \in [n]\}$ in $\mathbb{R}^p \times \mathbb{R}$, where $[n]$ stands for the index set $\{1, 2, \ldots, n\}$. Let $\boldsymbol{X} \in \mathbb{R}^{n \times p}$ denote the feature matrix whose i-th row contains $\boldsymbol{x}_i^\top$ and $\boldsymbol{y} \in \mathbb{R}^n$ denote the response vector whose i-th entry is y_i.

2.1 Ensemble estimator and prediction risk

For each $m \in [M]$, let I_m be a non-empty subset of $[n]$. Let $(\boldsymbol{X}_{I_m}, \boldsymbol{y}_{I_m})$ denote the corresponding random design matrix and response vector associated with the subsampled dataset $\{(\boldsymbol{x}_i, y_i) : i \in I_m\}$. For each of the subsampled datasets $(\boldsymbol{X}_{I_m}, \boldsymbol{y}_{I_m})$ for $m \in [M]$, we consider a penalized least-squares estimator:

$$\widehat{\boldsymbol{\beta}}_m \in \operatorname*{argmin}_{\boldsymbol{b} \in \mathbb{R}^p} \frac{1}{2|I_m|} \sum_{i \in I_m} (y_i - \boldsymbol{x}_i^\top \boldsymbol{b})^2 + g_m(\boldsymbol{b}) = \operatorname*{argmin}_{\boldsymbol{b} \in \mathbb{R}^p} \frac{1}{2|I_m|} \|\boldsymbol{L}_{I_m}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b})\|_2^2 + g_m(\boldsymbol{b}) \quad (2)$$

where $g_m : \mathbb{R}^p \to \mathbb{R}$ is a prespecified convex penalty and $\boldsymbol{L}_{I_m}$ is the diagonal projection matrix with entries $(\boldsymbol{L}_{I_m})_{ii} = I\{i \in I_m\}$. Following Stein (1981), we define the degrees of freedom of the estimator $\widehat{\boldsymbol{\beta}}_m$ as:

$$\widehat{\mathsf{df}}_m = \operatorname{tr}[(\partial/\partial\boldsymbol{y})\boldsymbol{X}\widehat{\boldsymbol{\beta}}_m]. \quad (3)$$

For the estimator defined in (2) using the full data $(\boldsymbol{x}_i, y_i)_{i \in [n]}$ with particular choices of the penalty function g, the quantity $\widehat{\mathsf{df}}$ has explicit expressions; see, for example, Zou and Hastie (2007); Tibshirani and Taylor (2012); Dossal et al. (2013); Vaiter et al. (2012, 2017); Bellec and Shen (2022)) among others. Table 1 presents the explicit known formulae for the lasso, ridge, and elastic net.

The ensemble estimator constructed using (2) from the subsample datasets $\{(\boldsymbol{X}_{I_m}, \boldsymbol{y}_{I_m})\}_{m \in [M]}$ is defined as:

$$\widetilde{\boldsymbol{\beta}}_M(\{I_m\}_{m=1}^M) := \frac{1}{M} \sum_{m=1}^M \widehat{\boldsymbol{\beta}}_m. \quad (4)$$

7

Table 1: Explicit formulae for the degrees of freedom in (3) of the estimator (2) for commonly used penalty functions. Here $\widehat{S} = \{j \in [p] : e_j^\top \widehat{\boldsymbol{\beta}}_m \neq 0\}$ is the set of active variables, $\boldsymbol{X}_{\widehat{S}}$ is the submatrix of $\boldsymbol{X}_{I_m}$ made of columns indexed in $\widehat{S}$. See Zou and Hastie (2007); Tibshirani and Taylor (2012); Dossal et al. (2013); Bellec and Zhang (2021), among others.

Estimator $\widehat{\boldsymbol{\beta}}_m$	Regularizer $g_m(\boldsymbol{b})$	Degrees of freedom $\widehat{\mathsf{df}}_m$
Lasso	$\lambda\|\boldsymbol{b}\|_1$	$|\widehat{S}|$
Ridge	$\frac{\lambda}{2}\|\boldsymbol{b}\|_2^2$	$\mathrm{tr}\left[\boldsymbol{X}_{I_m}\left(\boldsymbol{X}_{I_m}^\top\boldsymbol{X}_{I_m} + |I_m|\lambda\boldsymbol{I}\right)^{-1}\boldsymbol{X}_{I_m}^\top\right]$
Elastic net	$\lambda_1\|\boldsymbol{b}\|_1 + \frac{\lambda_2}{2}\|\boldsymbol{b}\|_2^2$	$\mathrm{tr}\left[\boldsymbol{X}_{\widehat{S}}\left(\boldsymbol{X}_{\widehat{S}}^\top\boldsymbol{X}_{\widehat{S}} + |I_m|\lambda_2\boldsymbol{I}\right)^{-1}\boldsymbol{X}_{\widehat{S}}^\top\right]$

Though it is possible to extend the analysis by using unequal weights for the ensemble estimator (4), we focus on the equal-weighted ensemble estimator for simplicity; see Remark 4 for more discussion. For brevity, we will omit the dependency on $\{I_m\}_{m\in[M]}$ for the ensemble estimator and simply write $\widetilde{\boldsymbol{\beta}}_M$ when it is clear from the context. By the linearity property of the trace operator, the degrees of freedom of the ensemble estimator can be easily shown to be the average of the individual degrees of freedom:

$$\widetilde{\mathsf{df}}_M = \frac{1}{M}\sum_{m=1}^{M}\widehat{\mathsf{df}}_m. \tag{5}$$

We assess the performance of the M-ensemble predictor $\widetilde{\boldsymbol{\beta}}_M$ via conditional squared prediction risk:

$$R_M := \mathbb{E}_{(\boldsymbol{x}_0, y_0)}\left[(y_0 - \boldsymbol{x}_0^\top\widetilde{\boldsymbol{\beta}}_M)^2 \mid (\boldsymbol{X}, \boldsymbol{y}), \{I_m\}_{m=1}^{M}\right], \tag{6}$$

where $(\boldsymbol{x}_0, y_0)$ is an independent test point sampled from the same distribution as the training dataset $(\boldsymbol{X}, \boldsymbol{y})$. Note that the conditional risk R_M is a scalar random variable that depends on both the dataset $(\boldsymbol{X}, \boldsymbol{y})$ and the random samples $\{I_m : m \in [M]\}$. Our goal is to construct estimators for R_M that work for all choices of ensemble sizes M.

2.2 GCV for ensemble estimator

Before we delve into our proposed estimators, let us first consider the ordinary Generalized Cross-Validation (GCV) estimator. Suppose that we have a linear predictor $\widehat{f}(\boldsymbol{x}) = \boldsymbol{x}^\top\widehat{\boldsymbol{\beta}}$ where $\widehat{\boldsymbol{\beta}}$ is obtained through (2) based on the dataset $(\boldsymbol{X}, \boldsymbol{y})$. With the notation $\widehat{\mathsf{df}}$ in (1), the ordinary GCV estimator for the prediction risk of this estimate $\widehat{\boldsymbol{\beta}}$ is defined as:

$$\widehat{R}^{\mathrm{gcv}} = \frac{\|\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}\|_2^2/n}{(1 - \widehat{\mathsf{df}}/n)^2}. \tag{7}$$

The numerator of the GCV estimator, representing the average training error, usually underestimates the true prediction risk. The denominator of the GCV estimator, smaller than one, is designed to correct this underestimation. On the other hand, the denominator of the GCV estimator accounts for this training "optimism". Note that this definition of the degree-of-freedom adjusted estimator coincides with the classic definition of GCV for linear smoother (Golub et al., 1979; Craven and Wahba, 1979), where the trace of the smoothing matrix represents the degrees of freedom of the linear predictor. Existing literature provides a rich theoretical background for the consistency and efficiency of the ordinary GCV estimator. For instance, its behaviors have been

analyzed by Bayati and Montanari (2011); Bayati et al. (2013); Miolane and Montanari (2021); Celentano et al. (2023) for the lasso and in Section 3 of Bellec (2023) for penalized M-estimator. For ridge regression, Patil et al. (2021) show the uniform consistency of the ordinary GCV even in the overparameterized regimes when the feature size p is larger than the sample size n. Beyond ridge regression, the consistency of the estimator (8) is shown to be consistent for other convex penalized estimators under suitable design conditions, see for example, Bayati and Montanari (2011); Bayati et al. (2013); Bellec (2023).

Two natural extensions of GCV for bagging can be considered. For the first extension ((8) below), the subsamples $I_1, \ldots, I_M$ are sampled without looking at the data, and subsequently only the training data $(\boldsymbol{x}_i, y_i)_{i \in I_{1:M}}$ is used for training, where $I_{1:M} = \cup_{m=1}^{M} I_m$. Then, since the training data is $(\boldsymbol{x}_i, y_i)_{i \in I_{1:M}}$, a natural extension of the ordinary GCV for the ensemble estimator (4) is the following estimator:

$$\frac{\|\boldsymbol{L}_{I_{1:M}}(\boldsymbol{y} - \boldsymbol{X}\widetilde{\boldsymbol{\beta}}_M)\|_2^2/|I_{1:M}|}{(1 - \widetilde{\mathsf{df}}_M/|I_{1:M}|)^2}, \tag{8}$$

where $\boldsymbol{L}_I$ is a diagonal matrix whose i-th diagonal entry is one if $i \in I$ and zero otherwise. In the special case when the ensemble size is $M = 1$ or the subsample size is $k = n$, the definition reduces to the ordinary GCV (7), for which consistency has been established under various settings. For ensemble size $M > 1$, there is less understanding of the behaviors of this estimator for general predictors. However, it is shown to be inconsistent when the ensemble size is $M = 2$ for ridge predictors (Du et al., 2023). The inconsistency in large part happens because, for a finite ensemble size M, the residuals computed using the bagged predictor contain non-negligible fractions of out-of-sample and in-sample, and all of them are treated equally.

For the second extension, we consider that all data $(\boldsymbol{x}_i, y_i)_{i \in [n]}$ is used even if some $i \in [n]$ does not belong to any of the subsamples $I_1, \ldots, I_M$. This leads to the ensemble GCV:

$$\widetilde{R}_M^{\mathrm{gcv}} = \frac{\|(\boldsymbol{y} - \boldsymbol{X}\widetilde{\boldsymbol{\beta}}_M)\|_2^2/n}{(1 - \widetilde{\mathsf{df}}_M/n)^2}. \tag{9}$$

As $M \to \infty$ with uniformly sampled $I_1, \ldots, I_M$, the union $I_{1:M}$ closely approaches the full set $[n]$ and the difference between (8) and (9) vanishes. Both (8) and (9) are consistent as M tends to infinity for ridge predictors (Du et al., 2023); namely, $\widetilde{R}_\infty^{\mathrm{gcv}} \overset{\mathrm{p}}{\to} R_\infty$. For finite ensemble sizes, however, (9) is generally preferred over (8) because one may gain efficiency from more observations (except when considering ensemble sizes that are neither extremely large nor when considering no ensemble). As we will show in Theorem 3, the ensemble GCV estimator, as defined in (9), is generally inconsistent for finite ensemble sizes. The primary goal of this paper is to fully understand this phenomenon, correct the inconsistency of GCV estimate (9), and develop a risk estimator that is consistent for all ensemble sizes M.

3 Proposed GCV correction

In this section, we aim to address the limitations of the naive GCV estimators (8) and (9), which, as we demonstrate, fail to produce consistent estimates for the true prediction risk (6) under finite ensemble sizes $M > 1$. By introducing term-by-term adjustments for each term of the decomposition (10) below, we derive corrected and consistent risk estimators that hold for any ensemble size M.

3.1 Intermediate risk estimators

We begin by considering the decomposition of (6) into its constituent components. The motivation behind our proposed risk estimators stems from this decomposition. It is easy to see that the risk of the M-ensemble estimator can be decomposed as follows:

$$R_M = \frac{1}{M^2} \sum_{m,\ell \in [M]} \mathbb{E}_{\boldsymbol{x}_0, y_0} \left[(y_0 - \boldsymbol{x}_0^\top \widehat{\boldsymbol{\beta}}_m)(y_0 - \boldsymbol{x}_0^\top \widehat{\boldsymbol{\beta}}_\ell) \mid (\boldsymbol{X}, \boldsymbol{y}), \{I_m, I_\ell\} \right]$$

$$= \frac{1}{M^2} \sum_{m,\ell \in [M]} \underbrace{(\widehat{\boldsymbol{\beta}}_m - \boldsymbol{\beta}_0)^\top \boldsymbol{\Sigma} (\widehat{\boldsymbol{\beta}}_\ell - \boldsymbol{\beta}_0) + \sigma^2}_{R_{m,\ell}}. \tag{10}$$

Here, $\boldsymbol{\beta}_0 := \mathbb{E}[\boldsymbol{x}_0 \boldsymbol{x}_0^\top]^{-1} \mathbb{E}[\boldsymbol{x}_0 y_0]$ denotes the coefficient vector of linear projection of y_0 onto $\boldsymbol{x}_0$, and $\sigma^2 := \mathbb{E}[(y_0 - \boldsymbol{x}_0^\top \boldsymbol{\beta}_0)^2]$ denotes the energy in the residual component in the response. As indicated in (10), we will denote the component in the risk decomposition corresponding to $\widehat{\boldsymbol{\beta}}_m$ and $\widehat{\boldsymbol{\beta}}_\ell$ by $R_{m,\ell}$. The fundamental idea behind our risk estimators is to estimate each component $R_{m,\ell}$ individually. The final risk estimator is then the sum of these estimated risk components.

We propose an adjustment for each term $R_{m,\ell}$ in the risk decomposition, which leads to a $\sqrt{n}$-consistent estimator. This adjustment is a key component of our proposed risk estimators and contributes to their improved performance over the ordinary GCV estimator. Before we present the corrected GCV estimators, we first introduce two intermediate estimators that give rise to our final correction.

Estimator using overlapping observations. Our first estimator is defined as analogous to R_M, with each of $R_{m,\ell}$ replaced by its estimate $\widehat{R}_{m,\ell}^{\mathrm{ovlp}}$ defined below:

$$\widetilde{R}_M^{\mathrm{ovlp}} = \frac{1}{M^2} \sum_{m,\ell \in [M]} \widehat{R}_{m,\ell}^{\mathrm{ovlp}}, \quad \widehat{R}_{m,\ell}^{\mathrm{ovlp}} = \frac{(\boldsymbol{y}_{I_m \cap I_\ell} - \boldsymbol{X}_{I_m \cap I_\ell} \widehat{\boldsymbol{\beta}}_m)^\top (\boldsymbol{y}_{I_m \cap I_\ell} - \boldsymbol{X}_{I_m \cap I_\ell} \widehat{\boldsymbol{\beta}}_\ell) / |I_m \cap I_\ell|}{(1 - \widehat{\mathsf{df}}_m / |I_m|)(1 - \widehat{\mathsf{df}}_\ell / |I_\ell|)}. \tag{11}$$

Here, the superscript "ovlp" is used because the numerator of $\widehat{R}_{m,\ell}^{\mathrm{ovlp}}$ uses only the overlapping observations of subsets I_m and I_ℓ when computing the residuals. In contrast, our next estimator will use all the available observations when computing similar residuals.

The intuition behind the ovlp-estimator (11) is that the summand in (11) is consistent for the corresponding summand in (10). This result is already known for $m = \ell$ under suitable conditions; see, for example, Bellec (2023) and references therein. In the present paper, we extend this consistency result from $m = \ell$ to any pair of $(m, \ell)_{m,\ell \in [M]}$, see Theorem 1 for a formal statement.

Remark 1 (Special cases of the ovlp-estimator)**.** We consider two special cases under equal subsample sizes with $|I_m| = k$ for all $m \in [M]$: (a) When $k = n$, the estimator in (11) matches with ordinary GCV estimator (7), which is consistent. (b) When $M = 1$, notice that $\widetilde{R}_M^{\mathrm{ovlp}}$ is equal to the GCV in (8) restricted to k subsampled observations. Thus, (11) is consistent for $M = 1$.

It is worth noting that the definition of the ovlp-estimator implicitly requires $|I_m \cap I_\ell| > 0$, which necessitates a large enough subsample size k to guarantee the overlapping sets $I_m \cap I_\ell$ are not empty. Furthermore, when the regularization is near zero (e.g., in ridgeless estimators), the term $1 - \widehat{\mathsf{df}}_m / |I_m|$ in the denominator can approach zero, further degrading the estimator's performance. To address these shortcomings, it is beneficial to use data from both I_m and I_ℓ, as well as other observations, to estimate the individual terms in the decomposition (10). This motivates us to construct the next estimator.

Estimator using all observations. We propose the following estimator, which estimates each of $R_{m,l}$ using all the data $(\boldsymbol{X}, \boldsymbol{y})$. The expression is given below:

$$\widetilde{R}_M^{\text{full}} = \frac{1}{M^2} \sum_{m,\ell \in [M]} \widehat{R}_{m,\ell}^{\text{full}}, \qquad \widehat{R}_{m,\ell}^{\text{full}} = \frac{(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}_m)^\top (\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}_\ell)/n}{1 - \widehat{\mathsf{df}}_m/n - \widehat{\mathsf{df}}_\ell/n + (\widehat{\mathsf{df}}_m\widehat{\mathsf{df}}_\ell/|I_m||I_\ell|) \cdot |I_m \cap I_\ell|/n}. \tag{12}$$

In contrast to the "ovlp" estimator in (11), we use superscript "full" in (12), which refers to the risk estimator that uses all the available observations ($\boldsymbol{y}$ and $\boldsymbol{X}$ are used in the numerator to compute residuals). Because the two intermediate estimators use different proportions of the samples to quantify the optimism, they require different degrees of freedom adjustments presented in the denominators.

Compared to $\widetilde{R}_M^{\text{ovlp}}$, the full-estimator offers several advantages. Firstly, it utilizes all available data, which is beneficial, especially when the subsample sizes $|I_m|$ and $|I_\ell|$ are small, or the union of subsets is limited. Secondly, its denominator is lower bounded by a positive constant, ensuring numerical stability, especially for base estimators with regularization close to zero. It is worth noting that hybrid estimators can be constructed that utilize data points in a range between the extreme cases of the intersection and the full set of observations. We will not focus on such hybrid strategies in this paper.

Remark 2 (Special cases of the full-estimator). As with Remark 1, we consider two special cases: (a) When $k = n$, observe from (13) that (12) matches with (9), which is consistent, as argued previously. (b) When $M = 1$, the full-estimator in (12), reduces to:

$$\widetilde{R}_1^{\text{full}} = \frac{(\boldsymbol{y} - \boldsymbol{X}\widetilde{\boldsymbol{\beta}}_1)^\top (\boldsymbol{y} - \boldsymbol{X}\widetilde{\boldsymbol{\beta}}_1)/n}{1 - 2\widetilde{\mathsf{df}}_1/n + (\widetilde{\mathsf{df}}_1)^2/(kn)} \tag{13}$$

$$= \frac{(\boldsymbol{y} - \boldsymbol{X}\widetilde{\boldsymbol{\beta}}_1)^\top \boldsymbol{L}_1(\boldsymbol{y} - \boldsymbol{X}\widetilde{\boldsymbol{\beta}}_1)/n + (\boldsymbol{y} - \boldsymbol{X}\widetilde{\boldsymbol{\beta}}_1)^\top \boldsymbol{L}_{1^c}(\boldsymbol{y} - \boldsymbol{X}\widetilde{\boldsymbol{\beta}}_1)/n}{1 - 2\widetilde{\mathsf{df}}_1/n + (\widetilde{\mathsf{df}}_1)^2/(kn)}, \tag{14}$$

where $\boldsymbol{L}_{1^c} = \boldsymbol{I}_n - \boldsymbol{L}_1$ and $\boldsymbol{L}_1 = \boldsymbol{L}_{I_1}$ for brevity. To see that (13) is consistent for any subsample size k, from the consistency of ordinary GCV for a single base predictor and no bagging, with observations $(\boldsymbol{x}_i, y_i)_{i \in I_1}$ and sample size $k = |I_1|$, from Section 3 of Bellec (2023), we have

$$\frac{(\boldsymbol{y} - \boldsymbol{X}\widetilde{\boldsymbol{\beta}}_1)^\top \boldsymbol{L}_1(\boldsymbol{y} - \boldsymbol{X}\widetilde{\boldsymbol{\beta}}_1)/k}{(1 - \widetilde{\mathsf{df}}_1/k)^2 R_1} \xrightarrow{\text{P}} 1, \tag{15}$$

where R_1 is the risk of $\widetilde{\boldsymbol{\beta}}_1$. Moreover, by the law of large numbers as $(n - k) \to \infty$,

$$\frac{(\boldsymbol{y} - \boldsymbol{X}\widetilde{\boldsymbol{\beta}}_1)^\top \boldsymbol{L}_{1^c}(\boldsymbol{y} - \boldsymbol{X}\widetilde{\boldsymbol{\beta}}_1)}{R_1(n - k)} \xrightarrow{\text{P}} 1. \tag{16}$$

Writing (15) and (16) as $1 + o_{\mathbb{P}}(1)$ and substituting into (14), we get that

$$\frac{\widetilde{R}_1^{\text{full}}}{R_1} = \frac{[1 + o_{\mathbb{P}}(1)] \cdot (1 - \widetilde{\mathsf{df}}_1/k)^2 \cdot (k/n) + [1 + o_{\mathbb{P}}(1)] \cdot (n - k)/n}{1 - 2\widetilde{\mathsf{df}}_1/n + (\widetilde{\mathsf{df}}_1)^2/(kn)} \xrightarrow{\text{P}} 1.$$

3.2 Corrected GCV for ensembles

Though the two intermediate estimators are consistent, the nature of the term-by-term correction requires enumeration over all pairs of $m, \ell \in [M]$. This quadratic complexity in M prevents them from being computationally tractable and practical for large M. To mitigate the computational drawbacks of the intermediate estimators, we further propose a corrected GCV estimator as below.
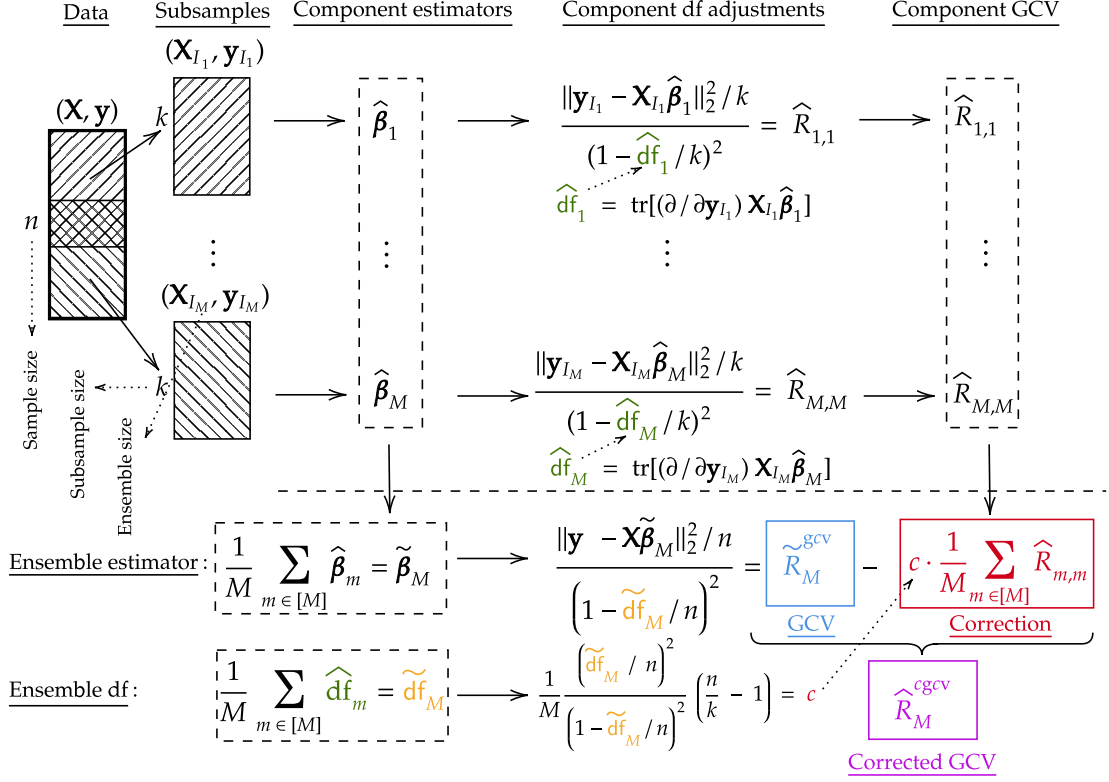
Figure 2: Illustration of the corrected generalized cross-validation for subsampled ensembles with equal subsample sizes k.

Definition 1 (Corrected GCV for ensemble estimator). The corrected GCV estimator for ensemble size M is defined as:

$$
\widetilde{R}_M^{\text{cgcv,ovlp}} = \underbrace{\frac{\|\boldsymbol{y} - \boldsymbol{X}\widetilde{\boldsymbol{\beta}}_M\|_2^2/n}{(1 - \widetilde{\mathsf{df}}_M/n)^2}}_{\widetilde{R}_M^{\text{gcv}}} - \underbrace{\frac{1}{M}\left\{\frac{(\widetilde{\mathsf{df}}_M/n)^2}{(1 - \widetilde{\mathsf{df}}_M/n)^2}\frac{1}{M}\sum_{m\in[M]}\left(\frac{n}{|I_m|} - 1\right)\widehat{R}_{m,m}^{\text{ovlp}}\right\}}_{\text{correction}}. \tag{17}
$$

We will also consider the alternative expression:

$$
\widetilde{R}_M^{\text{cgcv,full}} = \underbrace{\frac{\|\boldsymbol{y} - \boldsymbol{X}\widetilde{\boldsymbol{\beta}}_M\|_2^2/n}{(1 - \widetilde{\mathsf{df}}_M/n)^2}}_{\widetilde{R}_M^{\text{gcv}}} - \underbrace{\frac{1}{M}\left\{\frac{(\widetilde{\mathsf{df}}_M/n)^2}{(1 - \widetilde{\mathsf{df}}_M/n)^2}\frac{1}{M}\sum_{m\in[M]}\left(\frac{n}{|I_m|} - 1\right)\widehat{R}_{m,m}^{\text{full}}\right\}}_{\text{correction}}, \tag{18}
$$

where the only difference is the use of $\widehat{R}_{m,m}^{\text{full}}$ instead of $\widehat{R}_{m,m}^{\text{ovlp}}$ in the rightmost sum in (18).

The evaluation of (18) (and (17)) only requires computing three components: (1) the full training of the ensemble estimator; (2) the average $\widetilde{\mathsf{df}}_M$ defined in (5); and (3) the correction term that is built on the intermediate estimators. More importantly, this proposed estimator can be evaluated only in linear time with respect to M. See Figure 2 for an illustration of computing (18).

The correction term in (18) is derived from the insights gained through the intermediate estimators $\widetilde{R}_M^{\text{ovlp}}$ and $\widetilde{R}_M^{\text{full}}$. It allows us to alleviate the computational challenges posed by the pairwise

12

enumeration in the intermediate estimators, leading to an estimator that retains the advantages of both while being more computationally tractable. While the correction is defined with respect to GCV in (9), one can also derive another correction with respect to (8). We prefer to present the correction with respect to (9) because it leads to a slightly simpler expression and because, in this case, the correction is always positive. Our theoretical results in the next section show the corrected GCV estimator (18) is consistent (Theorem 2), and the naive extension of GCV in (9) *over-estimates the risk* with positive probability (Theorem 3).

Remark 3 (Special cases of corrected GCV). As with Remarks 1 and 2, we consider two special cases: (a) When $k = n$, observe that the correction term is zero and the first term $\widetilde{R}_1^{\mathrm{gcv}}$ is consistent as argued previously; thus, $\widetilde{R}_M^{\mathrm{cgcv,full}}$ is consistent. (b) When $M = 1$, it is also not hard to see that $\widetilde{R}_1^{\mathrm{cgcv,full}}$ is consistent. As argued in Section 3.1, both $\widehat{R}_{1,1}^{\mathrm{ovlp}}$ and $\widehat{R}_{1,1}^{\mathrm{full}}$ are consistent. Now writing (15) and (16) as $1 + o_{\mathbb{P}}(1)$,

$$\frac{\widetilde{R}_1^{\mathrm{cgcv,full}}}{R_1} = \frac{[1 + o_{\mathbb{P}}(1)](1 - \widetilde{\mathsf{df}}_1/k)^2 k/n + [1 + o_{\mathbb{P}}(1)](n - k)/n}{(1 - \widetilde{\mathsf{df}}_1/n)^2} - \frac{[1 + o_{\mathbb{P}}(1)](\widetilde{\mathsf{df}}_1/n)^2(n - k)/k}{(1 - \widetilde{\mathsf{df}}_1/n)^2} \xrightarrow{\mathrm{P}} 1.$$

Here, the last equality follows because

$$\frac{k}{n}(1 - \widetilde{\mathsf{df}}_1/k)^2 + \frac{n - k}{n} - \frac{n - k}{k}(\widetilde{\mathsf{df}}_1/n)^2 = 1 - 2\widetilde{\mathsf{df}}_1/n + \widetilde{\mathsf{df}}_1^2/(kn) - \frac{n - k}{k}(\widetilde{\mathsf{df}}_1/n)^2$$

$$= 1 - 2\widetilde{\mathsf{df}}_1/n + \widetilde{\mathsf{df}}_1^2/n^2 = (1 - \widetilde{\mathsf{df}}_1/n)^2.$$

Remark 4 (Unequal weighted ensemble). When defining the ensemble estimator in (4) with unequal weights, one can derive a similar corrected GCV estimator. This requires modifying the risk expansion in (10) and using the intermediate estimators presented in Section 3.1. For simplicity, we present and analyze the corrected GCV with equal weights. Because the individual estimators are i.i.d. conditioned on data, among all choices of deterministic weights that sum to 1, the optimal combination of weights is simply $1/M$ for the lowest expected risk.

One way to see this is to note that, since the penalties $g_1, \ldots, g_M$ are the same, the component estimators $\widehat{\boldsymbol{\beta}}_1, \ldots, \widehat{\boldsymbol{\beta}}_M$ are exchangeable. Because of exchangeability, for weights $(w_1, \ldots, w_M)$, the expected risk is a convex polynomial of degree 2 that is invariant by permutation of the M inputs. By symmetry, a minimum is achieved at equal weights $w_1 = \cdots = w_M$. Indeed, if a minimum is achieved at $w_1^*, \ldots, w_M^*$, it is also achieved at the average over all permutations $\frac{1}{M!}\sum_\pi (w_{\pi(1)}^*, \ldots, w_{\pi(M)}^*)$ by Jensen's inequality, and this average is proportional to $(1, 1, \ldots, 1)$. More broadly, this is a consequence of the Purkiss Principle (Waterhouse, 1983). Thus, if we further restrict the weights to sum to 1, the optimal choice for each weight is $1/M$.

Our goal in the rest of the paper is to establish consistency properties of the corrected GCV for any M and to provide convergence rates under various assumptions on the data and component estimators. We will provide a non-asymptotic analysis in Section 4 and an asymptotic analysis in Section 5. In both sections, we assume the following sampling scheme.

Assumption A (Sampling strategy). The subsample sets $\{I_m\}_{m=1}^M$ are independent of $(\boldsymbol{X}, \boldsymbol{y})$ and i.i.d., uniformly distributed over subsets of $[n]$ with cardinality k.

Before we dive into our analysis, it is worth reiterating that the intermediate estimators (11) and (12) are defined for ensembles fitted on subsamples $\mathcal{D}_{I_m}$ for $m \in [M]$ of arbitrary sizes and

13

component estimators fitted on different penalty functions g_m for $m \in [M]$. While our proofs are simplified by assuming equal subsample sizes, this condition can be relaxed for certain theoretical claims to follow in Section 4.1. The formulation of the modified GCV, on the other hand, assumes that the same penalty function is used for estimators across subsamples and that subsample sizes do not vary much to ensure the validity of certain degrees of freedom concentrations, as we will explain in Section 4.2.

4 Non-asymptotic analysis for general ensembles

In this section, we provide a non-asymptotic analysis of the estimators introduced in the previous section. Our analysis is applicable to general convex penalized estimators such as ridge regression and elastic net, among others.

To facilitate our analysis, we will need to impose certain assumptions on the data generation process. Specifically, we will assume Gaussian features and linear models as encapsulated in the assumptions below. These assumptions serve as a starting point but are not indispensable for more specialized estimators, as we discuss in Section 5.

> **Assumption B** (Feature structure)**.** The design matrix $\boldsymbol{X} \in \mathbb{R}^{n \times p}$ has i.i.d. rows $\boldsymbol{x}_i$ for $i \in [n]$ each drawn from a Gaussian distribution: $\boldsymbol{x}_i \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma})$.

> **Assumption C** (Response structure)**.** The response vector $\boldsymbol{y}$ has i.i.d. entries y_i for $i \in [n]$ each drawn from a linear model with Gaussian noise: $y_i = \boldsymbol{x}_i^\top \boldsymbol{\beta}_0 + \varepsilon_i$, where $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$ is independent of $\boldsymbol{x}_i$.

Although these assumptions may appear restrictive, they serve to simplify our analysis and enable us to derive the estimate of risk as well as prove explicit non-asymptotic bounds. These assumptions can be substantially relaxed for specialized base estimators. In Section 5, we will demonstrate that we can significantly relax these assumptions by focusing on a specific base estimator of ridge(less) regression.

> **Assumption D** (Penalty structure)**.** The penalty function $g_m : \mathbb{R}^p \to \mathbb{R}$ in (2) is μ-strongly convex with respect to $\boldsymbol{\Sigma}$ for $\mu > 0$, i.e., the function $\boldsymbol{b} \mapsto g_m(\boldsymbol{b}) - (\mu/2)\boldsymbol{b}^\top \boldsymbol{\Sigma} \boldsymbol{b}$ is convex in $\boldsymbol{b} \in \mathbb{R}^p$.

If g_m is twice continuously differentiable, Assumption D is equivalent to $\inf_{\boldsymbol{b} \in \mathbb{R}^p} \nabla^2 g_m(\boldsymbol{b}) \succeq \mu \boldsymbol{\Sigma}$, where $\nabla^2 g_m(\boldsymbol{b}) \in \mathbb{R}^{p \times p}$ is the Hessian matrix. For instance, if $g_m(\boldsymbol{b}) = \boldsymbol{b}^\top \boldsymbol{\Sigma}_w \boldsymbol{b}/2$ (generalized ridge penalty) with a positive definite matrix $\boldsymbol{\Sigma}_w \in \mathbb{R}^{p \times p}$, Assumption D holds with $\mu = \sigma_{\min}(\boldsymbol{\Sigma}_w)/\sigma_{\max}(\boldsymbol{\Sigma})$. In particular, $\mu = \lambda/\sigma_{\max}(\boldsymbol{\Sigma})$ when $g_m(\boldsymbol{b}) = \lambda\|\boldsymbol{b}\|_2^2/2$. Note that the strong convexity penalty guarantees a certain differentiable structure of the penalized least-squares estimator (2) with respect to $(\boldsymbol{X}, \boldsymbol{y})$ (see, for example, Theorem 1 of Bellec and Shen (2022) and Appendix A.1), which helps us in the proof to bound certain quantities in terms of μ. This strong convexity assumption can be relaxed in certain cases, including the lasso $g_m(\boldsymbol{b}) = \lambda\|\boldsymbol{b}\|_1$, as studied in Celentano et al. (2023); Bellec (2023); however, the proof argument will be largely different from the proof used in the current paper. Finally, note that in the case of $p/n < 1$, if the smallest

eigenvalue of $\boldsymbol{\Sigma}$ is uniformly lower bounded by a constant away from 0, then the strongly convex assumption is not needed; we only need the penalty to be convex. For an explicit statement and proof of this fact in the context of Theorem 1, see Appendix A.7.

4.1 Guarantees for intermediate risk estimators

Our first result provides a non-asymptotic guarantee on the intermediate estimators. Recall the notation in (10): the estimation target is $R_M = M^{-2} \sum_{m,\ell} R_{m,\ell}$, where $R_{m,\ell} = (\widehat{\boldsymbol{\beta}}_m - \boldsymbol{\beta}_0)^\top \boldsymbol{\Sigma} (\widehat{\boldsymbol{\beta}}_\ell - \boldsymbol{\beta}_0) + \sigma^2$. For notational convenience, we introduce the following notation for each $m, \ell \in [M]$,

$$d_{m,\ell}^{\text{ovlp}} := |I_m \cap I_\ell| \left(1 - \frac{\widehat{\mathsf{df}}_m}{|I_m|} \right) \left(1 - \frac{\widehat{\mathsf{df}}_\ell}{|I_\ell|} \right), \qquad d_{m,\ell}^{\text{full}} := n - \widehat{\mathsf{df}}_m - \widehat{\mathsf{df}}_\ell + \frac{\widehat{\mathsf{df}}_m \widehat{\mathsf{df}}_\ell}{|I_m||I_\ell|} |I_m \cap I_\ell|, \quad (19)$$

which are the denominators in the definitions of $\widehat{R}_{m,\ell}^{\text{ovlp}}$ and $\widehat{R}_{m,\ell}^{\text{full}}$ in (11) and (12). Now, we are ready to provide the theoretical results of $\widehat{R}_{m,\ell}^{\text{ovlp}}$ and $\widehat{R}_{m,\ell}^{\text{full}}$ in Theorem 1.

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:

$$\mathbb{E}\left[\left| \frac{d_{m,\ell}^{\text{ovlp}}}{n} \cdot \frac{\widehat{R}_{m,\ell}^{\text{ovlp}} - R_{m,\ell}}{\sqrt{R_{m,m} R_{\ell,\ell}}} \right| \right] \le C \frac{\gamma^{7/2}}{\tau^2 c^3 \sqrt{n}}, \qquad \mathbb{E}\left[\left| \frac{d_{m,\ell}^{\text{full}}}{n} \cdot \frac{\widehat{R}_{m,\ell}^{\text{full}} - R_{m,\ell}}{\sqrt{R_{m,m} R_{\ell,\ell}}} \right| \right] \le C \frac{\gamma^{5/2}}{\tau^2 c^2 \sqrt{n}}. \quad (20)$$

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

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

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

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

Theorem 1 shows the rate of convergence for both the ovlp- and full-estimator are $n^{-1/2}$. Fixing other problem parameters (τ, c, γ, M), this bound in $n^{-1/2}$ is tight up to constants; e.g., see argument around equation (2.6) of Bellec (2023) in the case of $M = 1$ and the Ordinary Least-Squares. It is worth noting that the upper bounds regarding $\widetilde{R}_M^{\text{ovlp}}$ and $\widetilde{R}_M^{\text{full}}$ in (20) and (21) are slightly different, especially in the dependence of the ratio $c = k/n$ and γ; the upper bounds of the full-estimator are tighter than the ovlp-estimator in terms of its dependencies on c and γ. While these upper bounds may not be optimal in the dependence on c and γ, we empirically observe that the relative error of the full-estimator is smaller than the ovlp-estimator, especially when c is small—see Figure 3. We do not claim that the dependence on the parameter M is tight; rather, it appears to be an artifact of our proof technique. Specifically, while the current bound suggests that $M \ll n^{1/6}$ is needed for the bound to approach zero, this condition is not necessary in practice. Empirical results in Section C.6 indicate that the proposed estimators perform well for all values of M. A major improvement for the current polynomial dependence in M would come from replacing the bounds (20) by novel exponential probability bounds for a single pair $(m, \ell) \in [M]^2$. Union bounds

over $[M]^2$ pairs would then only induce a logarithmic dependence in M. Exponential bounds for the risk $\|\boldsymbol{\Sigma}^{1/2}(\widehat{\boldsymbol{\beta}}_m - \boldsymbol{\beta}_0)\|^2$ and the residual $\|\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}_m\|^2$ are available, thanks to the CGMT (Celentano et al., 2023; Loureiro et al., 2021). To derive exponential bounds for (21), one would also need similar exponential concentration bounds for the inner products $(\widehat{\boldsymbol{\beta}}_m - \boldsymbol{\beta}_0)^\top\boldsymbol{\Sigma}(\widehat{\boldsymbol{\beta}}_\ell - \boldsymbol{\beta}_0)$ and for the degrees of freedom $\widehat{\mathsf{df}}_m$. One avenue to obtain exponential bounds for the inner products is the conditional Gordon inequality developed in (Celentano and Montanari, 2021, Lemma F.2 and appendix L). On the other hand, we are not aware at this point of any tool that can provide exponential inequalities for the degrees of freedom $\widehat{\mathsf{df}}_m$, except in the special case of the Lasso (Celentano et al., 2023, Theorem 8). The reason that the Lasso case is special and that this argument does not generalize to penalty functions different from the ℓ_1 norm is discussed in (Bellec, 2023, around equation (3.8)).

Note that the guarantees in Theorem 1 are in a multiplicative form. This is preferable to the more common additive bounds because, in the regimes we study, the risk R_M may have varied scales with different subsample sizes; for example, when the subsample size is near the feature size and the regularization parameter is small, in which cases the risk can be high (Patil et al., 2023). It is also worth noting that we do not assume either the pure signal energy (i.e., $\|\boldsymbol{\beta}_0\|_2^2$) or the effective signal energy (i.e., $\|\boldsymbol{\Sigma}^{1/2}\boldsymbol{\beta}_0\|_2^2$) is bounded. This is possible because of the multiplicative bounds.

Remark 5 (Diminishing strong convexity parameter). The upper bounds in (20) and (21) give explicit dependence on the strong convexity parameter μ of Assumption D through $\tau = \min(1, \mu)$. Our results thus allow to choose μ decreasing in n, for instance $\mu = n^{-1/8-\epsilon}$ is sufficient to ensure that the upper bounds of (20) and (21) go to 0 as $n \to \infty$ while M, c^{-1}, γ are bounded by constants.

Remark 6 (Extensions beyond strongly convex regularizers). Without any strong convexity assumption on the penalty (i.e., $\mu = 0$), results such as Theorem 4 for $M = 1$ cannot be established using known techniques even for the well-studied lasso, unless additional assumptions are made on the subsample size. In particular, Proposition 4 of Celentano et al. (2023) shows that in the current proportional asymptotic regime with $n \asymp p$, the risk of a single lasso may be unbounded if the number of samples used to train the lasso falls below the Donoho-Tanner phase transition described in Section 3 of Celentano et al. (2023). While the consistency of GCV for a single lasso is a consequence of this theory above this phase transition (Celentano et al., 2023, Theorem 9), below the phase transition both $\|\boldsymbol{\Sigma}^{1/2}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)\|_2$ and $1/(1 - \widehat{\mathsf{df}}/n)$ are unbounded for certain $\boldsymbol{\beta}_0$. Furthermore, the constants appearing in the upper bounds in Celentano et al. (2023) explode as the sample size approaches the phase transition. We are not aware of any currently known techniques suitable for studying GCV below this phase transition, and we believe new ideas are needed. Thus, in the present context, where we subsample I_m of size $k = |I_m|$, small values of k would eventually fall below the phase transition, in which case the theoretical analysis of bagging GCV is currently out of reach.

We numerically compare the performance of the ovlp- and full-estimators for the ridge ensemble in Figure 3. It is clear that the relative error of the full-estimator is smaller than the ovlp-estimator across different subsample size k and tuning parameter λ, especially when k is small and the tuning parameter λ is small. We also observe that, as k increases towards the full sample size n, the performance of ovlp-estimator and full-estimator are getting closer; this makes sense because the expressions of ovlp-estimator and full-estimator will be the same when $k = n$. For comparison between the ovlp- and full-estimators in the context of the elastic net ensemble and the lasso ensemble, we observe a similar comparative trend, and the results are shown in Appendix C.1.
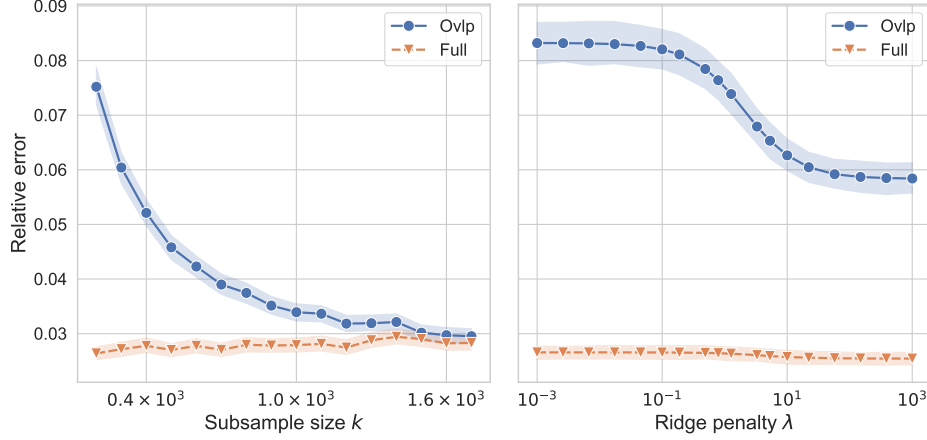
Figure 3: **Full-estimator $\widetilde{R}_M^{\text{full}}$ performs better than ovlp-estimator $\widetilde{R}_M^{\text{ovlp}}$ across different regularization and for small subsample size k.** Plots of relative errors of the "ovlp" and "full" estimators for ridge ensemble with ensemble size $M = 10$. 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. More details on the experimental setup can be found in Section 4.4.

4.2 Guarantees for corrected GCV

The next result provides non-asymptotic guarantees for corrected GCV. Recall the form of the two variants of CGCV, $\widetilde{R}_M^{\text{cgcv,full}}$ and $\widetilde{R}_M^{\text{cgcv,ovlp}}$ from (18) and (17), respectively:

$$\widetilde{R}_M^{\text{cgcv},\#} := \frac{\|\boldsymbol{y} - \boldsymbol{X}\widetilde{\boldsymbol{\beta}}_M\|_2^2/n}{(1 - \widetilde{\mathsf{df}}_M/n)^2} - \frac{(\widetilde{\mathsf{df}}_M/n)^2}{(1 - \widetilde{\mathsf{df}}_M/n)^2} \frac{1}{M^2}\left(\frac{n}{k} - 1\right)\sum_{m=1}^{M} \widehat{R}_{m,m}^{\#},$$

where $\# \in \{\text{full}, \text{ovlp}\}$, depending on which estimate is used in the rightmost sum to estimate $R_{m,m}$.

> **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:
>
> $$\mathbb{P}\left(\left|1 - \frac{\widetilde{R}_M^{\text{cgcv,ovlp}}}{R_M}\right| > \epsilon\right) \leq C\frac{M^4\gamma^{15/2}}{\epsilon\tau^6 c^6\sqrt{n}}, \qquad \mathbb{P}\left(\left|1 - \frac{\widetilde{R}_M^{\text{cgcv,full}}}{R_M}\right| > \epsilon\right) \leq C\frac{M^4\gamma^{13/2}}{\epsilon\tau^6 c^4\sqrt{n}}. \qquad (22)$$
>
> Thus, if $(M, \mu^{-1}, p/n, n/k)$ are bounded by constants independent of n, we have
>
> $$\widetilde{R}_M^{\text{cgcv,ovlp}}/R_M = 1 + \mathcal{O}_\mathbb{P}(n^{-1/2}), \qquad \widetilde{R}_M^{\text{cgcv,full}}/R_M = 1 + \mathcal{O}_\mathbb{P}(n^{-1/2}).$$

The upper bounds presented in Theorem 2 provide dependence of the accuracy of corrected GCV on several problem parameters, such as M, τ, c, and γ. The key message of Theorem 2 is that the corrected GCV estimators (both $\widetilde{R}_M^{\text{cgcv,ovlp}}$ and $\widetilde{R}_M^{\text{cgcv,full}}$) enjoy the rate of convergence $n^{1/2}$. As expected, numerical comparisons show that full-CGCV outperforms ovlp-CGCV, see Figures 11–13.

Remark 7 (GCV is consistent for infinite ensembles)**.** Note that for $M \to \infty$, it is easy to show that the correction term becomes 0 if $1 - \widetilde{\mathsf{df}}_M/n \geq \delta$ for some positive constant δ and $M^{-1} \sum_{m=1}^M \widehat{R}_{m,m}^{\#}$ is bounded from above. A special case of this result for ridge ensemble is proved in Du et al. (2023). While this literature showed that GCV is not consistent for ridge ensemble with $M = 2$, our results imply more than inconsistency by providing the correction term to make GCV consistent, and our results are applicable not only to ensemble ridge estimates but also to ensembles of any strongly-convex regularized least-squares, including elastic net estimates. In the next theorem, we show that GCV is inconsistent by providing a lower bound of the relative error of the GCV estimator.

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 $\widetilde{R}_M^{\mathrm{gcv}}$ in (9) over-estimates the risk R_M is lower bounded as:

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

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

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

In the context of ridge ensembles, Du et al. (2023) show that the GCV variant (8) is not consistent when ensemble size is $M = 2$. It is possible to extend the inconsistency results for the variant (8) for finite ensemble sizes greater than 1. We provide numerical experiments showing the inconsistency of this variant in Appendix C.7. As discussed in Section 2.2, for a moderate ensemble size M and subsample size k, we expect the GCV variant (8) to behave similarly to (9). Furthermore, the inconsistency of GCV implied by Theorem 3 applies even for other penalized least-squares estimator (2), extending the previous result to more generality.

Remark 8 (On the assumption of non-negligibility of degrees of freedom)**.** We argue here that assumption $\mathbb{P}(\widetilde{\mathsf{df}}_M/k > \delta_0) > \delta_0$ in Theorem 3 is unavoidable. If no such constant δ_0 exists as $n, p, k \to \infty$, by extracting a subsequence if necessary, we may assume $\widetilde{\mathsf{df}}_M/k \xrightarrow{\mathrm{P}} 0$. Herein, let $m \in [M]$ be fixed (e.g., take $m = 1$ throughout this remark). Then the positiveness of $(\widehat{\mathsf{df}}_m)_{m=1}^M$ (see (30) in Lemma 7) implies $\widehat{\mathsf{df}}_m/k \leq M \cdot \widetilde{\mathsf{df}}_M/k = o_p(1)$. The fact that GCV is consistent for $\widehat{\boldsymbol{\beta}}_m$, combined with $\widehat{\mathsf{df}}_m/k \xrightarrow{\mathrm{P}} 0$, gives $k^{-1} \sum_{i \in I_m} (y_i - \boldsymbol{x}_i^\top \widehat{\boldsymbol{\beta}}_m)^2/R_{m,m} \xrightarrow{\mathrm{P}} 1$. Now consider the deterministic $\bar{\boldsymbol{\beta}}$ defined as $\bar{\boldsymbol{\beta}} = \operatorname{argmin}_{\boldsymbol{b} \in \mathbb{R}^p} (\sigma^2 + \|\boldsymbol{\Sigma}^{1/2}(\boldsymbol{\beta}_0 - \boldsymbol{b})\|_2^2)/2 + g_m(\boldsymbol{b})$, where we replaced the objective function in (2) inside the minimum by its expectation. The strong convexity of the objective function of $\bar{\boldsymbol{\beta}}$ and the optimality condition of $\widehat{\boldsymbol{\beta}}_m$ give

$$\mu\|\boldsymbol{\Sigma}^{1/2}(\widehat{\boldsymbol{\beta}}_m - \bar{\boldsymbol{\beta}})\|_2^2 \leq [\|\boldsymbol{\Sigma}^{1/2}(\widehat{\boldsymbol{\beta}}_m - \boldsymbol{\beta}_0)\|_2^2 + \sigma^2] - [\|\boldsymbol{\Sigma}^{1/2}(\bar{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)\|_2^2 + \sigma^2] + 2(g_m(\widehat{\boldsymbol{\beta}}_m) - g_m(\bar{\boldsymbol{\beta}})),$$
$$0 \leq -\|\boldsymbol{L}_{I_m}(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}_m)\|_2^2/k + \|\boldsymbol{L}_{I_m}(\boldsymbol{y} - \boldsymbol{X}\bar{\boldsymbol{\beta}})\|_2^2/k + 2(g_m(\bar{\boldsymbol{\beta}}) - g_m(\widehat{\boldsymbol{\beta}}_m)).$$

Summing these two inequalities, the penalty terms cancel out, and using the law of large numbers for $\|\boldsymbol{L}_m(\boldsymbol{y}-\boldsymbol{X}\bar{\boldsymbol{\beta}})\|_2^2/k$, we get $\|\boldsymbol{\Sigma}^{1/2}(\widehat{\boldsymbol{\beta}}_m-\bar{\boldsymbol{\beta}})\|_2^2 \leq o_\mathbb{P}(1)[\sigma^2\|\boldsymbol{\Sigma}^{1/2}(\widehat{\boldsymbol{\beta}}_m-\boldsymbol{\beta}_0)\|_2^2 + \sigma^2\|\boldsymbol{\Sigma}^{1/2}(\bar{\boldsymbol{\beta}}-\boldsymbol{\beta}_0)\|_2^2]$ so

18

that $\|\mathbf{\Sigma}^{1/2}(\widehat{\boldsymbol{\beta}}_m - \bar{\boldsymbol{\beta}})\|_2^2/[\sigma^2 + \|\mathbf{\Sigma}^{1/2}(\bar{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)\|_2^2] \xrightarrow{\mathrm{P}} 0$. This means according to the norm $\|\mathbf{\Sigma}^{1/2}(\cdot)\|_2$, $\widehat{\boldsymbol{\beta}}_m$ behaves like the point mass $\bar{\boldsymbol{\beta}}$ up to a negligible multiplicative factor. This contradicts the typical asymptotics seen in the proportional regime with $p \asymp n$ (Bayati and Montanari, 2011; Thrampoulidis et al., 2018; Miolane and Montanari, 2021; Loureiro et al., 2022; Celentano et al., 2023; Wang et al., 2020, among others), so that the assumption $\mathbb{P}(\widetilde{\mathsf{df}}_M/k > \delta_0) > \delta_0$ is intrinsic to the proportional regime assumed throughout the paper.

4.3 Proof outlines

In this section, we provide the proof outlines and the ideas that lead to the definition of the ovlp and full estimators. Then we provide a heuristic derivation of the corrected GCV estimator from the full estimator.

By the risk decomposition (10), it is sufficient to study a single term in the double sum over $[M]^2$ and to focus on two estimators $\widehat{\boldsymbol{\beta}}$ and $\widetilde{\boldsymbol{\beta}}$, trained respectively on subsampled datasets $(\boldsymbol{x}_i, y_i)_{i \in I}$ and $(\boldsymbol{x}_i, y_i)_{i \in \widetilde{I}}$ for two subsets $I \subset [n]$ and $\widetilde{I} \subset [n]$ independent of $(\boldsymbol{X}, \boldsymbol{y})$, that is,

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

To study the success or failure of GCV in (7) or (8), it is natural to study the inner products between the residuals of $\widehat{\boldsymbol{\beta}}$ and $\widetilde{\boldsymbol{\beta}}$, of the form

$$\sum_{i \in J} \left(y_i - \boldsymbol{x}_i^\top \widehat{\boldsymbol{\beta}} \right) \left(y_i - \boldsymbol{x}_i^\top \widetilde{\boldsymbol{\beta}} \right) \tag{23}$$

for some subset $J \subset [n]$, as such inner products naturally appear by expanding the square in the numerators of the GCV estimates in (7) or (8). There are several natural candidates for the set J of observations to use in the inner product above, including $J = [n]$ or $J = \cup_{m \in [M]} I_m$, or, as we study specific estimators trained over I and $\widetilde{I}$, the intersection $I \cap \widetilde{I}$ as well as I or $\widetilde{I}$ themselves. The following proof sketch explains why the intermediate estimators (11)-(12) were defined in this manner, including the choice of set J in residual inner products of the form (23) and the various degrees of freedom adjustments visible in the denominators of (11)-(12).

The ovlp estimator. Define $V_{ii} = 1 - \frac{\partial \boldsymbol{x}_i^\top \widehat{\boldsymbol{\beta}}}{\partial y_i}$ as well as $\widetilde{V}_{ii} = 1 - \frac{\partial \boldsymbol{x}_i^\top \widetilde{\boldsymbol{\beta}}}{\partial y_i}$ for all $i \in [n]$. The degrees of freedom for the individual estimators $\widehat{\boldsymbol{\beta}}$ and $\widetilde{\boldsymbol{\beta}}$ are given by $\widehat{\mathsf{df}} = |I| - \sum_{i \in I} V_{ii}$ and $\widetilde{\mathsf{df}} = |\widetilde{I}| - \sum_{i \in \widetilde{I}} \widetilde{V}_{ii}$, respectively. By variants of Stein's formula with respect to $(\boldsymbol{X}, \boldsymbol{\epsilon})$ formally stated in Lemmas 8 and 9, the inner product (23) is well approximated as follows,

$$\sum_{i \in J} \left(y_i - \boldsymbol{x}_i^\top \widehat{\boldsymbol{\beta}} \right) \left(y_i - \boldsymbol{x}_i^\top \widetilde{\boldsymbol{\beta}} \right)$$
$$\approx -\left(\frac{\widetilde{\mathsf{df}}}{|\widetilde{I}| - \widetilde{\mathsf{df}}} \right) \sum_{i \in J \cap \widetilde{I}} \left(y_i - \boldsymbol{x}_i^\top \widehat{\boldsymbol{\beta}} \right) \left(y_i - \boldsymbol{x}_i^\top \widetilde{\boldsymbol{\beta}} \right) + \left(\sum_{i \in J} V_{ii} \right) \left(\sigma^2 + (\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)^\top \mathbf{\Sigma} (\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \right). \tag{24}$$

The right-most term is the unknown quantity of interest, $(\sigma^2 + (\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)^\top \boldsymbol{\Sigma}(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_0))$, that appears in the risk decomposition (10). This motivates rearranging (24) as

$$
\left(\sum_{i \in J} V_{ii}\right)\left(\sigma^2 + (\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)^\top \boldsymbol{\Sigma}(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)\right)
$$
$$
\approx \sum_{i \in J}\left(y_i - \boldsymbol{x}_i^\top \widehat{\boldsymbol{\beta}}\right)\left(y_i - \boldsymbol{x}_i^\top \widetilde{\boldsymbol{\beta}}\right) + \left(\frac{\widetilde{\mathsf{df}}}{|\widetilde{I}| - \widetilde{\mathsf{df}}}\right) \sum_{i \in J \cap \widetilde{I}}\left(y_i - \boldsymbol{x}_i^\top \widehat{\boldsymbol{\beta}}\right)\left(y_i - \boldsymbol{x}_i^\top \widetilde{\boldsymbol{\beta}}\right). \tag{25}
$$

Now, to simplify the second line in (25), we consider $J \subset [n]$ such that $J = J \cap \widetilde{I}$ (e.g., $J = I \cap \widetilde{I}$ or $J = \widetilde{I}$), so that the second line simplifies to $\left(\frac{|\widetilde{I}|}{|\widetilde{I}| - \widetilde{\mathsf{df}}}\right)\sum_{i \in J}\left(y_i - \boldsymbol{x}_i^\top \widehat{\boldsymbol{\beta}}\right)\left(y_i - \boldsymbol{x}_i^\top \widetilde{\boldsymbol{\beta}}\right)$. Dividing both sides by $\left(\sum_{i \in J} V_{ii}\right)$ then gives an estimator of the prediction risk. Specifically, we first consider $J = I \cap \widetilde{I}$. If given $(|\widetilde{I} \cap I|, I)$ the intersection $\widetilde{I} \cap I$ is uniformly distributed over subsets of I with cardinality $|\widetilde{I} \cap I|$, then Lemma 21 (on the concentration of sampling without replacement) gives

$$
\frac{1}{|I \cap \widetilde{I}|} \sum_{i \in I \cap \widetilde{I}} V_{ii} \approx \frac{1}{|I|} \sum_{i \in I} V_{ii} = 1 - \frac{\widehat{\mathsf{df}}}{|I|}
$$

to approximate the factor in front of $(\sigma^2 + (\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)^\top \boldsymbol{\Sigma}(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_0))$. Combining these approximations with $J = I \cap \widetilde{I}$ yields

$$
|I \cap \widetilde{I}|\left(1 - \widehat{\mathsf{df}}/|I|\right)\left(\sigma^2 + (\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)^\top \boldsymbol{\Sigma}(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)\right) \approx \frac{1}{1 - \widetilde{\mathsf{df}}/|\widetilde{I}|} \sum_{I \cap \widetilde{I}}\left(y_i - \boldsymbol{x}_i^\top \widehat{\boldsymbol{\beta}}\right)\left(y_i - \boldsymbol{x}_i^\top \widetilde{\boldsymbol{\beta}}\right),
$$

which gives rise to the definition of overlapping estimator (11). The superscript "ovlp" in (11) indicates that only overlapping observations are used in the residual inner product. These successive approximations motivate the definition of $\mathsf{Rem}_1, \mathsf{Rem}_2, \mathsf{Rem}_3$ in (32) of the formal proof of Theorem 1 for the "ovlp" estimator.

The full estimator. When the subsample size is small, the choice of $J = I \cap \widetilde{I}$ may not be desirable because it only uses residuals of the overlapping data, and J can be an empty set when I and $\widetilde{I}$ are disjoint. Observations that are neither in I nor in $\widetilde{I}$, i.e., $i \in [n] \setminus (I \cap \widetilde{I})$ would be particularity useful since they produce exactly unbiased estimates, with $\mathbb{E}[(y_i - \boldsymbol{x}_i^\top \widehat{\boldsymbol{\beta}})(y_i - \boldsymbol{x}_i^\top \widetilde{\boldsymbol{\beta}}) \mid (\widehat{\boldsymbol{\beta}}, \widetilde{\boldsymbol{\beta}})] = \sigma^2 + (\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)^\top \boldsymbol{\Sigma}(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)$ for $i \notin I \cap \widetilde{I}$ by independence. Ideally, in order to reduce variance, we would leverage a residual inner product over all $[n]$ samples. This motivates us to consider the approximation (24) with $J = [n]$ and $J = \widetilde{I}$. By choosing a specific weighted combination of these two approximations that make all residual inner product over $\widetilde{I}$ cancel out, we are left with an approximation between the residual inner product over $[n]$ and the target of interest, namely,

$$
\sum_{i \in [n]}\left(y_i - \boldsymbol{x}_i^\top \widehat{\boldsymbol{\beta}}\right)\left(y_i - \boldsymbol{x}_i^\top \widetilde{\boldsymbol{\beta}}\right) \approx \left[-\frac{\widetilde{\mathsf{df}}}{|\widetilde{I}|}\left(\sum_{i \in \widetilde{I}} V_{ii}\right) + \left(\sum_{i \in [n]} V_{ii}\right)\right]\left(\sigma^2 + (\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)^\top \boldsymbol{\Sigma}(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)\right).
$$

If I and $\widetilde{I}$ are two different independent subsets I_m, I_ℓ (corresponding to an off-diagonal term $m \leq \ell$ in (10)), then by the independence of $\widetilde{I}$ and $(V_{ii})_{i \in [n]}$ the concentration of sampling without replacement in Lemma 21 further gives $\frac{1}{|\widetilde{I}|}\sum_{i \in \widetilde{I}} V_{ii} \approx \frac{1}{n}\sum_{i=1}^n V_{ii} = 1 - \widehat{\mathsf{df}}/n$ so that the square bracket factor on the right-hand side simplifies to $n(-\widehat{\mathsf{df}}/n(1 - \widehat{\mathsf{df}}/n) + 1 - \widehat{\mathsf{df}}/n) = n(1 - \widehat{\mathsf{df}}/n)(1 -$

$\widehat{\mathsf{df}}/n$). This is only valid if I and $\widetilde{I}$ are independent. On the other hand, if $I = \widetilde{I}$, (corresponding to a diagonal term $m = \ell$ in (10)), then $\frac{1}{|\widetilde{I}|} \sum_{i \in \widetilde{I}} V_{ii} = 1 - \widehat{\mathsf{df}}/|I|$ and the square bracket above equals $n - 2\widehat{\mathsf{df}} + \widehat{\mathsf{df}}^2/k$. This means that a different correction needs to be used for the diagonal and off-diagonal terms. These approximations motivate the definition of the "full" estimator in (12).

The CGCV estimator. Next, we provide some heuristics behind the derivation of the corrected GCV from the full estimator. The key ingredients are the concentration of $|I_m \cap I_\ell|$ and $\widehat{\mathsf{df}}_m$ below (formally stated in Lemma 22 and Lemma 13, respectively):

$$|I_m \cap I_\ell| \approx \frac{|I_m||I_\ell|}{n} \quad \text{for all } m \neq \ell, \qquad \widehat{\mathsf{df}}_m \approx \widetilde{\mathsf{df}}_M = \frac{1}{M} \sum_{m' \in [M]} \widehat{\mathsf{df}}_{m'} \quad \text{for all } m.$$

By those concentration properties, we observe that $d_{m,\ell}^{\mathrm{full}}$, the denominator in the full-estimator defined in (19), has the following approximation depending on whether $m = \ell$ or not:

$$\frac{d_{m,\ell}^{\mathrm{full}}}{n} \approx \begin{cases} 1 - 2\dfrac{\widetilde{\mathsf{df}}_M}{n} + \left(\dfrac{\widetilde{\mathsf{df}}_M}{n}\right)^2 & \text{if } m \neq \ell \\[3mm] 1 - 2\dfrac{\widetilde{\mathsf{df}}_M}{n} + \left(\dfrac{\widetilde{\mathsf{df}}_M}{n}\right)^2 \dfrac{n}{|I_m|} & \text{if } m = \ell \end{cases} = \left(1 - \frac{\widetilde{\mathsf{df}}_M}{n}\right)^2 + \mathbb{1}_{\{m=\ell\}}\left(\frac{n}{|I_m|} - 1\right)\left(\frac{\widetilde{\mathsf{df}}_M}{n}\right)^2.$$

Combining the above display and $\widehat{R}_{m,\ell}^{\mathrm{full}} \approx R_{m,\ell}$ by (20), we have

$$\frac{\|\boldsymbol{y} - \boldsymbol{X}\widetilde{\boldsymbol{\beta}}_M\|_2^2}{n} = \frac{1}{M^2} \sum_{m,\ell} \frac{(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}_m)^\top(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}_\ell)}{n} = \frac{1}{M^2} \sum_{m,\ell} \frac{d_{m,\ell}^{\mathrm{full}}}{n} \cdot \widehat{R}_{m,\ell}^{\mathrm{full}}$$

$$\approx \left(1 - \frac{\widetilde{\mathsf{df}}_M}{n}\right)^2 R_M + \left(\frac{\widetilde{\mathsf{df}}_M}{n}\right)^2 \frac{1}{M^2} \sum_{m=1}^{M}\left(\frac{n}{|I_m|} - 1\right) R_{m,m}.$$

Dividing by $(1 - \widetilde{\mathsf{df}}_M/n)^2$ and plugging in $\widehat{R}_{m,m}^{\mathrm{ovlp}}$ or $\widehat{R}_{m,m}^{\mathrm{full}}$ to estimate $R_{m,m}$ inside the rightmost sum, we obtain the corrected GCV estimators given in Definition 1.

4.4 Numerical illustrations

In this section, we corroborate our theoretical results with numerical experiments on synthetic data, showing the validity of our proposed corrected GCV ($\widetilde{R}_M^{\mathrm{cgcv,full}}$ in (18)). The purpose is to empirically show our proposal, the corrected GCV estimator (shown as CGCV), accurately estimates the true risk R_M (shown as Risk), while the ordinary GCV in (9) (shown as GCV) is a poor estimator of the true risk R_M.

Linear models with Gaussian designs. We first describe the simulation setting. We set $(n, p) = (2000, 500)$ and consider different combinations of (k, M) and tuning parameters $(\lambda, \lambda_1, \lambda_2)$ for the three penalty functions in Table 1. We consider a linear model satisfying Assumptions B and C with noise variance $\sigma^2 = 1$. To be specific, we consider a non-isotropic covariance matrix $\boldsymbol{\Sigma}$, whose diagonal entries are evenly spaced in $[0.1, 10]$ and other entries are 0. We generate the true regression coefficient $\boldsymbol{\beta}_0$ as follows: generate $\boldsymbol{b}_0 \in \mathbb{R}^p$ with its first $p - 100$ rows sampled from i.i.d. $\mathcal{N}(0, 1)$, and set the remaining rows to 0. We then scale $\boldsymbol{b}_0$ to obtain the coefficient $\boldsymbol{\beta}_0$ by $\boldsymbol{\beta}_0 = \boldsymbol{b}_0(\sigma^2/\boldsymbol{b}_0^\top\boldsymbol{\Sigma}\boldsymbol{b}_0)^{1/2}$ so that the signal-to-noise ratio $\boldsymbol{\beta}_0^\top\boldsymbol{\Sigma}\boldsymbol{\beta}_0/\sigma^2$ is 1. For the ridge and lasso
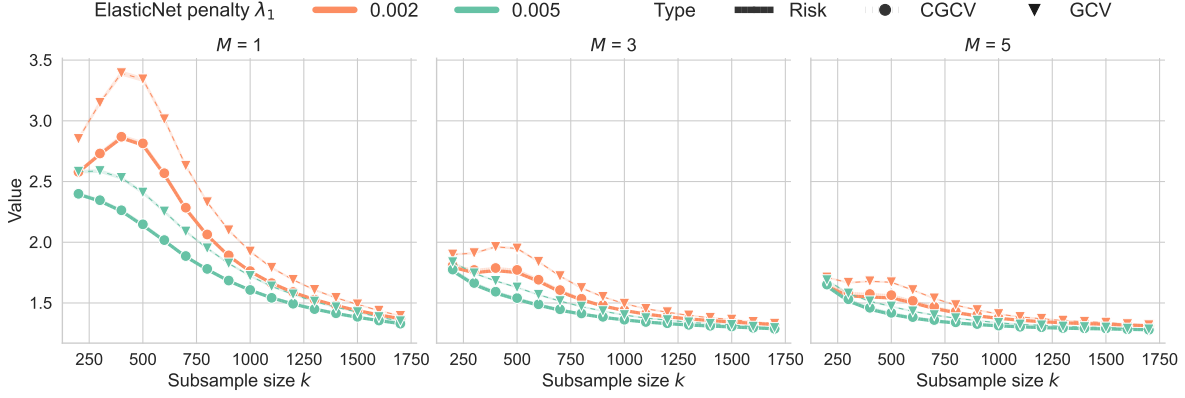
Figure 4: **CGCV is consistent for finite ensembles across different subsample sizes.** Plot of risks versus the subsample size k for Elastic Net ensemble with different λ and M and fixed $\lambda_2 = 0.01$.

ensembles, the tuning parameter λ of the corresponding penalty in Table 1 is fixed across all the subsample estimates $\widehat{\boldsymbol{\beta}}_m$. For the elastic net ensemble, the tuning parameter λ_2 is fixed as 0.01 and the same sequence of λ_1 is used across all subsample estimates $\widehat{\boldsymbol{\beta}}_m$. For each simulation setting, we perform 1000 dataset repetitions and report the average of relevant risks.

We present the simulation results of the elastic net ensemble in Figures 4 and 5, and leave the results of the ridge and lasso ensembles to the Appendix C of the supplement. Figure 4 shows the risk curve and its estimators as a function of the ensemble size M. We observe a clear gap between the curves of the GCV estimator and the true risk for different settings of (M, k, λ_1), especially for small subsample size k or small ensemble size M. In stark contrast, our proposed corrected GCV aligns closely with the true risk curve. Figure 5 again confirms that our proposal outperforms the GCV in estimating the true risk across different values of M and different combinations of k and λ_1. We provide more simulation results for the ridge and lasso ensembles in Appendix C.
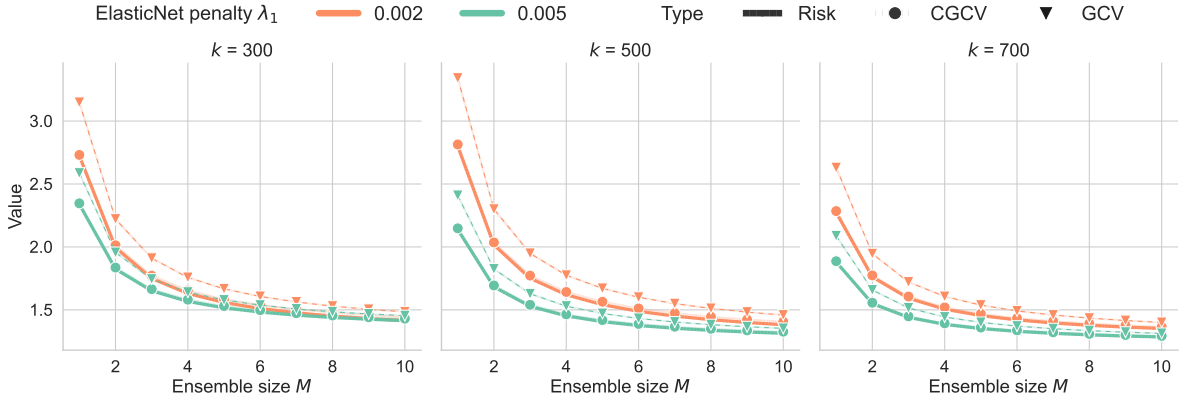

Figure 5: **GCV gets closer to the risk as ensemble size increases across different subsample sizes.** Plot of risks versus the Elastic Net ensemble size M for ridge ensemble with different λ and k and fixed $\lambda_2 = 0.01$.

Linear models with non-Gaussian designs. To demonstrate the robustness of our results beyond Gaussian distribution, we also consider scenarios where $\boldsymbol{X}$ is not Gaussian distributed but is sub-Gaussian, for example, the Rademacher and uniform distributions. The simulation results for

these non-Gaussian features closely resemble those for Gaussian features presented in the current section, confirming that the proposed CGCV accurately estimates the actual risk, whereas GCV exhibits a non-ignorable bias. We provide the simulation results for non-Gaussian features in Appendix D.

5   Asymptotic analysis for ridge ensembles

For ridge ensembles, the GCV estimator (8) for ridge ensembles can be inconsistent when the ensemble size is greater than 1, as mentioned in Section 2.2. This inconsistency still exists when features are sampled from non-Gaussian distributions (Du et al., 2023). In this section, we will show that the result of the previous section can be generalized to a wide class of data distributions when focusing on ridge predictors and using standard results from random matrix theory.

Recall the base estimator $\widehat{\boldsymbol{\beta}}_m$ defined in (2) for $m = 1, \ldots, M$. If we use the ridge penalty with regularization parameter $\lambda > 0$, then the base estimator reduces to the *ridge* estimator fitted on $\mathcal{D}_{I_m}$:

$$\widehat{\boldsymbol{\beta}}_{m,\lambda} = \underset{\boldsymbol{\beta} \in \mathbb{R}^p}{\operatorname{argmin}} \sum_{i \in I_m} (y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta})^2 / k + \lambda \|\boldsymbol{\beta}\|_2^2 = (\boldsymbol{X}^\top \boldsymbol{L}_{I_m} \boldsymbol{X} / k + \lambda \boldsymbol{I}_p)^{-1} \boldsymbol{X}^\top \boldsymbol{L}_{I_m} \boldsymbol{y} / k. \tag{26}$$

Taking $\lambda \to 0^+$, $\widehat{\boldsymbol{\beta}}_{m,0} := (\boldsymbol{X}^\top \boldsymbol{L}_{I_m} \boldsymbol{X} / |I_m|)^+ \boldsymbol{X}^\top \boldsymbol{L}_{I_m} \boldsymbol{y} / |I_m|$ becomes the so-called *ridgeless* estimator, or minimum-norm interpolator, where $\boldsymbol{A}^+$ denotes the Moore-Penrose inverse of matrix $\boldsymbol{A}$.

To prepare for our upcoming results, we impose the following assumptions on the responses, features, and subsample aspect ratios.

Assumption E (Response structure). The response y satisfies $\mathbb{E}[y] = 0$ and $\mathbb{E}[y^{4+\delta}] \leq C_0$ for some $\delta > 0$ and $C_0 > 0$.

Assumption F (Feature structure). The feature vector decomposes as $\boldsymbol{x} = \boldsymbol{\Sigma}^{1/2} \boldsymbol{z}$, where $\boldsymbol{z} \in \mathbb{R}^p$ contains i.i.d. entries with mean 0, variance 1, bounded moments of order $4 + \delta$ for some $\delta > 0$, and $\boldsymbol{\Sigma} \in \mathbb{R}^{p \times p}$ is deterministic and symmetric with eigenvalues uniformly bounded between $r_{\min} > 0$ and $r_{\max} < \infty$.

We make several remarks about the data assumptions in our analysis. First, we do not impose any specific model assumptions between the response variable y and the feature vector $\boldsymbol{x}$. Thus, our results are model-free, relying solely on the bounded moment assumption stated in Assumption E, which ensures the generality of our subsequent findings. Although we assume zero mean for y, this is purely for simplicity, as centering can always be achieved by subtracting the sample mean in practice. Furthermore, the condition of bounded moments can also be satisfied by imposing a stronger distributional assumption, such as sub-Gaussianity.

Assumption G (Proportional asymptotics). The sample size n, subsample size k, and feature dimension p tend to infinity such that $p/n \to \phi \in (0, \infty)$ and $p/k \to \psi \in [\phi, \infty]$.

Even though we assume that both the data aspect ratio p/n and the subsample aspect ratio p/k converge to fixed constants, this is for simplicity of exposition and proofs, and is not essential. It suffices only to require that they scale proportionally such that $0 < \liminf p/n \leq \liminf p/n \leq \liminf p/k < \infty$ as, by compactness, we can always extract a subsequence of regression problems in which p/n and p/k have a finite limit.

Both Assumption F, concerning the feature vector, and Assumption G, clarifying the proportional asymptotics regime, are assumptions commonly employed in the study of random matrix theory (Bai and Silverstein, 2010; El Karoui, 2010) and the analysis of ridge and ridgeless regression (Karoui and Kösters, 2011; El Karoui, 2013; Dobriban and Wager, 2018; Hastie et al., 2022).

Remark 9 (Extreme points in proportional asymptotics)**.** In Assumption G, we allow the limiting data aspect ratio p/n and subsample aspect ratio p/k to be in $(0, \infty)$ and $[\phi, \infty]$, respectively. In the extreme case where $p/k \to \infty$, the prediction risks of the ridge ensembles admit simple asymptotic formulas. More specifically, both the prediction risk and the estimates converge to the null risk $\mathbb{E}[y^2]$, the risk of a full predictor that always outputs zero.

5.1 Guarantees for intermediate risk estimators

Since the ridge predictor is a linear smoother, its degrees of freedom $\widehat{\mathsf{df}}_m$ defined in (3) is equivalent to the trace of its smoothing matrix $\boldsymbol{S}_m = \boldsymbol{X}_{I_m}(\boldsymbol{X}_{I_m}^\top \boldsymbol{X}_{I_m}/k + \lambda \boldsymbol{I}_p)^{-1}\boldsymbol{X}_{I_m}^\top/k$; namely, $\widehat{\mathsf{df}}_m = \mathrm{tr}[\boldsymbol{S}_m]$. For ridge regression, the ovlp-estimator $\widehat{R}_{m,\ell,\lambda}^{\mathrm{ovlp}}$ and the full-estimator $\widehat{R}_{m,\ell,\lambda}^{\mathrm{full}}$ as defined in (11) and (12) are respectively given explicitly as follows:

$$\widehat{R}_{m,\ell,\lambda}^{\mathrm{ovlp}} = \frac{(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}_{m,\lambda})^\top \boldsymbol{L}_m \boldsymbol{L}_\ell (\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}_{\ell,\lambda})/|I_m \cap I_\ell|}{1 - \dfrac{\mathrm{tr}[\boldsymbol{S}_m]}{|I_m|} - \dfrac{\mathrm{tr}[\boldsymbol{S}_\ell]}{|I_\ell|} + \dfrac{\mathrm{tr}[\boldsymbol{S}_m]}{|I_m|}\dfrac{\mathrm{tr}[\boldsymbol{S}_\ell]}{|I_\ell|}},$$

$$\widehat{R}_{m,\ell,\lambda}^{\mathrm{full}} = \frac{(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}_{m,\lambda})^\top (\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}_{\ell,\lambda})/n}{1 - \dfrac{\mathrm{tr}[\boldsymbol{S}_m]}{n} - \dfrac{\mathrm{tr}[\boldsymbol{S}_\ell]}{n} + \dfrac{\mathrm{tr}[\boldsymbol{S}_m]}{|I_m|}\dfrac{\mathrm{tr}[\boldsymbol{S}_\ell]}{|I_\ell|} \cdot \dfrac{|I_m \cap I_\ell|}{n}}.$$

Here, we use the subscript λ to indicate the dependence of the estimators for the risk component on the ridge penalty parameter λ. Similarly, we write $\widetilde{R}_{M,\lambda}^{\mathrm{ovlp}}$, $\widetilde{R}_{M,\lambda}^{\mathrm{full}}$, and $R_{M,\lambda}$ for the two risk estimators and the risk, respectively. The following theorem shows that both the two estimators are pointwise consistent for the risk component $R_{m,\ell,\lambda}$.

> **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 $|\widehat{R}_{m,\ell,\lambda}^{\mathrm{ovlp}} - R_{m,\ell,\lambda}| \xrightarrow{\text{a.s.}} 0$ and $|\widehat{R}_{m,\ell,\lambda}^{\mathrm{full}} - R_{m,\ell,\lambda}| \xrightarrow{\text{a.s.}} 0$. Consequently, it holds that $|\widetilde{R}_{M,\lambda}^{\mathrm{ovlp}} - R_{M,\lambda}| \xrightarrow{\text{a.s.}} 0$ and $|\widetilde{R}_{M,\lambda}^{\mathrm{full}} - R_{M,\lambda}| \xrightarrow{\text{a.s.}} 0$.

For the first estimator to work in the extreme case when $\psi = \infty$, we also need $\sqrt{n}/k = \mathcal{O}(1)$. We note that this is not required for the full-estimator precisely because the full-estimator uses both subsample observations and out-of-subsample observations. So, even if the number of overlapping subsample observations is small, the out-of-subsample observations are large, and the estimator is able to track the risk. The regime when k is small is the regime where the full-estimator is substantially better than the ovlp-estimator, as we have seen in Figure 3.

One natural question that remains for ridge predictors is how to select the ridge penalty λ. Building on the previous results, we are able to show a stronger notion of the consistency of the two proposed estimators over $\lambda \in \Lambda := [0, \infty]$. When $\lambda = \infty$, the ensemble ridge predictor reduces to a null predictor that always outputs zero. Under the above assumptions, uniform consistency is established through the following theorem.

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

The restriction on ψ is because both the risk and the risk estimate diverge when $\psi = 1$ for the ridgeless predictor when $\lambda = 0$. This divergence phenomenon has been studied in several recent works; see, e.g., Hastie et al. (2022).

5.2   Guarantees for corrected GCV

The average degrees of freedom (5) reduces to $\widetilde{\mathsf{df}}_M = M^{-1} \sum_{m=1}^{M} \mathrm{tr}(\boldsymbol{S}_m)$. The corrected GCV from (18) and (17) for ridge regression are given for $\# \in \{\mathrm{ovlp}, \mathrm{full}\}$ by:

$$\widehat{R}_{M,\lambda}^{\mathrm{cgcv},\#} := \widetilde{R}_{M,\lambda}^{\mathrm{gcv}} - \frac{1}{M} \left\{ \frac{(M^{-1} \sum_{m=1}^{M} \mathrm{tr}(\boldsymbol{S}_m)/n)^2}{(1 - M^{-1} \sum_{m=1}^{M} \mathrm{tr}(\boldsymbol{S}_m)/n)^2} \left(\frac{n}{k} - 1 \right) \frac{1}{M} \sum_{m=1}^{M} \widehat{R}_{m,m}^{\#} \right\}, \tag{27}$$

where $\widetilde{R}_{M,\lambda}^{\mathrm{gcv}} = (1 - M^{-1} \sum_{m=1}^{M} \mathrm{tr}(\boldsymbol{S}_m)/n)^{-2} \|\boldsymbol{y} - \boldsymbol{X}\widetilde{\boldsymbol{\beta}}_M\|_2^2/n$ is the GCV estimator and the correction term $\widehat{R}_{m,m,\lambda}^{\#}$ is given, for $\# \in \{\mathrm{ovlp}, \mathrm{full}\}$, by:

$$\widehat{R}_{m,m,\lambda}^{\mathrm{ovlp}} = \frac{\|\boldsymbol{y}_{I_m} - \boldsymbol{X}_{I_m}\widehat{\boldsymbol{\beta}}_{m,\lambda}\|_2^2/k}{(1 - M^{-1} \sum_{m=1}^{M} \mathrm{tr}(\boldsymbol{S}_m)/k)^2},$$

$$\widehat{R}_{m,m,\lambda}^{\mathrm{full}} = \frac{\|\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}_{m,\lambda}\|_2^2/n}{(1 - M^{-1} \sum_{m=1}^{M} \mathrm{tr}(\boldsymbol{S}_m)/n)^2 + (M^{-1} \sum_{m=1}^{M} \mathrm{tr}(\boldsymbol{S}_m)/n)^2 \cdot (n/k - 1)}.$$

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

Theorem 6 specializes the results of Theorem 2 for a general convex penalty to the ridge predictor. In particular, the correction term decreases as $1/M$. On the other hand, compared to Theorem 2, it also extends the corrected GCV to ridgeless predictors. Some remarks are as follows.

Remark 10 (Consistency of GCV under general models). The GCV estimator $\widetilde{R}_{M,\lambda}^{\mathrm{gcv}}$ is shown to be consistent for estimating the prediction risk R_M of the ridge predictors (Patil et al., 2021) (when $M = 1$ or $k = n$) and in the infinite ensemble (Du et al., 2023). Therein, the infinite ensemble is defined by letting the ensemble size M tend to infinity for any given sample size n and the feature dimension p. Since the correction term (27) vanishes as M tends to infinity, Theorem 6 above

25

also indirectly shows that the corrected GCV estimator is consistent in the infinite ensemble under more general data models. On the other hand, the correction term in (27) implies that the GCV estimator $\widetilde{R}_{M,\lambda}^{\mathrm{gcv}}$ is generally inconsistent for finite ensemble sizes when $n > k$ and $\widetilde{\mathsf{df}}_M > 0$. In the case of ridge regression, the latter condition is generally satisfied because for $\lambda \in [0, \infty)$ because

$$\widetilde{\mathsf{df}}_M = M^{-1} \sum_{m=1}^{M} \mathrm{tr}(\boldsymbol{S}_m) = M^{-1} \sum_{m=1}^{M} \mathrm{tr}[\boldsymbol{X}_{I_m}^{\top} \boldsymbol{X}_{I_m}/k(\boldsymbol{X}_{I_m}^{\top} \boldsymbol{X}_{I_m}/k + \lambda \boldsymbol{I}_p)^{-1}] > 0,$$

unless all the singular values of $\boldsymbol{X}_{I_m}$ are zero for all $m \in [M]$. Compared to Du et al. (2023, Proposition 3), which shows the inconsistency of $\widetilde{R}_{M,\lambda}^{\mathrm{gcv}}$ for $M = 2$ and $\lambda = 0$ under a well-specified linear model, our result is more general because it allows for any ridge parameter $\lambda \geq 0$, any ensemble size $M \geq 2$, and an arbitrary response model.

Remark 11 (Model selection). The ensemble ridge predictors involve three key parameters: the ridge penalty parameter λ, the subsample size k, and the ensemble size M. Questions related to the selection of the model of ridge ensembles over the subsample size k and the ensemble size M have already been discussed by Patil et al. (2023); Du et al. (2024). Tuning over M is relatively easy because the risk is decreasing in M (Patil et al., 2023). For tuning over λ, Theorem 6 implies that the tuned risk using the corrected GCV estimator converges to the optimal risk in a grid of ridge penalties used to fit the ensemble. In other words, the optimality of the data-dependent model selection can be guaranteed by the corrected GCV estimators.

Remark 12 (Non-asymptotic versus asymptotic analyses). When restricted to ridge estimators, the asymptotic analysis presented in the current section reveals some benefits compared to the finite-sample analysis in Section 4. First, the asymptotic analysis in the current section relaxes the assumptions on the data-generating process. More specifically, no explicit response-feature model is needed, except for the bounded moment assumptions and the feature structure assumed in Assumptions E and F. Second, the consistency provided in the current section does not require strong convexity as in Assumption D. It also applies to the ridgeless ensembles when $\lambda = 0$, in which case the base predictor reduces to the minimum ℓ_2-norm least squares. Last but not least, the asymptotic analysis enables the uniformity over $\lambda \in [0, \infty]$.

5.3 Proof outlines

There are three key steps involved in proving the uniform consistency (in λ) of the intermediate and corrected GCV estimator in Theorems 5 and 6, respectively, for ridge ensembles.

1. Deriving the asymptotic limit of the prediction risk: The first step involves deriving asymptotic equivalents for the risk of the ridge ensemble predictor. We build upon prior results on the risk analysis of ridge ensembles in Du et al. (2023); Patil and Du (2023) and derive these asymptotic risk equivalences, as done in Lemma 24.

2. Deriving the asymptotic limit of the intermediate and CGCV estimates: The second step involves deriving asymptotic equivalents for each of the proposed intermediate and corrected GCV estimators. Since each of these estimators is a ratio of terms that involve a version of training error and a denominator correction, we obtain each of these asymptotics separately. This is done in Lemmas 25–28.

3. Establishing the pointwise consistency in λ by matching the two limits and then lifting to uniform convergence in λ: In the last step, we show that the asymptotic limits obtained in the first two steps match with each other, which shows the pointwise consistency in λ

(Theorem 4). The matching uses the analytic properties of parameters related to certain fixed-point equations. To establish uniform convergence, we lift the pointwise consistency by showing that the family of functions under consideration forms an equicontinuous family and appeals to a stochastic version of the Arzela-Ascoli theorem. This is done in Lemma 29.

Throughout the proofs, the notion of asymptotic equivalence plays a crucial role (see Appendix B.1 for more details). We refer the reader to Patil et al. (2023); Patil and Du (2023) for background and further calculus of asymptotic equivalents. In our proofs, we build on these equivalents to obtain equivalents of various bias and variance components of the intermediate and corrected GCV estimators (see Appendix B.4 for more details). Finally, all our results are model-free. A component of this analysis involves an intermediate step showing quadratic concentration with uncorrelated components in Lemma 32 that extends the previous results of Bartlett et al. (2021, Lemma A.16) and Patil and Du (2023, Lemma D.3), which may be of independent interest.

5.4 Numerical illustrations: Non-linear models with non-Gaussian designs

To numerically evaluate our proposed corrected GCV estimator, we generate data from a nonlinear model $y = \boldsymbol{x}^\top \boldsymbol{\beta}_0 + (\|\boldsymbol{x}\|_2^2/p - 1) + \epsilon$, where $\boldsymbol{x} = \boldsymbol{\Sigma}^{-1/2}\boldsymbol{z}$ with $z_1, \ldots, z_p \overset{\text{i.i.d.}}{\sim} t_5$ follows a t-distribution with degrees of freedom 5 and $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_{\text{AR1},\rho=0.25} = (\rho^{|i-j|})_{i,j\in[p]}$ is the covariance matrix of the AR(1) process, $\boldsymbol{\beta}_0$ is the average of the top-5 eigenvectors of $\boldsymbol{\Sigma}$, and $\epsilon \sim \mathcal{N}(0,1)$. This setup has a linearized signal-to-noise ratio of 1.67. We set the sample size $n = 6000$ and the feature dimension $p = 1200$.

Our first experiment is to compare the naive GCV estimator $\widetilde{R}_M^{\text{gcv}}$ and the corrected GCV estimator $\widetilde{R}_M^{\text{cgcv,full}}$ as a function of the subsample size k, or equivalently, the subsample aspect ratio p/k. In Figure 6, we visualize the results for both the ridge and lasso predictors with different ensemble sizes M. For ridge predictors, we see that the naive GCV estimator is consistent when $M = 1$. However, it is extremely unstable when the subsample aspect ratio p/k is close to 1, which is also the interpolation threshold for the ridgeless predictor. For the lasso predictors, the naive GCV estimates are not even close to the prediction risks when the subsample size is small, even for ensemble size $M = 1$. Nevertheless, the corrected GCV estimates are in line with the true prediction risks for different ensemble sizes M. Although our theoretical results do not cover the lasso, the numerical results in the current subsection indicate the generality and robustness of our proposed corrected GCV estimator beyond ridge ensembles. On the other hand, as shown in Figure 7, the inconsistency of the naive GCV vanishes as the ensemble size M tends to infinity.

Finally, we compare the two estimators as a function of the ridge penalty λ in Figure 8. As we can see, for ensemble size $M = 1$, both estimators can track the true prediction risk well in either underparameterized or overparameterized regimes. However, when $M > 1$, the naive GCV estimator does not provide a valid estimate for the true risk. The inconsistency becomes more significant when either the subsample size k is small relative to the feature dimension p or the ridge penalty is close to zero. Similar observations are also present for the lasso estimators; see Figure 34. On the other hand, our corrected GCV estimator is uniformly consistent in various values of λ for all ensemble sizes M, as proved in Theorem 6.

6 Discussion

The primary goal of this paper has been to examine GCV for the risk estimation of ensembles of penalized estimators. Our main contribution is to show that ordinary GCV fails to be consistent for finite ensembles and to formulate a corrected GCV that is consistent for any finite ensemble.
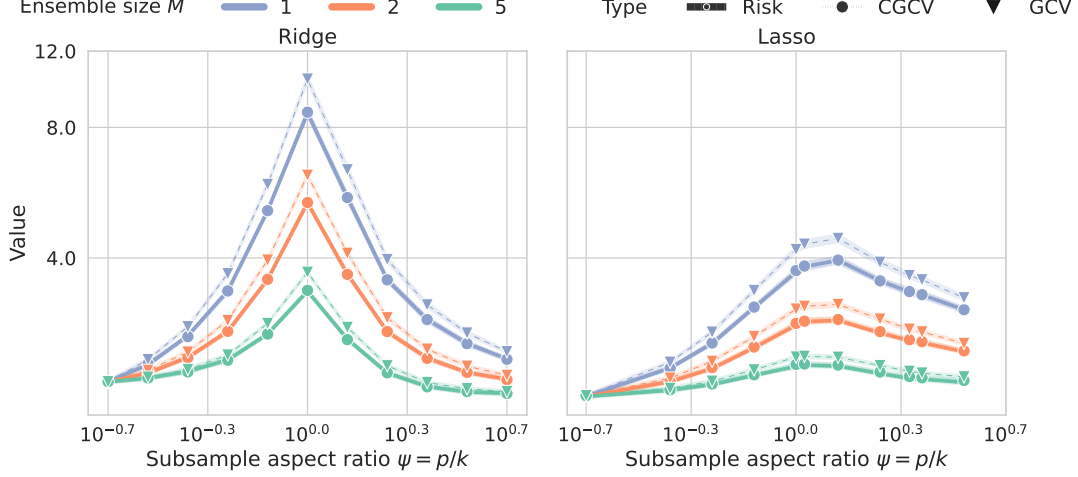
Figure 6: **CGCV is consistent for finite ensembles across different subsample aspect ratios.** The GCV estimates for the ridge and lasso ensembles with penalty $\lambda = 10^{-2}$ and varying subsample size in a problem setup of Section 5.4, over 50 repetitions of the datasets. The left panel shows the results for ridge predictors. The right panel shows the results for the lasso predictors.
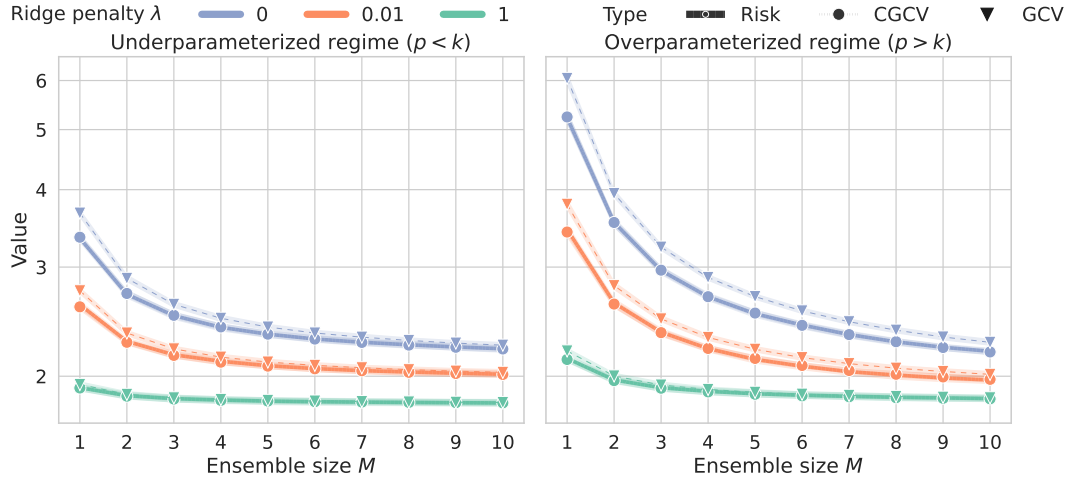

Figure 7: **GCV gets closer to the risk as ensemble size increases.** The GCV estimates for ridge ensemble with varying ensemble size in a problem setup of Section 5.4 over 50 repetitions of the datasets. Here, the feature dimension is $p = 1200$. The left panel shows the case when the subsample sizes are $k = 2400$ (an underparameterized regime). The right panel shows the case when the subsample sizes are $k = 800$ (an overparameterized regime).

The main high-level idea in constructing the correction is deriving term-by-term corrections for the ordinary GCV estimator, which subsequently leads to the form of corrected GCV. Assuming Gaussianity and linear model, we derived non-asymptotic bounds for our corrected GCV estimator that enjoys $\sqrt{n}$-consistency and provides explicit dependence on the subsample size k, ensemble size M, and the form and level of regularization penalty λ. These assumptions, although seemingly stringent, are amenable to a wide range of general convex penalized estimators. Moreover, for ensembles of ridge regression, we prove the asymptotic consistency of our estimator under very mild assumptions and, in particular, without assuming Gaussianity and linear response model.

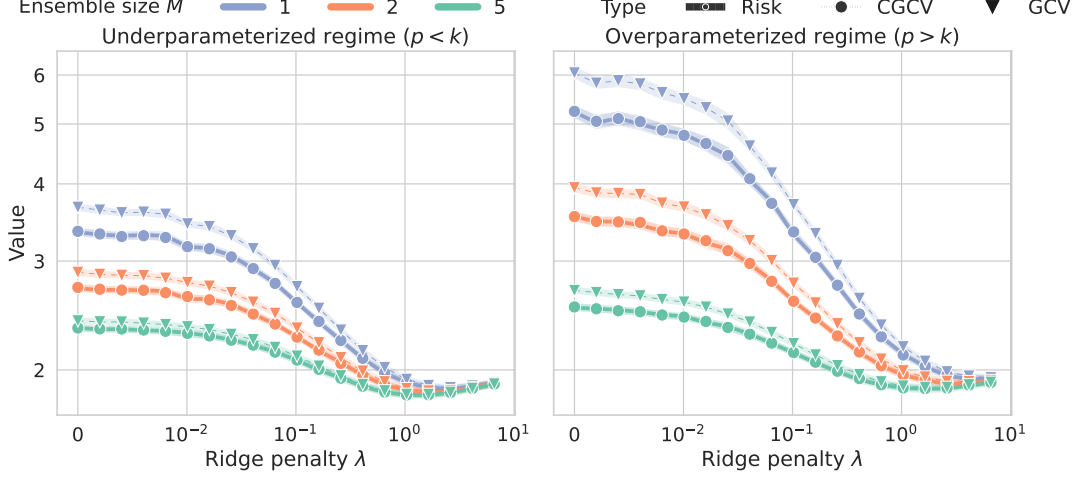As alluded to earlier, our intermediate risk estimators can incorporate several extensions and mod-

Figure 8: **CGCV is consistent for finite ensembles across different regularization levels.** The GCV estimates for ridge ensemble with varying ridge 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 panel shows the case when the subsample sizes are $k = 2400$ (an underparameterized regime). The right panel shows the case when the subsample sizes are $k = 800$ (an overparameterized regime).

ifications, thereby further broadening their applicability. First, we can use different subsample sizes for different ensemble estimators. Second, we can use different types of penalties or different regularization parameters for different base estimators. In such cases, the components of the ovlp-estimator (11) and the full-estimator (12) are still consistent. Note that additive bounds, as indicated in Theorems 1 and 2, already accommodate this flexibility. It is also possible to establish the multiplicative bounds assuming certain restrictions on the component risks. Finally, provided that the average cardinality concentrates and the cardinality of the intersection is lower-bounded with high probability, a slightly general form of CGCV is expected to work.

Although the techniques of the present paper can be applied to estimate the risk of hybrid ensembles (with the I_m having different cardinalities and the $\widehat{\boldsymbol{\beta}}_m$ having different penalty functions), it poses challenges when it comes to tuning the hyperparameters. Tuning an ensemble model, especially when equipped with a plethora of hyperparameters, is computationally challenging. In such cases, a greedy strategy often employed is to minimize the risk for each subsample estimator separately. However, it is important to note that the optimal regularization parameter λ for an individual ensemble component (i.e., for a fixed $m \in [M]$) may not be the best choice for the entire ensemble (i.e., after averaging over $m \in [M]$). This computational complexity of tuning ensembles becomes especially challenging, as ensemble components employ different penalties or regularization amounts. For this reason, we recommend training each individual estimate $\widehat{\boldsymbol{\beta}}_m$ with the same penalty and tuning parameter (say, λ) and selecting the tuning parameter λ for the ensemble average $\widetilde{\boldsymbol{\beta}}_M = M^{-1} \sum_{m=1}^{M} \widehat{\boldsymbol{\beta}}_m$ in (4) that yields the smallest corrected GCV criterion. Furthermore, one can also combine CGCV with the extrapolated CV method of Du et al. (2024) to tune over k and M.

Our work opens up several avenues for future investigation. We discuss three of them below.

Beyond squared loss for training. We have considered penalized estimators trained using squared loss in (2). One may consider penalized estimators trained on different loss functions, such as the Huber loss or other robust loss functions, especially if the noise random variables ϵ_i

are heavy-tailed and/or have infinite variance. Estimators of the squared risk in this context are obtained in Rad and Maleki (2020); Wang et al. (2018); Bellec (2023); Bellec and Shen (2022) for the non-ensemble setting, with the method in (Rad and Maleki, 2020; Wang et al., 2018) being applicable beyond the squared risk. It would be of interest to extend (with necessary modifications) our corrected GCV criterion to enable ensembles of base estimators $\widehat{\boldsymbol{\beta}}_m$ trained on robust loss functions. We leave this research direction for future work.

Beyond squared loss for testing. Our primary focus in this paper has been centered on evaluating the squared prediction risk. The special form of the squared risk allows the useful decomposition in (10), which subsequently leads to the estimators (11), (12) and corrected GCV. In practice, one may be interested in error metrics other than squared error. More broadly, one may be interested in functionals of the out-of-sample error distribution, such as quantiles of the error distribution. For regular penalized estimators, one can construct estimators for such functionals by constructing estimators for the out-of-sample distribution first and then subsequently using the plug-in functionals of these distributions. See Patil et al. (2022) for an example that uses GCV-based correction of the in-sample residuals. Another avenue for loss estimation beyond the square risk is (Rad and Maleki, 2020; Wang et al., 2018), wherein the proposed Approximate Leave-One-Out handles the non-ensemble setting. Whether one can further additively correct these residuals and construct consistent estimators of the out-of-sample distribution for the ensemble of penalized estimators is an interesting direction of future work.

Risk estimation for generic ensembles. The subsamples $I_1, \ldots, I_M$ are assumed to be sampled without replacement throughout this paper. It is of interest to investigate risk estimators for other ensemble techniques like bagging (where we sample with replacement) or other random weighting schemes, i.e., where the data-fitting loss in (2) is replaced by the loss $\sum_{i=1}^n w_{m,i}(y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta})^2$ for weights $(w_{m,i})_{i\in[n]}$ sampled independently of the data $(\boldsymbol{X}, \boldsymbol{y})$. This includes weights $w_{m,i} \overset{\text{i.i.d.}}{\sim}$ Poisson(1) or $(w_{m,1}, \ldots, w_{m,n}) \sim$ Multinomial(n, n, n^{-1}) typically used in the pair bootstrap; see El Karoui and Purdom (2018) for a study of such weights in unregularized estimation in the proportional regime.

Acknowledgements

References

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

Adlam, B. and Pennington, J. (2020b). Understanding double descent requires a fine-grained bias-variance decomposition. In *Advances in Neural Information Processing Systems*.

Ando, R. and Komaki, F. (2023). On high-dimensional asymptotic properties of model averaging estimators. *arXiv preprint arXiv:2308.09476*.

Auddy, A., Zou, H., Rad, K. R., and Maleki, A. (2023). Approximate leave-one-out cross validation for regression with ℓ_1 regularizers (extended version). *arXiv preprint arXiv:2310.17629*.

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

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

Bayati, M., Erdogdu, M. A., and Montanari, A. (2013). Estimating lasso risk and noise level. *Advances in Neural Information Processing Systems*, 26.

Bayati, M. and Montanari, A. (2011). The lasso risk for gaussian matrices. *IEEE Transactions on Information Theory*, 58(4):1997–2017.

Bellec, P. C. (2022). Observable adjustments in single-index models for regularized M-estimators. *arXiv preprint arXiv:2204.06990*.

Bellec, P. C. (2023). Out-of-sample error estimation for m-estimators with convex penalty. *Information and Inference: A Journal of the IMA*, 12(4):2782–2817.

Bellec, P. C. and Romon, G. (2021). Chi-square and normal inference in high-dimensional multi-task regression. *arXiv preprint arXiv:2107.07828*.

Bellec, P. C. and Shen, Y. (2022). Derivatives and residual distribution of regularized M-estimators with application to adaptive tuning. In *Conference on Learning Theory*.

Bellec, P. C. and Zhang, C.-H. (2021). Second-order stein: Sure for sure and other applications in high-dimensional inference. *The Annals of Statistics*, 49(4):1864–1903.

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

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

Breiman, L. (2001). Random forests. *Machine learning*, 45(1):5–32.

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

Celentano, M. and Montanari, A. (2021). Cad: Debiasing the lasso with inaccurate covariate model. *arXiv preprint arXiv:2107.14172*.

Celentano, M., Montanari, A., and Wei, Y. (2023). The lasso with general Gaussian designs with applications to hypothesis testing. *The Annals of Statistics*, 51(5):2194–2220.

Chaudhuri, A. (2014). *Modern Survey Sampling*. CRC Press.

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

Davidson, J. (1994). *Stochastic Limit Theory: An Introduction for Econometricians*. OUP Oxford.

Davidson, K. R. and Szarek, S. J. (2001). Local operator theory, random matrices and banach spaces. *Handbook of the geometry of Banach spaces*, 1(317-366):131.

Dietterich, T. G. (1998). An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting and randomization. *Machine learning*, 32:1–22.

Dietterich, T. G. (2000). Ensemble methods in machine learning. In *International workshop on multiple classifier systems*.

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

Dossal, C., Kachour, M., Fadili, M., Peyré, G., and Chesneau, C. (2013). The degrees of freedom of the lasso for general design matrix. *Statistica Sinica*, pages 809–828.

Du, J.-H., Patil, P., and Kuchibhotla, A. K. (2023). Subsample ridge ensembles: Equivalences and generalized cross-validation. In *International Conference on Machine Learning*.

Du, J.-H., Patil, P., Roeder, K., and Kuchibhotla, A. K. (2024). Extrapolated cross-validation for randomized ensembles. *Journal of Computational and Graphical Statistics*, pages 1–12.

El Karoui, N. (2010). The spectrum of kernel random matrices. *The Annals of Statistics*, 38(1):1–50.

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

El Karoui, N. and Purdom, E. (2018). Can we trust the bootstrap in high-dimensions? The case of linear models. *The Journal of Machine Learning Research*, 19(1):170–235.

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

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

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

Han, Q. and Xu, X. (2023). The distribution of ridgeless least squares interpolators. *arXiv preprint arXiv:2307.02044*.

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

Hastie, T., Tibshirani, R., and Friedman, J. H. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer. Second edition.

Karoui, N. E. and Kösters, H. (2011). Geometric sensitivity of random matrix results: Consequences for shrinkage estimators of covariance and related statistical methods. *arXiv preprint arXiv:1105.1404*.

Knowles, A. and Yin, J. (2017). Anisotropic local laws for random matrices. *Probability Theory and Related Fields*, 169:257–352.

Leeb, H. (2008). Evaluation and selection of models for out-of-sample prediction when the sample size is small relative to the complexity of the data-generating process. *Bernoulli*, pages 661–690.

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

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

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

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

Lopes, M. E. (2019). Estimating the algorithmic variance of randomized ensembles via the bootstrap. *The Annals of Statistics*, 47(2):1088–1112.

Lopes, M. E., Wu, S., and Lee, T. C. (2020). Measuring the algorithmic convergence of randomized ensembles: The regression setting. *SIAM Journal on Mathematics of Data Science*, 2(4):921–943.

Loureiro, B., Gerbelot, C., Cui, H., Goldt, S., Krzakala, F., Mezard, M., and Zdeborová, L. (2021). Learning curves of generic features maps for realistic datasets with a teacher-student model. *Advances in Neural Information Processing Systems*, 34:18137–18151.

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

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

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

Mücke, N., Reiss, E., Rungenhagen, J., and Klein, M. (2022). Data-splitting improves statistical performance in overparameterized regimes. In *International Conference on Artificial Intelligence and Statistics*.

Patil, P. and Du, J.-H. (2023). Generalized equivalences between subsampling and ridge regularization. *Thirty-seventh Conference on Neural Information Processing Systems*.

Patil, P., Du, J.-H., and Kuchibhotla, A. K. (2023). Bagging in overparameterized learning: Risk characterization and risk monotonization. *Journal of Machine Learning Research*, 24(319):1–113.

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

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

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

Stein, C. M. (1981). Estimation of the mean of a multivariate normal distribution. *The Annals of Statistics*, 9(6):1135–1151.

Tan, K. and Bellec, P. C. (2023). Multinomial logistic regression: Asymptotic normality on null covariates in high-dimensions. *arXiv preprint arXiv:2305.17825*.

Tan, K., Romon, G., and Bellec, P. C. (2022). Noise covariance estimation in multi-task high-dimensional linear models. *arXiv preprint arXiv:2206.07256*.

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

Tibshirani, R. J. and Taylor, J. (2012). Degrees of freedom in lasso problems. *The Annals of Statistics*, 40(2):1198–1232.

Vaiter, S., Deledalle, C., Fadili, J., Peyré, G., and Dossal, C. (2017). The degrees of freedom of partly smooth regularizers. *Annals of the Institute of Statistical Mathematics*, 69:791–832.

Vaiter, S., Deledalle, C., Peyré, G., Fadili, J., and Dossal, C. (2012). The degrees of freedom of the group lasso for a general design. *arXiv preprint arXiv:1212.6478*.

Wang, S., Weng, H., and Maleki, A. (2020). Which bridge estimator is the best for variable selection? *The Annals of Statistics*, 48(5):2791 – 2823.

Wang, S., Zhou, W., Maleki, A., Lu, H., and Mirrokni, V. (2018). Approximate leave-one-out for high-dimensional non-differentiable learning problems. *arXiv preprint arXiv:1810.02716*.

Wasserman, L. (2006). *All of Nonparametric Statistics*. Springer.

Waterhouse, W. C. (1983). Do symmetric problems have symmetric solutions? *The American Mathematical Monthly*, 90(6):378–387.

Wei, A., Hu, W., and Steinhardt, J. (2022). More than a toy: Random matrix models predict how real-world neural representations generalize. In *International Conference on Machine Learning*.

Wolpert, D. H. (1992). Stacked generalization. *Neural networks*, 5(2):241–259.

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

Zou, H. and Hastie, T. (2007). On the degrees of freedom of the lasso. *The Annals of Statistics*, 35(5):2173–2192.

<div style="text-align: center;">

Supplement

</div>

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 2.

Organization

The content of this appendix is organized as follows.

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

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

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

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

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

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

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

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

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

Section	Content	Purpose
Appendix 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 2: 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 $\{1, \ldots, n\}$ for a natural number n
$\|I\|$	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 $\boldsymbol{v}$
$\|\boldsymbol{u}\|_p$	The ℓ_p norm of a vector $\boldsymbol{u}$ for $p \geq 1$
$\|f\|_{L_p}$	The L_p norm of a function f for $p \geq 1$
$\|\boldsymbol{A}\|_{\mathrm{op}}$	Operator (or spectral) norm of a real matrix $\boldsymbol{A}$
$\|\boldsymbol{A}\|_{\mathrm{tr}}$	Trace (or nuclear) norm of a real matrix $\boldsymbol{A}$
$\|\boldsymbol{A}\|_F$	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
$X = \mathcal{O}_\alpha(Y)$	$\|X\| \leq C_\alpha Y$ 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 Appendix B for more details)
$\overset{\mathrm{d}}{=}$	Equality in distribution
$\overset{\mathrm{d}}{\to}$	Denotes convergence in distribution
$\overset{\mathrm{p}}{\to}$	Denotes convergence in probability
$\overset{\mathrm{a.s.}}{\longrightarrow}$	Denotes almost sure convergence

Note: 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 be the ℓ_2 norm of $\boldsymbol{u}$. If a proof of a statement is separated from the statement, the statement is restated (while keeping the original numbering) along with the proof for convenience.

A   Proofs in Section 4

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 $|I_m| = |I_\ell| = k \leq n$, 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.,

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

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 Appendix A:

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

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

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

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:

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

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 Appendix 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:
>
> $$\mathbb{E}\left[\left|\frac{d_{m,\ell}^{\text{ovlp}}}{n} \cdot \frac{\widehat{R}_{m,\ell}^{\text{ovlp}} - R_{m,\ell}}{\sqrt{R_{m,m}R_{\ell,\ell}}}\right|\right] \leq C\frac{\gamma^{7/2}}{\tau^2 c^3 \sqrt{n}}, \qquad \mathbb{E}\left[\left|\frac{d_{m,\ell}^{\text{full}}}{n} \cdot \frac{\widehat{R}_{m,\ell}^{\text{full}} - R_{m,\ell}}{\sqrt{R_{m,m}R_{\ell,\ell}}}\right|\right] \leq C\frac{\gamma^{5/2}}{\tau^2 c^2 \sqrt{n}}. \quad (20)$$
>
> 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
>
> $$\mathbb{P}\left(\left|1 - \frac{\widetilde{R}_M^{\text{ovlp}}}{R_M}\right| > \epsilon\right) \leq C\frac{M^3 \gamma^{11/2}}{\epsilon \tau^4 c^7 \sqrt{n}}, \qquad \mathbb{P}\left(\left|1 - \frac{\widetilde{R}_M^{\text{full}}}{R_M}\right| > \epsilon\right) \leq C\frac{M^3 \gamma^{9/2}}{\epsilon \tau^4 c^2 \sqrt{n}}. \quad (21)$$
>
> Thus, if $(M, \mu^{-1}, p/n, n/k)$ are bounded from above by a constant independent of n, we have
>
> $$\widetilde{R}_M^{\text{ovlp}}/R_M = 1 + \mathcal{O}_{\mathbb{P}}(n^{-1/2}), \qquad \widetilde{R}_M^{\text{full}}/R_M = 1 + \mathcal{O}_{\mathbb{P}}(n^{-1/2}).$$

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

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

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

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

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

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

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

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

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

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

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

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 $(\text{RHS})^2$ given $(|I \cap \widetilde{I}|, I, \boldsymbol{V})$ is bounded from above as

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

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

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

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

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

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

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

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

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

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

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 Appendix 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 $\boldsymbol{L}_J\boldsymbol{r}\boldsymbol{h}^\top \in \mathbb{R}^{n \times (p+1)}$.

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

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

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

Control of Rem_3. Note

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

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

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

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

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

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 $e^x \ge \sqrt{x/2}$ 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 $\xi_1 = (1+\mathrm{tr}[\boldsymbol{B}])\|\boldsymbol{Lr}\|^2/\|\boldsymbol{h}\|^2-\mathrm{tr}[\boldsymbol{LV}]$ and $\xi_2 = k-(1+\mathrm{tr}[\boldsymbol{B}])^2\|\boldsymbol{Lr}\|^2/\|\boldsymbol{h}\|^2$. Then
>
> $$\mathbb{E}|k - \mathrm{tr}[\boldsymbol{LV}](1 + \mathrm{tr}[\boldsymbol{B}])| = \mathbb{E}|(1 + \mathrm{tr}[\boldsymbol{B}])\xi_1 + \xi_2| \lesssim \sqrt{k}\tau^{-2}(1+p/k)^{5/2}.$$

Lemma 9 is proved in Appendix 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

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

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

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

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

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

Proof for the full-*estimator*. Using the notation in Appendix 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:

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

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

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. $\qquad\square$

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

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

$$
\left| \frac{\widetilde{R}_M^\#}{R_M} - 1 \right| = \frac{|\sum_{m,\ell} (\widehat{R}_{m,\ell}^\# - \boldsymbol{h}_m^\top \boldsymbol{h}_\ell)|}{\|\sum_{m'} \boldsymbol{h}_{m'}\|^2} \leq \frac{(\sum_{m'} \|\boldsymbol{h}_{m'}\|)^2}{\|\sum_{m'} \boldsymbol{h}_{m'}\|^2} \cdot \sum_{m,\ell} \frac{\|\boldsymbol{h}_m\|\|\boldsymbol{h}_\ell\|}{(\sum_{m'} \|\boldsymbol{h}_{m'}\|)^2} \cdot \frac{|\widehat{R}_{m,\ell}^\# - \boldsymbol{h}_m^\top \boldsymbol{h}_\ell|}{\|\boldsymbol{h}_m\|\|\boldsymbol{h}_\ell\|}
$$

$$
\leq \frac{(\sum_{m'} \|\boldsymbol{h}_{m'}\|)^2}{\|\sum_{m'} \boldsymbol{h}_{m'}\|^2} \cdot \max_{m,\ell} \frac{|\widehat{R}_{m,\ell}^\# - \boldsymbol{h}_m^\top \boldsymbol{h}_\ell|}{\|\boldsymbol{h}_m\|\|\boldsymbol{h}_\ell\|} \leq M \cdot \frac{\sum_{m'} \|\boldsymbol{h}_{m'}\|^2}{\|\sum_{m'} \boldsymbol{h}_{m'}\|^2} \cdot \max_{m,\ell} \frac{|\widehat{R}_{m,\ell}^\# - \boldsymbol{h}_m^\top \boldsymbol{h}_\ell|}{\|\boldsymbol{h}_m\|\|\boldsymbol{h}_\ell\|}
$$

$$
= M \cdot \mathrm{Ratio} \cdot \max_{m,\ell} \frac{n}{|d_{m,\ell}^\#|} |E_{m,\ell}^\#|. \tag{37}
$$

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:

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

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

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

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

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

Lemma 10 is proved in Appendix 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

$$
\sum_m \|\boldsymbol{h}_m\|^2 - \|\sum_m \boldsymbol{h}_m\|^2 = \sum_{m \neq \ell} (-\boldsymbol{h}_m^\top \boldsymbol{h}_\ell + \eta_{m,\ell}\|\boldsymbol{h}_m\|\|\boldsymbol{h}_\ell\| - \eta_{m,\ell}\|\boldsymbol{h}_m\|\|\boldsymbol{h}_\ell\|) \leq \sum_{m \neq \ell} U_{m,\ell}\|\boldsymbol{h}_m\|\|\boldsymbol{h}_\ell\|
$$

$$
\leq \left(\sum_{m \neq \ell} U_{m\ell}^2 \right)^{1/2} \left(\sum_{m \neq \ell} \|\boldsymbol{h}_m\|^2\|\boldsymbol{h}_\ell\|^2 \right)^{1/2} \leq \left(\sum_{m,\ell} U_{m\ell}^2 \right)^{1/2} \sum_m \|\boldsymbol{h}_m\|^2.
$$

43

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

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

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

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

$$(41)$$

which gives the upper bound of $\mathrm{Ratio} = \sum_m \|\boldsymbol{h}_m\|^2/\|\sum_m \boldsymbol{h}_m\|^2$.

Control of $|d_{m,\ell}^{\#}|^{-1}$. The lemma below gives an lower bound of $d_{m,\ell}^{\#}$ for $\# \in \{\mathrm{ovlp}, \mathrm{full}\}$.

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

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

Lemma 11 is proved in Appendix 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 $\epsilon > 0$:

$$\mathbb{P}\Big(\Big|\frac{\widetilde{R}_M^{\#}}{R_M} - 1\Big| > \epsilon\Big) \leq \mathbb{P}\Big(\mathrm{Ratio} \cdot \max_{m,\ell} \frac{n}{|d_{m,\ell}^{\#}|} |E_{m,\ell}^{\#}| > \frac{\epsilon}{M}\Big)$$

$$\lesssim \frac{M^2\gamma}{n\tau^2 c^2} + \mathbb{P}\Big(2\max_{m,\ell} \frac{n}{|d_{m,\ell}^{\#}|} |E_{m,\ell}^{\#}| > \frac{\epsilon}{M}\Big)$$

$$\leq \frac{M^2\gamma}{n\tau^2 c^2} + \sum_{m,\ell} \mathbb{P}\Big(\frac{n}{|d_{m,\ell}^{\#}|} |E_{m,\ell}^{\#}| > \frac{\epsilon}{2M}\Big)$$

$$\leq \frac{M^2\gamma}{n\tau^2 c^2} + \sum_{m,\ell} \Big\{\mathbb{P}\Big(\frac{d_{m,\ell}^{\#}}{n} < Cd^{\#}(\tau, c, \gamma)\Big) + \mathbb{P}\Big(|E_{m,\ell}^{\#}| > \frac{\epsilon Cd^{\#}(\tau, c, \gamma)}{2M}\Big)\Big\}$$

$$\lesssim \frac{M^2\gamma}{n\tau^2 c^2} + M^2 \Big[\frac{\gamma^4}{nc^2\tau^4} + \frac{2M}{\epsilon Cd^{\#}(\tau, c, \gamma)} \frac{1}{\sqrt{n}\tau^2} \cdot \begin{cases} c^{-3}\gamma^{7/2} & \# = \mathrm{ovlp} \\ c^{-2}\gamma^{5/2} & \# = \mathrm{full}, \end{cases}\Big]$$

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

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

$$\lesssim \frac{M^2\gamma^4}{nc^2\tau^4} + \frac{M^3}{\epsilon\sqrt{n}\tau^4} \times \begin{cases} c^{-7}\gamma^{11/2} & \# = \mathrm{ovlp} \\ c^{-2}\gamma^{9/2} & \# = \mathrm{full} \end{cases}, \qquad (42)$$

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

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

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

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

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:

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

Equation (45) is proved in Appendix 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 Appendix 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

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

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

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 $\widetilde{R}^{\mathrm{cgcv},\#}$ in (43) and $\widetilde{R}^{\mathrm{cgcv}}$ in (44),

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

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

$$\left|\frac{\widetilde{R}_M^{\mathrm{cgcv}} - \widetilde{R}_M^{\mathrm{cgcv},\#}}{R_M}\right| \leq \frac{c}{(1 - \widetilde{\mathsf{df}}_M/n)^2} \frac{\sum_m \|\boldsymbol{h}_m\|^2}{\|\sum_m \boldsymbol{h}_m\|^2} \max_m \left|\frac{\widehat{R}_{mm}^{\#}}{\|\boldsymbol{h}_m\|^2} - 1\right| = c \cdot U \cdot \max_m V_m^{\#},$$

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

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

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

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

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

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

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

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

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 $\epsilon \in (0,1)$,

$$\mathbb{P}\left(\left|\frac{\widetilde{R}_M^{\mathrm{cgcv}} - \widetilde{R}_M^{\mathrm{cgcv},\#}}{R_M}\right| > \epsilon\right) \leq \mathbb{P}(cU \max_m V_m^{\#} > \epsilon) \leq \mathbb{P}\left(U > \frac{2\gamma^2}{C^2\tau^2}\right) + \mathbb{P}\left(c\frac{2\gamma^2}{C^2\tau^2} \max_m V_M^{\#} > \epsilon\right)$$

$$\lesssim \frac{M^2\gamma^2}{n\tau^2 c^2} + \sum_m \mathbb{P}\Big(V_m^\# > \frac{\epsilon C^2\tau^2}{2c\gamma^2}\Big) \ \text{(by (47))}$$

$$\lesssim \frac{M^2\gamma^2}{n\tau^2 c^2} + \frac{M\gamma^4}{n\tau^4 c^2} + \frac{M}{\sqrt{n}\tau^4}\Big(\frac{\epsilon C^2\tau^2}{2c\gamma^2}\Big)^{-1} \cdot \begin{cases} c^{-7}\gamma^{11/2} & \# = \text{ovlp} \\ c^{-2}\gamma^{9/2} & \# = \text{full} \end{cases} \ \text{(by (48))}$$

$$\lesssim \frac{M^2\gamma^4}{n\tau^4 c^2} + \frac{M}{\sqrt{n}\epsilon\tau^6} \times \begin{cases} c^{-6}\gamma^{15/2} & \# = \text{ovlp} \\ c^{-1}\gamma^{13/2} & \# = \text{full} \end{cases} \ \text{(thanks to } \tau \in (0,1] \text{ and } \gamma \geq 1)$$

$$\lesssim \frac{M^2}{\sqrt{n}\epsilon\tau^6} \times \begin{cases} c^{-6}\gamma^{15/2} & \# = \text{ovlp} \\ c^{-2}\gamma^{13/2} & \# = \text{full} \end{cases} \ \text{(thanks to } \epsilon \in (0,1))$$

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

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

$$\leq M \cdot \frac{1}{(1 - \widetilde{\mathsf{df}}_M/n)^2} \frac{\sum_m \|\boldsymbol{h}_m\|^2}{\|\sum_m \boldsymbol{h}_m\|^2} \cdot \max_{m,\ell}\Big(\frac{|\boldsymbol{r}_m^\top \boldsymbol{r}_\ell - d_{m,\ell}^{\text{full}} \boldsymbol{h}_m^\top \boldsymbol{h}_\ell|}{n\|\boldsymbol{h}_m\|\|\boldsymbol{h}_\ell\|} + \frac{|d_{m,\ell}^{\text{cgcv}} - d_{m,\ell}^{\text{full}}|}{n}\Big)$$

$$= M \cdot U \cdot \max_{m,\ell}\Big(|E_{m,\ell}^{\text{full}}| + |d_{m,\ell}^{\text{cgcv}} - d_{m,\ell}^{\text{full}}|/n\Big),$$

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 $\# = \text{full}$ and (47): there exists a absolute constant C such that

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

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

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

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

$$\mathbb{P}\Big(\big|\frac{\widetilde{R}_M^{\mathrm{cgcv}}}{R_M} - 1\big| > \epsilon\Big) \leq \mathbb{P}\Big(U \cdot \max_{m,\ell}(|E_{m,\ell}^{\mathrm{full}}| + \frac{|d_{m,\ell}^{\mathrm{cgcv}} - d_{m,\ell}^{\mathrm{full}}|}{n}) > \frac{\epsilon}{M}\Big)$$

$$\leq \mathbb{P}\Big(U > \frac{C\gamma^2}{\tau^2}\Big) + \mathbb{P}\Big(\max_{m,\ell}(|E_{m,\ell}^{\mathrm{full}}| + \frac{|d_{m,\ell}^{\mathrm{cgcv}} - d_{m,\ell}^{\mathrm{full}}|}{n}) > \frac{\epsilon}{M}\frac{\tau^2}{C\gamma^2}\Big)$$

$$\lesssim \frac{M^2\gamma^2}{n\tau^2 c^2} + \sum_{m,\ell}\mathbb{P}\Big(|E_{m,\ell}^{\mathrm{full}}| > \frac{\epsilon\tau^2}{2CM\gamma^2}\Big) + \mathbb{P}\big(\frac{|d_{m,\ell}^{\mathrm{cgcv}} - d_{m,\ell}^{\mathrm{full}}|}{n} > \frac{\epsilon\tau^2}{2CM\gamma^2}\big)$$

$$\lesssim \frac{M^2\gamma^2}{n\tau^2 c^2} + M^2\Big\{\frac{2CM\gamma^2}{\epsilon\tau^2} \cdot \frac{\gamma^{5/2}}{\sqrt{n}\tau^2 c^2} + Me^{-nc/2} + \frac{2CM\gamma^2}{\epsilon\tau^2} \cdot \frac{M\gamma^{9/2}}{\sqrt{n}\tau^3 c^4}\Big\}$$

$$\lesssim \frac{M^4\gamma^{13/2}}{\sqrt{n}\epsilon\tau^5 c^4} \quad \text{(thanks to } \epsilon, c, \tau \in (0, 1] \text{ and } \gamma \geq 1\text{),}$$

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

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

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

Below we prove $d_{m,\ell}^{\mathrm{full}}/n \approx d_{m,\ell}^{\mathrm{cgcv}}/n$. 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

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

The proof of Lemma 13 is given in Appendix A.5.6. From Lemma 13, it suffices to bound $|d_{m,\ell}^{\mathrm{full}} - d_{m,\ell}^{\mathrm{cgcv}}|/n$ 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

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

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 $\epsilon > 0$,

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

48

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

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

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

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

$$\leq c^2\Big|\frac{|I_m \cap I_\ell| \cdot n}{k^2} - 1\Big| + \frac{|\widehat{\mathsf{df}}_m - \widetilde{\mathsf{df}}_M|}{n} + \frac{|\widehat{\mathsf{df}}_\ell - \widetilde{\mathsf{df}}_M|}{n}.$$

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

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

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

$$\mathbb{P}\Big(\frac{|d_{m,\ell}^{\mathrm{full}} - d_{m,\ell}^{\mathrm{cgcv}}|}{n} > \epsilon\Big) \leq \mathbb{P}\Big(c^2\Big|\frac{|I_m \cap I_\ell| \cdot n}{k^2} - 1\Big| > \frac{\epsilon}{3}\Big) + \mathbb{P}\Big(\frac{|\widehat{\mathsf{df}}_m - \widetilde{\mathsf{df}}_M|}{n} > \frac{\epsilon}{3}\Big) + \mathbb{P}\Big(\frac{|\widehat{\mathsf{df}}_\ell - \widetilde{\mathsf{df}}_M|}{n} > \frac{\epsilon}{3}\Big)$$

$$\lesssim \frac{c}{\epsilon\sqrt{n}} + M\Big(e^{-nc/2} + \frac{\gamma^{9/2}}{\epsilon\sqrt{n}\tau^3 c^4}\Big) \lesssim M\Big(e^{-nc/2} + \frac{\gamma^{9/2}}{\epsilon\sqrt{n}\tau^3 c^4}\Big),$$

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

A.4 Proof of Theorem 3

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

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

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

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

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

$$\mathbb{P}\Big(\frac{\widetilde{\mathsf{df}}_M}{k} \geq \delta\Big) \leq \mathbb{P}\Big(\mathrm{Correction} \geq \frac{\delta^2 c(1-c)}{M}\Big)$$

$$\leq \mathbb{P}\left(\frac{\widetilde{R}_M^{\text{gcv}}}{R_M} - \text{Correction} - 1 \leq -\frac{\delta^2 c(1-c)}{2M}\right) + \mathbb{P}\left(\frac{\widetilde{R}_M^{\text{gcv}}}{R_M} - 1 > \frac{\delta^2 c(1-c)}{2M}\right)$$

$$\leq C\left(\frac{\delta^2 c(1-c)}{2M}\right)^{-1}\frac{M^4\gamma^{13/2}}{\sqrt{n}\tau^5 c^4} + \mathbb{P}\left(\frac{\widetilde{R}_M^{\text{gcv}}}{R_M} > 1 + \frac{c(1-c)\delta^2}{2M}\right) \text{ (by (51) with } \epsilon = \frac{\delta^2 c(1-c)}{2M})$$

$$\leq C'\frac{M^5}{\sqrt{n}\tau^5\delta^2}\frac{\gamma^{13/2}}{c^5(1-c)} + \mathbb{P}\left(\frac{\widetilde{R}_M^{\text{gcv}}}{R_M} > 1 + \frac{c(1-c)\delta^2}{2M}\right),$$

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

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

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:

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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 $\|\boldsymbol{B}\|_F^2 \leq \mathrm{rank}(\boldsymbol{B})\|\boldsymbol{B}\|_{\mathrm{op}}^2 \leq p(k\mu)^{-2}$ 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

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

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. $\qquad\square$

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

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

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

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

where $\mathbb{E}[\|\boldsymbol{\varrho}\|^2\|\widetilde{\boldsymbol{\eta}}\|^2] \leq \mathbb{E}[\|\boldsymbol{L}_J\boldsymbol{Z}\|_{\mathrm{op}}^2]$ thanks to $\|\widetilde{\boldsymbol{\eta}}\|^2 = 1$ and $\|\boldsymbol{\varrho}\|^2 \leq \|\boldsymbol{L}_J\boldsymbol{Z}\|_{\mathrm{op}}^2$. 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 $\|\boldsymbol{h}\|^2 \neq 0$, we have

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

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

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

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

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

Therefore, the RHS of (53) is bounded from above by $|J| + \mathbb{E}[\|\boldsymbol{L}_J\boldsymbol{Z}\|_{\mathrm{op}}^2(1 + \Xi_J + \widetilde{\Xi}_J)]$ 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 $2\|\boldsymbol{L}_J\boldsymbol{Z}\|_{\mathrm{op}}^2(\Xi_J + \widetilde{\Xi}_J)$. This finishes the proof. $\qquad\square$

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

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

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

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

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

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

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

This finishes the proof. $\qquad\square$

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

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

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

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

$$\mathbb{E}[|k - \|\boldsymbol{a}\|^2|] \leq \mathbb{E}[2(\sqrt{k}\|\boldsymbol{a} - \boldsymbol{b}\| + |k - \|\boldsymbol{b}\|^2| + \|\boldsymbol{a} - \boldsymbol{b}\|^2]$$
$$\lesssim \sqrt{k\mathbb{E}[\|\boldsymbol{a} - \boldsymbol{b}\|^2]} + \mathbb{E}[\|\boldsymbol{a} - \boldsymbol{b}\|^2] + \mathbb{E}[|k - \|\boldsymbol{b}\|^2|]. \tag{55}$$

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

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

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

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

This concludes the proof. $\qquad\square$

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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,

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

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):

$$\sum_{j \in [p+1]} \sum_{i \in I \setminus \widetilde{I}} \left(\widetilde{\boldsymbol{v}}^\top \frac{\partial \boldsymbol{v}}{\partial z_{ij}}\right)^2 = \frac{1}{\|\boldsymbol{h}\|^2} \sum_{j \in [p+1]} \sum_{i \in I} (\widetilde{\boldsymbol{v}}^\top \boldsymbol{P}^\perp \boldsymbol{B} \boldsymbol{e}_j (\boldsymbol{L}\boldsymbol{r})_i - \widetilde{\boldsymbol{v}}^\top \boldsymbol{P}^\perp \boldsymbol{B} \boldsymbol{Z}^\top \boldsymbol{L} \boldsymbol{e}_i h_j)^2$$

$$\lesssim \|\boldsymbol{B}\|_{\mathrm{op}}^2 \|\boldsymbol{L}\boldsymbol{Z}\|_{\mathrm{op}}^2 + \|\boldsymbol{L}\boldsymbol{Z}\boldsymbol{B}^\top\|_{\mathrm{op}}^2 \lesssim (k\mu)^{-2} \|\boldsymbol{L}\boldsymbol{Z}\|_{\mathrm{op}}^2.$$

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

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

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

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

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

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

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

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

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

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

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

which implies

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

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

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

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

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

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

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 $m \neq \ell$, we decompose

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

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

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

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

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

Thus, Markov's inequality leads to

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

This concludes the proof. $\qquad\square$

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

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

thanks to $\tau = \min(1, \mu)$, where $\boldsymbol{G} \in \mathbb{R}^{n \times p}$ has i.i.d. Gaussian entries. Since $\|\boldsymbol{LG}\|_{\mathrm{op}} \overset{\mathrm{d}}{=} \|\boldsymbol{G}'\|_{\mathrm{op}}$ 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

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

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. $\square$

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

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

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

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

Then, multiple applications of the triangle inequality lead to

$$\mathbb{P}(|\widehat{\mathsf{df}}_m - \widetilde{\mathsf{df}}_M| > n\epsilon) \leq \mathbb{P}\left(\left|\frac{\widehat{\mathsf{df}}_m}{n} - d_n\right| > \frac{\epsilon}{2}\right) + \mathbb{P}\left(\left|\frac{M^{-1}\sum_m\widehat{\mathsf{df}}_m}{n} - d_n\right| > \frac{\epsilon}{2}\right)$$

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

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 $\widehat{\mathsf{df}}_m = k - \mathrm{tr}[\boldsymbol{L}_m\boldsymbol{V}_m]$, our goal is to bound

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

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

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

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

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

By the above displays, we obtain

$$\mathbb{E}\left[\left|\frac{\mathrm{tr}[\boldsymbol{L}_m\boldsymbol{V}_m]^2}{n^2} - \mathbb{E}[W_m]\right|\right] \leq \mathbb{E}\left[\left|\frac{\mathrm{tr}[\boldsymbol{L}_m\boldsymbol{V}_m]^2}{n^2} - W_m\right|\right] + \mathbb{E}\left[\left|W_m - \mathbb{E}[W_m]\right|\right]$$

$$\leq \mathbb{E}\left[\left|\frac{\mathrm{tr}[\boldsymbol{L}_m\boldsymbol{V}_m]^2}{n^2} - W_m\right|\right] + \sqrt{\mathrm{Var}[W_m]}$$

$$\lesssim \frac{\gamma^{7/2}}{\sqrt{n}\tau^2 c^2} + \sqrt{\frac{\gamma^3}{n\tau^2}} \lesssim \frac{\gamma^{7/2}}{\sqrt{n}\tau^2 c^2}.$$

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

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

Notice that under Ω,

$$\left|\frac{\mathrm{tr}[\boldsymbol{L}_m\boldsymbol{V}_m]^2}{n^2} - \mathbb{E}[W_m]\right| = \left|\frac{\mathrm{tr}[\boldsymbol{L}_m\boldsymbol{V}_m]}{n} - \sqrt{\mathbb{E}[W_m]}\right|\cdot\left|\frac{\mathrm{tr}[\boldsymbol{L}_m\boldsymbol{V}_m]}{n} + \sqrt{\mathbb{E}[W_m]}\right|$$

$$= \left|\frac{\widehat{\mathsf{df}}_m}{n} - d_n\right|\cdot\left|\frac{\mathrm{tr}[\boldsymbol{L}_m\boldsymbol{V}_m]}{n} + \sqrt{\mathbb{E}[W_m]}\right| \geq \left|\frac{\widehat{\mathsf{df}}_m}{n} - d_n\right|C\gamma^{-1}c^2\tau.$$

Combining the above displays together, we obtain

$$\mathbb{P}\left(\left|\frac{\widehat{\mathsf{df}}_m}{n} - d_n\right| > \epsilon\right) \leq \mathbb{P}(\Omega^c) + \mathbb{P}\left(\left|\frac{\mathrm{tr}[\boldsymbol{L}_m\boldsymbol{V}_m]^2}{n^2} - \mathbb{E}[W_m]\right| > \epsilon C\gamma^{-1}c^2\tau\right)$$

$$\leq e^{-nc/2} + \frac{\gamma}{\epsilon C c^2\tau}\mathbb{E}\left[\left|\frac{\mathrm{tr}[\boldsymbol{L}_m\boldsymbol{V}_m]^2}{n^2} - \mathbb{E}[W_m]\right|\right]$$

$$\lesssim e^{-nc/2} + \frac{\gamma}{\epsilon c^2\tau}\cdot\frac{\gamma^{7/2}}{\sqrt{n}\tau^2 c^2} = e^{-nc/2} + \frac{\gamma^{9/2}}{\epsilon\sqrt{n}\tau^3 c^4}.$$

This finishes the proof. □

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

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

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

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

$$\sum_{ij}(\mathsf{Rem}^2_{ij})^2 = \frac{1}{\|\boldsymbol{h}\|^2}\sum_{ij}(\boldsymbol{v}^\top\boldsymbol{Z}^\top\boldsymbol{L}\boldsymbol{Z}\boldsymbol{P}^\perp(\boldsymbol{B}\boldsymbol{e}_j(\boldsymbol{L}\boldsymbol{r})_i - \boldsymbol{B}\boldsymbol{Z}^\top\boldsymbol{L}\boldsymbol{e}_i h_j))^2$$

$$\lesssim \frac{1}{\|\boldsymbol{h}\|^2}(\|\boldsymbol{B}^\top\boldsymbol{P}^\perp\boldsymbol{Z}^\top\boldsymbol{L}\boldsymbol{Z}\boldsymbol{v}\|^2\|\boldsymbol{L}\boldsymbol{r}\|^2 + \|\boldsymbol{L}\boldsymbol{Z}\boldsymbol{B}^\top\boldsymbol{P}^\perp\boldsymbol{Z}^\top\boldsymbol{L}\boldsymbol{Z}\boldsymbol{v}\|^2\|\boldsymbol{h}\|^2)$$

$$\lesssim \|\boldsymbol{B}\|^2_{\mathrm{op}}\|\boldsymbol{L}\boldsymbol{Z}\|^6_{\mathrm{op}} \lesssim (k\mu)^{-2}\|\boldsymbol{L}\boldsymbol{Z}\|^6_{\mathrm{op}}.$$

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 $\gamma = \max(1, p/n)$. $\qquad\square$

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 $\mathbb{P}(\|\boldsymbol{G}\|_{\mathrm{op}} > \sqrt{k} + \sqrt{q} + t) \leq \Phi(-t) \leq e^{-t^2/2}$ holds for all $t \in \mathbb{R}$, where $\Phi(\cdot)$ is the CDF of the standard normal. Therefore, we have $\mathbb{E}[\|\boldsymbol{G}\|^m_{\mathrm{op}}] \leq C(m)(\sqrt{k} + \sqrt{q})^r$ for all $m \geq 1$, 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. $\qquad\square$

Lemma 21 (Simple random sampling properties; see, e.g., page 13 of Chaudhuri (2014))**.** Fix an array $(x_i)_{i=1}^M$ of length $M \geq 1$ and let μ_M be the mean $M^{-1}\sum_{i\in[M]}x_i$ and σ^2_M 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 $m \leq M$. Then, the mean and variance of the sample mean $\widehat{\mu}_J = \sum_{i\in J}x_i$ are given by

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

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

Proof. A random variable X is said to follow a hypergeometric distribution, denoted as $X \sim$ 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)):

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

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

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
>
> $$\check{\boldsymbol{b}}_m = \underset{\boldsymbol{b} \in \mathbb{R}^p}{\operatorname{argmin}} \frac{1}{2|I_m|}\|\boldsymbol{L}_{I_m}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b})\|_2^2 + g_m(\boldsymbol{b}) + \mu_n\|\boldsymbol{\Sigma}^{1/2}(\boldsymbol{b} - \boldsymbol{\beta}_0)\|^2$$
>
> 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
>
> $$\frac{\sigma^2 + \|\boldsymbol{\Sigma}^{1/2}(\check{\boldsymbol{b}}_m - \boldsymbol{\beta}_0)\|^2}{\sigma^2 + \|\boldsymbol{\Sigma}^{1/2}(\widehat{\boldsymbol{\beta}}_m - \widehat{\boldsymbol{\beta}}_0)\|^2} \xrightarrow{\text{P}} 1, \quad \frac{1 - \check{\mathsf{df}}_m/k}{1 - \widehat{\mathsf{df}}_m/k} \xrightarrow{\text{P}} 1, \quad \frac{\|\boldsymbol{y} - \boldsymbol{X}\check{\boldsymbol{b}}_m\|}{\|\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}_m\|} \xrightarrow{\text{P}} 1, \quad \frac{\|\boldsymbol{L}_{I_m}(\boldsymbol{y} - \boldsymbol{X}\check{\boldsymbol{b}}_m)\|}{\|\boldsymbol{L}_{I_m}(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}_m)\|} \xrightarrow{\text{P}} 1$$
>
> as well as
>
> $$\frac{\|\boldsymbol{\Sigma}^{1/2}(\widehat{\boldsymbol{\beta}}_m - \check{\boldsymbol{b}}_m)\|}{\min\{\|\boldsymbol{\Sigma}^{1/2}(\widehat{\boldsymbol{\beta}}_m - \boldsymbol{\beta}_0)\|, \|\boldsymbol{\Sigma}^{1/2}(\check{\boldsymbol{b}}_m - \boldsymbol{\beta}_0)\|\}} \xrightarrow{\text{P}} 0.$$
>
> 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
>
> $$\frac{\widetilde{R}_M^{\text{cgcv,ovlp}}}{R_M} \xrightarrow{\text{P}} 1, \qquad \frac{\widetilde{R}_M^{\text{cgcv,full}}}{R_M} \xrightarrow{\text{P}} 1,$$
>
> 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

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

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

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

Since $p/k \leq \gamma < 1$, 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,

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

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

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

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 $\frac{1-\widehat{\mathsf{df}}_m/k}{1-\widecheck{\mathsf{df}}_m/k}$ 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

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

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

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

By (41), we further have, for $\check{\boldsymbol{b}}_1, \ldots \check{\boldsymbol{b}}_M$ 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 $(\widehat{\boldsymbol{\beta}}_m, \widehat{\mathsf{df}}_m)$. $\qquad\square$

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:

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

where $f_{\mathrm{LI}}(\boldsymbol{x}) = \boldsymbol{\beta}_0^\top \boldsymbol{x}$ is the oracle linear predictor, and $f_{\mathrm{NL}}(\boldsymbol{x}) = y - f_{\mathrm{LI}}(\boldsymbol{x})$ is the nonlinear component that is not explained by $f_{\mathrm{LI}}(\boldsymbol{x})$. The best linear projection has the useful property that $f_{\mathrm{LI}}(\boldsymbol{x})$ is (linearly) uncorrelated with $f_{\mathrm{NL}}(\boldsymbol{x})$, 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:

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

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

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

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

while its two estimators are defined as

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

where

$$N_{m,\ell}^{\mathrm{ovlp}} = |I_m \cap I_\ell|^{-1}(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}_m)^\top \boldsymbol{L}_{m\cap\ell}(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}_\ell$$

$$D_{m,\ell}^{\mathrm{ovlp}} = 1 - \frac{\mathrm{tr}(\boldsymbol{S}_m)}{|I_m|} - \frac{\mathrm{tr}(\boldsymbol{S}_\ell)}{|I_\ell|} + \frac{\mathrm{tr}(\boldsymbol{S}_m)\,\mathrm{tr}(\boldsymbol{S}_\ell)}{|I_m||I_\ell|}$$

$$N_{m,\ell}^{\mathrm{full}} = n^{-1}(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}_m)^\top(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}_\ell)$$

$$D_{m,\ell}^{\mathrm{full}} = 1 - \frac{\mathrm{tr}(\boldsymbol{S}_m)}{n} - \frac{\mathrm{tr}(\boldsymbol{S}_\ell)}{n} + \frac{\mathrm{tr}(\boldsymbol{S}_m)\,\mathrm{tr}(\boldsymbol{S}_\ell)|I_m \cap I_\ell|}{n|I_m||I_\ell|},$$

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 $\lim_{p\to\infty} |\mathrm{tr}[\boldsymbol{C}_p(\boldsymbol{A}_p - \boldsymbol{B}_p)]| = 0$ 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 $|\widehat{R}_{m,\ell,\lambda}^{\mathrm{ovlp}} - R_{m,\ell,\lambda}| \xrightarrow{\text{a.s.}} 0$ and $|\widehat{R}_{m,\ell,\lambda}^{\mathrm{full}} - R_{m,\ell,\lambda}| \xrightarrow{\text{a.s.}} 0$. Consequently, it holds that $|\widetilde{R}_{M,\lambda}^{\mathrm{ovlp}} - R_{M,\lambda}| \xrightarrow{\text{a.s.}} 0$ and $|\widetilde{R}_{M,\lambda}^{\mathrm{full}} - R_{M,\lambda}| \xrightarrow{\text{a.s.}} 0$.

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

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

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

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 $\{R_{m,\ell} - \mathscr{R}_{m,\ell}\}_{p\in\mathbb{N}}$, $\{\widehat{R}_{m,\ell}^{\text{ovlp}} - \mathscr{R}_{m,\ell}\}_{p\in\mathbb{N}}$, and $\{\widehat{R}_{m,\ell}^{\text{full}} - \mathscr{R}_{m,\ell}\}_{p\in\mathbb{N}}$ are uniformly equicontinuous on $\lambda \in \Lambda = [0,\infty]$ when $\psi \neq 1$. This implies that the sequences of functions $\{R_{m,\ell} - \widehat{R}_{m,\ell}^{\text{ovlp}}\}_{p\in\mathbb{N}}$ and $\{R_{m,\ell} - \widehat{R}_{m,\ell}^{\text{full}}\}_{p\in\mathbb{N}}$ 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

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

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

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

$$\widehat{\beta}_m - \beta_0 = M_m \frac{X^\top L_m}{k}(X\beta_0 + f_{\text{NL}}) - \beta_0$$

$$= -\lambda M_m \beta_0 + M_m \frac{X^\top L_m}{k} f_{\text{NL}}.$$

Then it follows that

$$\|f_{\text{NL}}\|_{L_2}^2 + (\widehat{\beta}_m - \beta_0)^\top \Sigma(\widehat{\beta}_\ell - \beta_0)$$
$$= \lambda^2 \beta_0^\top M_m \Sigma M_\ell \beta_0$$
$$\quad + (\|f_{\text{NL}}\|_{L_2}^2 + f_{\text{NL}}^\top \frac{L_m X}{k} M_m \Sigma M_\ell \frac{X^\top L_\ell}{k} f_{\text{NL}})$$
$$\quad + (\lambda \beta_0^\top M_\ell \Sigma M_\ell \frac{X^\top L_m}{k} f_{\text{NL}} + \lambda \beta_0^\top M_\ell \Sigma M_m \frac{X^\top L_m}{k} f_{\text{NL}})$$
$$= T_B + T_V + T_C.$$

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 $f_0 = L_{m\cap\ell} f_{\text{NL}}$:

$$T_V \simeq \|f_{\text{NL}}\|_{L_2}^2 + \frac{1}{k^2} f_0^\top X M_m \Sigma M_\ell X^\top f_0 \simeq \|f_{\text{NL}}\|_{L_2}^2(1 + \widetilde{v}_p(-\lambda; \varphi_{m\ell}, \psi)),$$

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. □

> **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

$$L_{m\cap\ell}(y - X\widehat{\beta}_m) = L_{m\cap\ell}\left(I_p - XM_m\frac{X^\top L_m}{k}\right)y$$

$$= \lambda L_{m\cap\ell}XM_m\beta_0 + L_{m\cap\ell}\left(I_p - XM_m\frac{X^\top L_m}{k}\right)f_{\text{NL}}.$$

Then it follows that

$$\frac{1}{|I_m \cap I_\ell|}(y - X\widehat{\beta}_m)^\top L_{m\cap\ell}(y - X\widehat{\beta}_\ell)$$
$$= \lambda^2 \beta_0^\top M_m \widehat{\Sigma}_{m\cap\ell} M_\ell \beta_0$$
$$+ \frac{1}{|I_m \cap I_\ell|}(L_m f_{\text{NL}})^\top \left(I_p - \frac{XM_mX^\top}{k}\right) L_{m\cap\ell}\left(I_p - \frac{XM_\ell X^\top}{k}\right)(L_\ell f_{\text{NL}})$$
$$+ \frac{\lambda}{|I_m \cap I_\ell|}\left(\beta_0^\top M_\ell X^\top L_{m\cap\ell}\left(I_p - XM_m\frac{X^\top L_m}{k}\right)f_{\text{NL}} + \beta_0^\top M_\ell X^\top L_{m\cap\ell}\left(I_p - XM_\ell\frac{X^\top L_\ell}{k}\right)f_{\text{NL}}\right)$$
$$= T_B + T_V + T_C.$$

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 $f_0 = L_{m\cap\ell}f_{\text{NL}}$:

$$T_V \simeq \frac{1}{|I_m \cap I_\ell|}(L_{m\cap\ell}f_{\text{NL}})^\top\left(I_p - \frac{XM_mX^\top}{k}\right)L_{m\cap\ell}\left(I_p - \frac{XM_\ell X^\top}{k}\right)(L_{m\cap\ell}f_{\text{NL}})$$

$$= \frac{1}{|I_m \cap I_\ell|}\|f_0\|_2^2 - \frac{1}{k}\sum_{j\in\{m,\ell\}}f_0^\top\frac{XM_jX^\top}{|I_m \cap I_\ell|}f_0 + \frac{|I_m \cap I_\ell|}{k^2}f_0^\top\frac{XM_m\widehat{\Sigma}_{m\cap\ell}M_\ell X^\top}{|I_m \cap I_\ell|}f_0$$

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

where the last equivalence is from Lemma 31.

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

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 $I_m, I_\ell \overset{\text{SRS}}{\sim} \mathcal{I}_k$, it holds that

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

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

Proof. Since $I_m, I_\ell \overset{\text{SRS}}{\sim} \mathcal{I}_k$, we have that $|I_m| = |I_\ell| = k$. From Du et al. (2023, Lemma C.1), we have that
$$\frac{\text{tr}(\boldsymbol{S}_j)}{|I_j|} = \frac{1}{k^2}\text{tr}(\boldsymbol{X}\boldsymbol{M}_j\boldsymbol{X}^\top\boldsymbol{L}_j) = \frac{1}{k}\text{tr}(\boldsymbol{M}_j\boldsymbol{\Sigma}_j) \simeq 1 - \lambda v_p(-\lambda; \psi).$$

By continuous mapping theorem,

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

which completes the proof. $\square$

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

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

where

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

$\varphi_{m\ell} = \psi \mathbb{1}_{\{m=\ell\}} + \phi \mathbb{1}_{\{m \neq \ell\}}$, 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

$$\frac{1}{n}(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}_m)^\top(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}_\ell)$$

$$= \frac{1}{n}\lambda^2 \boldsymbol{\beta}_0^\top \boldsymbol{M}_m \widehat{\boldsymbol{\Sigma}} \boldsymbol{M}_\ell \boldsymbol{\beta}_0$$

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

$$+ \frac{\lambda}{n}\left(\boldsymbol{\beta}_0^\top \boldsymbol{M}_\ell\boldsymbol{X}^\top\left(\boldsymbol{I}_p - \boldsymbol{X}\boldsymbol{M}_m\frac{\boldsymbol{X}^\top\boldsymbol{L}_m}{k}\right)\boldsymbol{f}_{\text{NL}} + \boldsymbol{\beta}_0^\top \boldsymbol{M}_\ell\boldsymbol{X}^\top\left(\boldsymbol{I}_p - \boldsymbol{X}\boldsymbol{M}_\ell\frac{\boldsymbol{X}^\top\boldsymbol{L}_\ell}{k}\right)\boldsymbol{f}_{\text{NL}}\right)$$

$$= T_B + T_V + T_C.$$

66

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 $\boldsymbol{f}_1 = \boldsymbol{L}_{m\cup\ell}\boldsymbol{f}_{\mathrm{NL}}$ and $\boldsymbol{f}_0 = \boldsymbol{L}_{m\cap\ell}\boldsymbol{f}_{\mathrm{NL}}$, respectively:

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

From Lemma 31, it follows that

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

Combining the above results completes the proof. $\qquad\square$

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 $I_m, I_\ell \overset{\mathrm{SRS}}{\sim} \mathcal{I}_k$, it holds that

$$D_{m,\ell}^{\mathrm{full}} \simeq \mathscr{D}_{m,\ell}^{\mathrm{full}} := \begin{cases} \dfrac{\psi - \phi}{\psi} + \dfrac{\phi}{\psi}\lambda^2 v_p(-\lambda; \psi)^2, & m = \ell, \\[3mm] \left(\dfrac{\psi - \phi}{\psi} + \dfrac{\phi}{\psi}\lambda v_p(-\lambda; \psi)\right)^2, & m \neq \ell, \end{cases}$$

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

Proof. From the proof of Lemma 26, we have $\mathrm{tr}(\boldsymbol{S}_j)/|I_j| \xrightarrow{\text{a.s.}} 1 - \lambda v_p(-\lambda; \psi)$, 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

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

67

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

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

$$= \left(\frac{\psi - \phi}{\psi} + \frac{\phi}{\psi}\lambda v_p(-\lambda;\psi) \right)^2,$$

which finishes the proof. $\square$

> **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 $\mathscr{R}_{m,\ell}, \mathscr{N}^{\text{ovlp}}_{m,\ell}, \mathscr{D}^{\text{ovlp}}_{m,\ell}, \mathscr{N}^{\text{full}}_{m,\ell}, \mathscr{D}^{\text{full}}_{m,\ell}$ defined in Lemmas 24 and 26–28, the sequences of functions $\{R_{m,\ell} - \mathscr{R}_{m,\ell}\}_{p \in \mathbb{N}}$, $\{N^{\text{ovlp}}_{m,\ell}/D^{\text{ovlp}}_{m,\ell} - \mathscr{N}^{\text{ovlp}}_{m,\ell}/\mathscr{D}^{\text{ovlp}}_{m,\ell}\}_{p \in \mathbb{N}}$, and $\{N^{\text{full}}_{m,\ell}/D^{\text{full}}_{m,\ell} - \mathscr{N}^{\text{full}}_{m,\ell}/\mathscr{D}^{\text{full}}_{m,\ell}\}_{p \in \mathbb{N}}$ 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

$$\frac{1}{|I_m \cap I_\ell|}(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}_m)^\top \boldsymbol{L}_{m \cap \ell}(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}_\ell)$$

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

From law of large numbers, $\|\boldsymbol{L}_{m \cap \ell}\boldsymbol{f}_{\text{NL}}\|_2^2/|I_m \cap I_\ell| \xrightarrow{\text{a.s.}} \|f_{\text{NL}}\|_{L_2}^2$ 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), $\boldsymbol{f}_{\text{NL}}^\top \boldsymbol{L}_{m \cap \ell}\boldsymbol{X}\boldsymbol{\beta}_0/|I_m \cap I_\ell| \xrightarrow{\text{a.s.}} 0$.

For the other term, note that,

$$\|\widehat{\boldsymbol{\beta}}_m\|_2 \leq \|(\boldsymbol{X}^\top \boldsymbol{L}_m \boldsymbol{X}/k + \lambda \boldsymbol{I}_p)^{-1}(\boldsymbol{X}^\top \boldsymbol{L}_m \boldsymbol{y}/k)\|_2$$

$$\leq \|(\boldsymbol{X}^\top \boldsymbol{L}_m \boldsymbol{X}/k + \lambda \boldsymbol{I}_p)^{-1}\boldsymbol{X}^\top \boldsymbol{L}_m/\sqrt{k}\| \cdot \|\boldsymbol{L}_m \boldsymbol{y}/\sqrt{k}\|_2$$

$$\leq C\sqrt{\mathbb{E}[y_1^2]} \cdot \|(\boldsymbol{X}^\top \boldsymbol{L}_m \boldsymbol{X}/k + \lambda \boldsymbol{I}_p)^{-1}\boldsymbol{X}^\top \boldsymbol{L}_m/\sqrt{k}\|_{\text{op}},$$

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 $\|\widehat{\boldsymbol{\Sigma}}\|_{\text{op}}$ 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 $\operatorname{tr}(\boldsymbol{S}_j)/k \xrightarrow{\text{a.s.}} 0$. 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

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

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 $\|\boldsymbol{S}_j\|_{\mathrm{op}} = \|\boldsymbol{L}_j\boldsymbol{X}(\boldsymbol{X}^{\top}\boldsymbol{L}_j\boldsymbol{X}/k + \lambda\boldsymbol{I}_p)^{+}\boldsymbol{X}^{\top}\boldsymbol{L}_j/|I_j|\|_{\mathrm{op}} \leq 1$ 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 $\|\partial\boldsymbol{S}_j/\partial\lambda\|_{\mathrm{op}}$ 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 $Q_{p,0} := \lim_{\lambda\to 0^+} Q_{p,\lambda} = (1 - \psi)^2$, 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,

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

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

For the numerator, similarly, we define

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

and $Q'_{p,\lambda} = \mathcal{N}^{\text{ovlp}}_{m,\ell}$. By the similar argument above, the sequence of function $P'_{p,\lambda} - Q'_{p,\lambda}$ is equicontinuous on Λ and thus $Q'_{p,0} = \lim_{\lambda \to 0^+} Q'_{p,\lambda}$ 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 $P''_{p,\lambda} := N^{\text{ovlp}}_{m,\ell}/D^{\text{ovlp}}_{m,\ell}$ and $Q''_{p,\lambda} := \mathcal{N}^{\text{ovlp}}_{m,\ell}/\mathcal{D}^{\text{ovlp}}_{m,\ell}$. From the proof of Part (1), we know that $D^{\text{ovlp}}_{m,\ell}$ is positive and upper bounded almost surely, and $N^{\text{ovlp}}_{m,\ell}$ 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 $\psi \neq 1$, $\mathcal{D}^{\text{ovlp}}_{m,\ell} = (1 - \psi)^2$; when $\lambda > 0$, $\mathcal{D}^{\text{ovlp}}_{m,\ell}$ 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 $Q''_{p,\lambda}$:

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

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. $\qquad\square$

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 $|\widehat{R}^{\text{cgcv,ovlp}}_{M,\lambda} - R_{M,\lambda}| \xrightarrow{\text{a.s.}} 0$ and $|\widehat{R}^{\text{cgcv,full}}_{M,\lambda} - R_{M,\lambda}| \xrightarrow{\text{a.s.}} 0$. Moreover, when $\psi \neq 1$, it holds that $\sup_{\lambda \in \Lambda} |\widehat{R}^{\text{cgcv,ovlp}}_{M,\lambda} - R_{M,\lambda}| \xrightarrow{\text{a.s.}} 0$ and $\sup_{\lambda \in \Lambda} |\widehat{R}^{\text{cgcv,full}}_{M,\lambda} - R_{M,\lambda}| \xrightarrow{\text{a.s.}} 0$.

Pointwise consistency

From Lemma 24 and Lemma 27, we have that

$$R_M = \|f_{\text{NL}}\|^2_{L_2} + \frac{1}{M^2} \sum_{m=1}^{M} \|\widehat{\boldsymbol{\beta}}(\mathcal{D}_{I_m}) - \boldsymbol{\beta}_0\|^2_{\boldsymbol{\Sigma}}$$

$$+ \frac{1}{M^2} \sum_{m,\ell=1}^{M} (\widehat{\boldsymbol{\beta}}(\mathcal{D}_{I_m}) - \boldsymbol{\beta}_0)^\top \boldsymbol{\Sigma}(\widehat{\boldsymbol{\beta}}_\ell - \boldsymbol{\beta}_0))$$

$$\simeq \left(\frac{1}{M}(1 + \widetilde{v}_p(-\lambda; \psi, \psi)) \right.$$

$$\left. + \frac{M-1}{M}(1 + \widetilde{v}_p(-\lambda; \phi, \psi)) \right) (\|f_{\text{NL}}\|^2_{L_2} + \widetilde{c}_p(-\lambda; \psi)), \qquad (63)$$

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

and

$$\frac{1}{n}\|\boldsymbol{y} - \boldsymbol{X}\widetilde{\boldsymbol{\beta}}_M\|^2_2$$

$$= \frac{1}{M^2} \sum_{m=1}^{M} \frac{1}{n} \|\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}_m)\|_2^2 + \frac{1}{M^2} \sum_{m,\ell=1}^{M} \frac{1}{n} (\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}_m)^\top (\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}_\ell)$$

$$\simeq \left(\frac{1}{M} d_p(-\lambda; \psi, \psi)(1 + \widetilde{v}_p(-\lambda; \psi, \psi)) \right.$$

$$\left. + \frac{M-1}{M} d_p(-\lambda; \phi, \psi)(1 + \widetilde{v}_p(-\lambda; \phi, \psi)) \right) (\|f_{\mathrm{NL}}\|_{L_2}^2 + \widetilde{c}_p(-\lambda; \psi)). \tag{65}$$

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:

$$R_M - \frac{\|\boldsymbol{y} - \boldsymbol{X}\widetilde{\boldsymbol{\beta}}_M\|_2^2/n}{(1 - \widetilde{\mathsf{df}}/n)^2}$$

$$\simeq \frac{1}{M} \left(1 - \frac{d_p(-\lambda; \psi, \psi)}{d_p(-\lambda; \phi, \psi)} \right) (1 + \widetilde{v}_p(-\lambda; \psi, \psi))(\|f_{\mathrm{NL}}\|_{L_2}^2 + \widetilde{c}_p(-\lambda; \psi)). \tag{68}$$

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

$$\frac{1}{(1 - \widetilde{\mathsf{df}}/n)^2} \cdot \frac{1}{M} \cdot \frac{(\psi - \phi)(\widetilde{\mathsf{df}}/n)^2}{\phi(1 - \widetilde{\mathsf{df}}/n)^2 + (\psi - \phi)(\widetilde{\mathsf{df}}/n)^2} \cdot \frac{\|\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}_m\|_2^2}{n}$$

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

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

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

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

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

71

and

$$\widetilde{R}_M^{\mathrm{gcv}} \simeq \frac{1}{M^2} \sum_{m,\ell \in [M]} \frac{\mathscr{N}_{m,\ell}^{\mathrm{full}}}{\mathscr{D}_{1,2}^{\mathrm{full}}} \simeq \frac{1}{M^2} \sum_{m,\ell \in [M]} R_{m,\ell} - \frac{1}{M^2} \sum_{m \in [M]} \frac{\mathscr{D}_{m,m}^{\mathrm{full}}}{\mathscr{D}_{1,2}^{\mathrm{full}}} R_m$$

$$\widehat{R}_{m,m}^{\#} \simeq R_{m,m}.$$

Then we have

$$\widehat{R}_M^{\mathrm{cgcv},\#} = \widetilde{R}_M^{\mathrm{gcv}} - \frac{1}{M} \frac{(\widetilde{\mathsf{df}}/n)^2}{(1-\widetilde{\mathsf{df}}/n)^2} \frac{\psi - \phi}{\phi} \frac{1}{M} \sum_{m=1}^M \widehat{R}_{m,m}^{\#}$$

$$\simeq \widetilde{R}_M^{\mathrm{gcv}} - \frac{1}{M^2} \sum_{m=1}^M \frac{\mathscr{D}_{1,1}^{\mathrm{full}} - \mathscr{D}_{1,2}^{\mathrm{full}}}{\mathscr{D}_{1,2}^{\mathrm{full}}} R_{m,m}$$

$$\simeq \frac{1}{M^2} \sum_{m,\ell \in [M]} R_{m,\ell}$$

$$= R_M,$$

which establishes the point-wise consistency.

Similar to the proof of Theorem 5, the uniform equicontinuity of $\widetilde{R}_M^{\mathrm{gcv}}$ and $R_{m,\ell}$ to their asymptotic limits follows from Lemma 29. And the uniformity for $|\widehat{R}_M^{\mathrm{cgcv},\#} - R_M|$ 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) $\lambda^2 \boldsymbol{\beta}_0^\top \boldsymbol{M}_m \widehat{\boldsymbol{\Sigma}} \boldsymbol{M}_\ell \boldsymbol{\beta}_0 \simeq \begin{cases} \left(\dfrac{\psi - \phi}{\psi} + \dfrac{\phi}{\psi}\lambda^2 v_p(-\lambda; \psi)^2\right) \widetilde{c}_p(-\lambda; \psi)(1 + \widetilde{v}_p(-\lambda; \varphi_{m\ell}, \psi)), & m = \ell, \\[3mm] \left(\dfrac{\psi - \phi}{\psi} + \dfrac{\phi}{\psi}\lambda v_p(-\lambda; \psi)\right)^2 \widetilde{c}_p(-\lambda; \psi)(1 + \widetilde{v}_p(-\lambda; \varphi_{m\ell}, \psi)), & m \neq \ell. \end{cases}$

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

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

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

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

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. $\square$

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

(2) $\frac{1}{|I_m \cap I_\ell|} \boldsymbol{f}_0^\top \left(\boldsymbol{I}_p - \frac{\boldsymbol{X} \boldsymbol{M}_m \boldsymbol{X}^\top}{k} \right) \boldsymbol{L}_{m\cap\ell} \left(\boldsymbol{I}_p - \frac{\boldsymbol{X} \boldsymbol{M}_\ell \boldsymbol{X}^\top}{k} \right) \boldsymbol{f}_0 \simeq \|f_{\mathrm{NL}}\|_{L_2}^2 \lambda^2 v_p(-\lambda; \psi))^2 (1 + \widetilde{v}_p(-\lambda; \varphi_{m\ell}, \psi))$.

(3) $\frac{|I_m \cup I_\ell|}{n} \boldsymbol{f}_1^\top \left(\boldsymbol{I}_p - \frac{\boldsymbol{X} \boldsymbol{M}_m \boldsymbol{X}^\top}{k} \right) \boldsymbol{L}_{m\cup\ell} \left(\boldsymbol{I}_p - \frac{\boldsymbol{X} \boldsymbol{M}_\ell \boldsymbol{X}^\top}{k} \right) \boldsymbol{f}_1$
$\quad + \frac{|(I_m \cup I_\ell)^c|}{n} \|\boldsymbol{f}_0\|_2^2 + \frac{|(I_m \cup I_\ell)^c|^2}{nk^2} \boldsymbol{f}_0^\top \boldsymbol{X} \boldsymbol{M}_m \widehat{\boldsymbol{\Sigma}}_{(m\cup\ell)^c} \boldsymbol{M}_\ell \boldsymbol{X}^\top \boldsymbol{f}_0$

$$\simeq \begin{cases} \|f_{\mathrm{NL}}\|_{L_2}^2 \left(\dfrac{\psi - \phi}{\psi} + \dfrac{\phi}{\psi} \lambda^2 v_p(-\lambda; \psi)^2 \right) (1 + \widetilde{v}_p(-\lambda; \varphi_{m\ell}, \psi)), & m = \ell, \\[2ex] \|f_{\mathrm{NL}}\|_{L_2}^2 \left(\dfrac{\psi - \phi}{\psi} + \dfrac{\phi}{\psi} \lambda v_p(-\lambda; \psi) \right)^2 (1 + \widetilde{v}_p(-\lambda; \varphi_{m\ell}, \psi)), & m \neq \ell. \end{cases}$$

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

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

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

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

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

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

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

$$= \widetilde{v}_p(-\lambda; \varphi_{m\ell}, \psi)).$$

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

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

Part (2) Note that

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

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

74

We next analyze the three terms separately for $m \neq \ell$. By the law of large numbers, the first term converges as $\boldsymbol{f}_0^\top \boldsymbol{f}_0 / |I_m \cap I_\ell| \xrightarrow{\text{a.s.}} \|f_{\mathrm{NL}}\|_{L_2}^2$. 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

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

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

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

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

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

when $m \neq \ell$.

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

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

$$\simeq \|f_{\mathrm{NL}}\|_{L_2}^2 (1 - \frac{2}{k} \operatorname{tr}(\boldsymbol{M}\widehat{\boldsymbol{\Sigma}}_m) + \frac{1}{k} \operatorname{tr}(\boldsymbol{M}_m\widehat{\boldsymbol{\Sigma}}_m\boldsymbol{M}_m\widehat{\boldsymbol{\Sigma}}_m))$$

$$\simeq \|f_{\mathrm{NL}}\|_{L_2}^2 \lambda^2 v_p(-\lambda; \psi)^2 (1 + \widetilde{v}_p(-\lambda; \psi, \psi)).$$

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

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

$$\simeq \lambda^2 v_p(-\lambda; \psi)^2 \left(\frac{2(\psi - \phi)}{2\psi - \phi} \frac{1}{\lambda v_p(-\lambda; \psi)} + \frac{\phi}{2\psi - \phi} \right) (1 + \widetilde{v}_p(-\lambda; \psi, \psi)).$$

From Lemma 32, it then follows that

$$\frac{|I_m \cup I_\ell|}{n} \boldsymbol{f}_1^\top \left(\boldsymbol{I}_p - \frac{\boldsymbol{X}\boldsymbol{M}_m\boldsymbol{X}^\top}{k} \right) \boldsymbol{L}_{m \cup \ell} \left(\boldsymbol{I}_p - \frac{\boldsymbol{X}\boldsymbol{M}_\ell\boldsymbol{X}^\top}{k} \right) \boldsymbol{f}_1$$

$$\simeq \|f_{\mathrm{NL}}\|_{L^2}^2 \frac{\phi(2\psi - \phi)}{\psi^2} \lambda^2 v_p(-\lambda; \psi)^2 \left(\frac{2(\psi - \phi)}{2\psi - \phi} \frac{1}{\lambda v_p(-\lambda; \psi)} + \frac{\phi}{2\psi - \phi} \right) (1 + \widetilde{v}_p(-\lambda; \psi, \psi)).$$

From strong law of large numbers, the second term converges $\frac{|(I_m \cup I_\ell)^c|}{n} \|\boldsymbol{f}_0\|_2^2 \xrightarrow{\text{a.s.}} \|f_{\mathrm{NL}}\|_{L^2}^2$. 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

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

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

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

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

$$\simeq \|f_{\mathrm{NL}}\|_{L_2}^2 \frac{k}{n} \left(1 - \frac{2}{k} \operatorname{tr}(\boldsymbol{M}\widehat{\boldsymbol{\Sigma}}_m) + \frac{1}{k} \operatorname{tr}(\boldsymbol{M}_m\widehat{\boldsymbol{\Sigma}}_m\boldsymbol{M}_m\widehat{\boldsymbol{\Sigma}}_m) \right)$$

$$+ \|f_{\mathrm{NL}}\|_{L_2}^2 \frac{n - k}{n} \left(1 + \frac{1}{n} \operatorname{tr}(\boldsymbol{M}_m\widehat{\boldsymbol{\Sigma}}_m\boldsymbol{M}_m\widehat{\boldsymbol{\Sigma}}_{m^c}) \right)$$

$$\simeq \|f_{\mathrm{NL}}\|_{L_2}^2 \frac{\phi}{\psi} \lambda^2 v_p(-\lambda;\psi)^2(1+\widetilde{v}_p(-\lambda;\varphi_{m\ell},\psi)) + \frac{\psi-\phi}{\psi}\left(1+\frac{1}{k}\operatorname{tr}(\boldsymbol{M}_m\widehat{\boldsymbol{\Sigma}}_m\boldsymbol{M}_m\boldsymbol{\Sigma})\right)$$

$$\simeq \|f_{\mathrm{NL}}\|_{L_2}^2 \frac{\phi}{\psi} \lambda^2 v_p(-\lambda;\psi)^2(1+\widetilde{v}_p(-\lambda;\varphi_{m\ell},\psi)) + \frac{\psi-\phi}{\psi}(1+\widetilde{v}_p(-\lambda;\psi,\psi))$$

$$\simeq \|f_{\mathrm{NL}}\|_{L_2}^2 \left(\frac{\psi-\phi}{\psi} + \frac{\phi}{\psi}\lambda^2 v_p(-\lambda;\psi)^2\right)(1+\widetilde{v}_p(-\lambda;\varphi_{m\ell},\psi)).$$

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

> **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 $\limsup\|\boldsymbol{A}\|_{\mathrm{op}} \le M_0$ 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
>
> $$\frac{1}{n}\boldsymbol{f}^\top \boldsymbol{X}\boldsymbol{A}\boldsymbol{X}^\top \boldsymbol{f} \simeq \frac{1}{n}\|f_{\mathrm{NL}}\|_{L_2}^2 \operatorname{tr}(\boldsymbol{X}\boldsymbol{A}\boldsymbol{X}^\top). \tag{73}$$

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 $U_{ij} = [\boldsymbol{f}]_i[\boldsymbol{f}]_j\mathbb{1}\{i \ne j\}$. We then have

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

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, $\|\boldsymbol{U}\|_{\mathrm{tr}}/n \leq 2\|\boldsymbol{f}\|_2^2/n \xrightarrow{\mathrm{a.s.}} 2\|f_{\mathrm{NL}}\|_{L_2}^2 < \infty$ 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 $n^{-1}|\langle \boldsymbol{B}_1^{-1}\boldsymbol{B}_2\boldsymbol{B}_1^{-1}, \boldsymbol{U}\rangle| \xrightarrow{\mathrm{a.s.}} 0$. Therefore,

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

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

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

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

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

which finishes the proof. $\qquad\square$

C Additional numerical illustrations in Section 4

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

Elastic net:
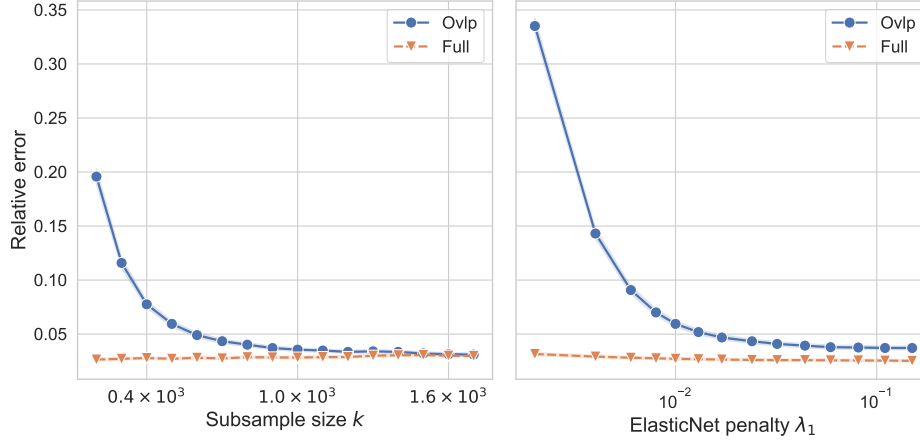
Figure 9: **Full-estimator performs better than ovlp-estimator across different subsample sizes.** The relative errors of the "ovlp" and "full" estimators for ridge ensemble with ensemble size $M = 10$. The left panel shows the results with ridge penalty $\lambda = 1$ and varying subsample size k. The 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.

Lasso:
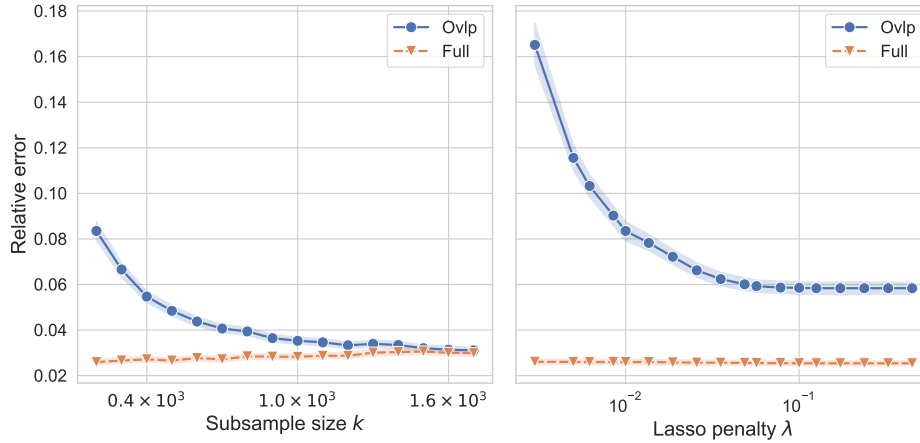
Figure 10: **Full-estimator performs better than ovlp-estimator across different subsample sizes.** The relative errors of the "ovlp" and "full" estimators for ridge ensemble with ensemble size $M = 10$. The left panel shows the results with ridge penalty $\lambda = 1$ and varying subsample size k. The 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
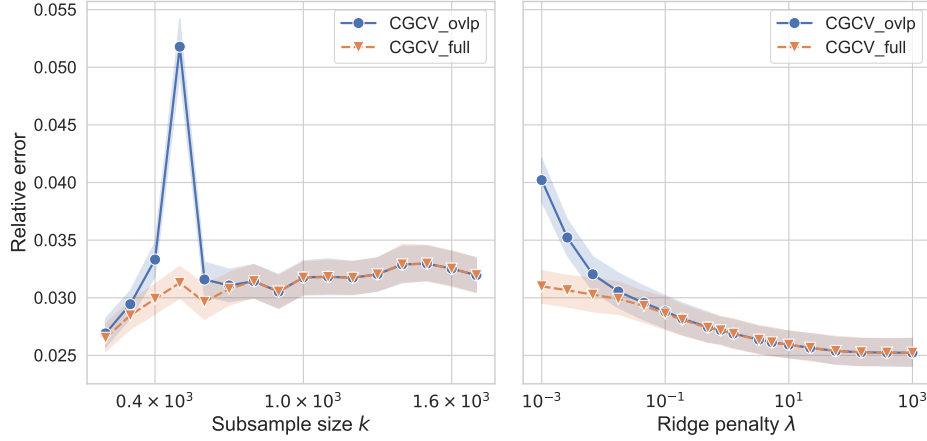
Ridge:

Figure 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.
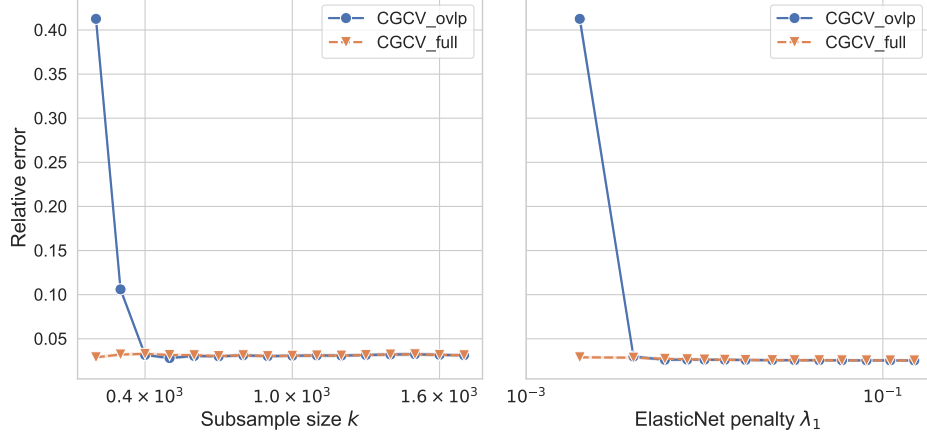
Elastic net:

Figure 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.
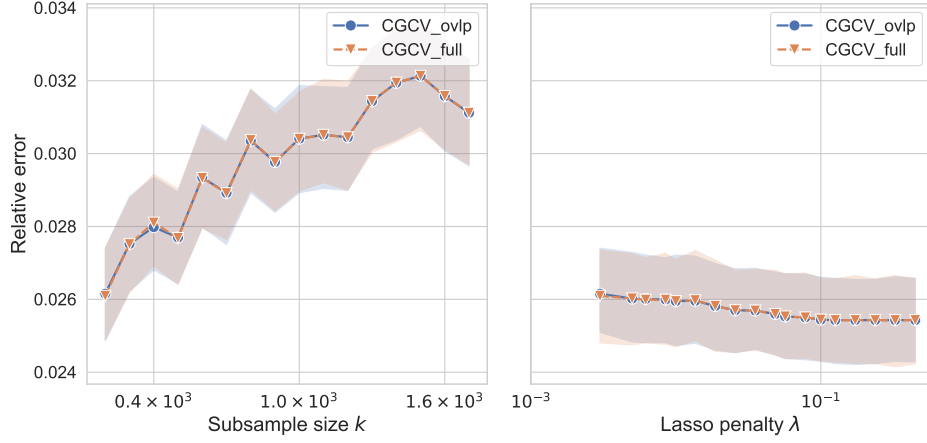
Lasso:

Figure 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
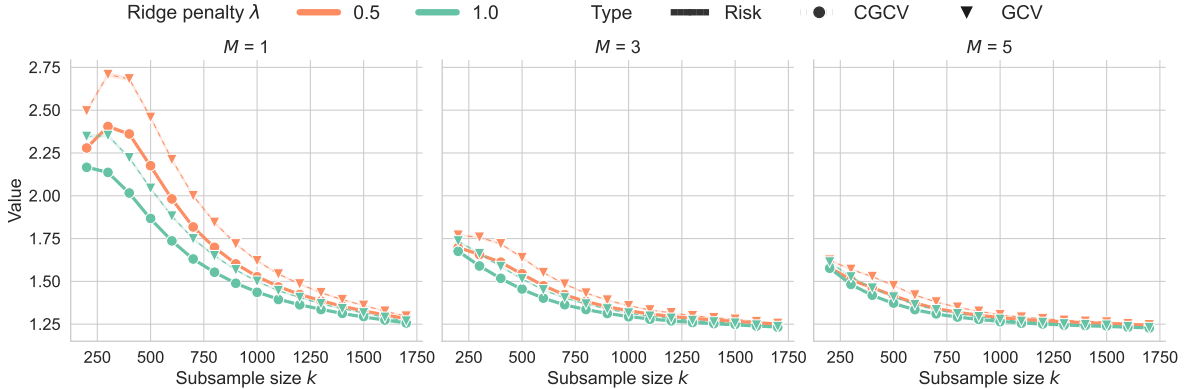
Ridge:

Figure 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.
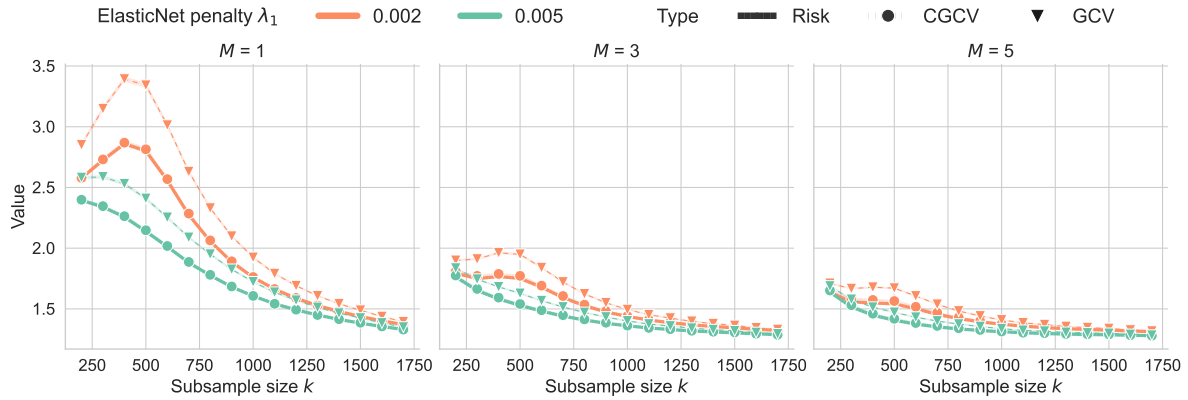
Elastic Net:

Figure 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.
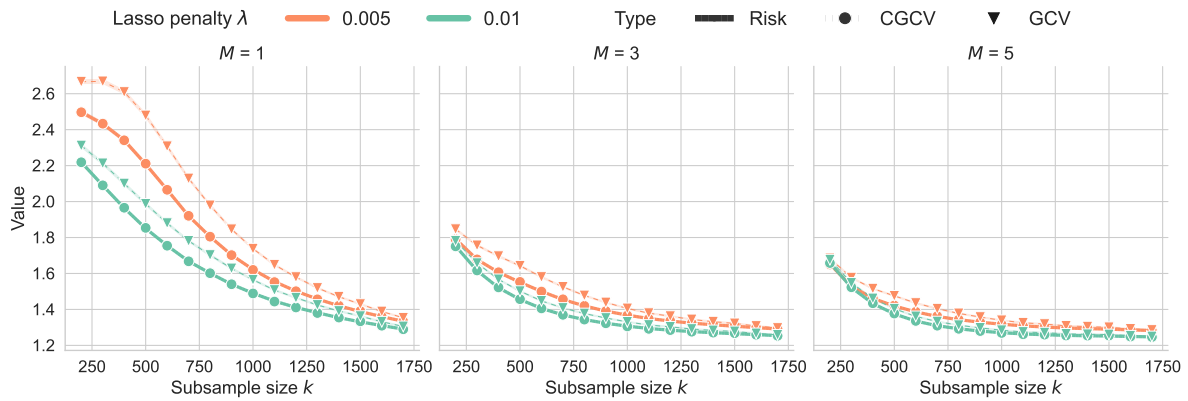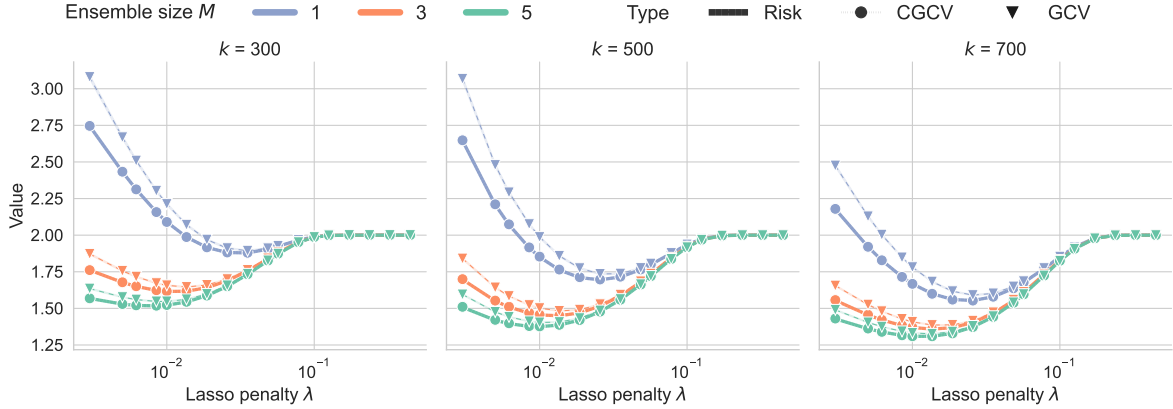
Lasso:

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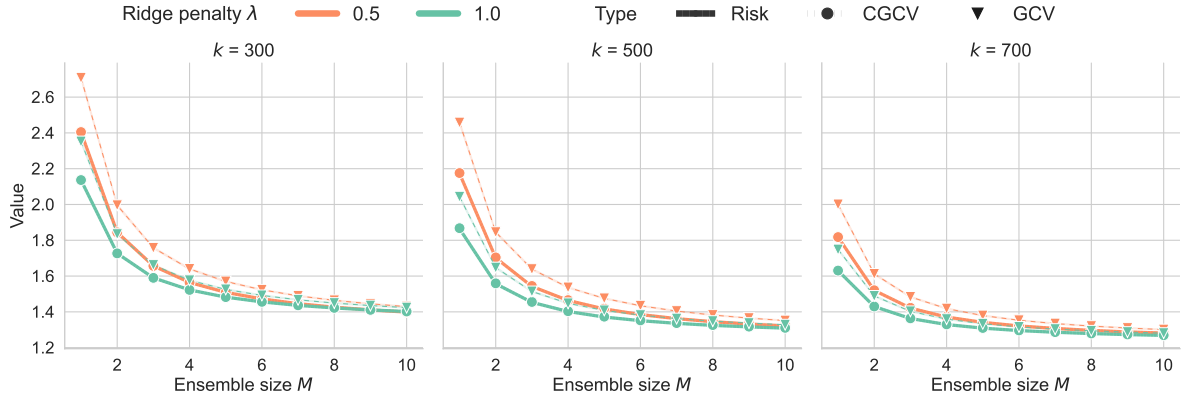
Figure 17: **GCV gets closer to the risk as ensemble size increases across different subsample sizes.** Plot of risks versus the subsample size k for the lasso ensemble with different λ and M. The data generating process is the same as in Figure 1.

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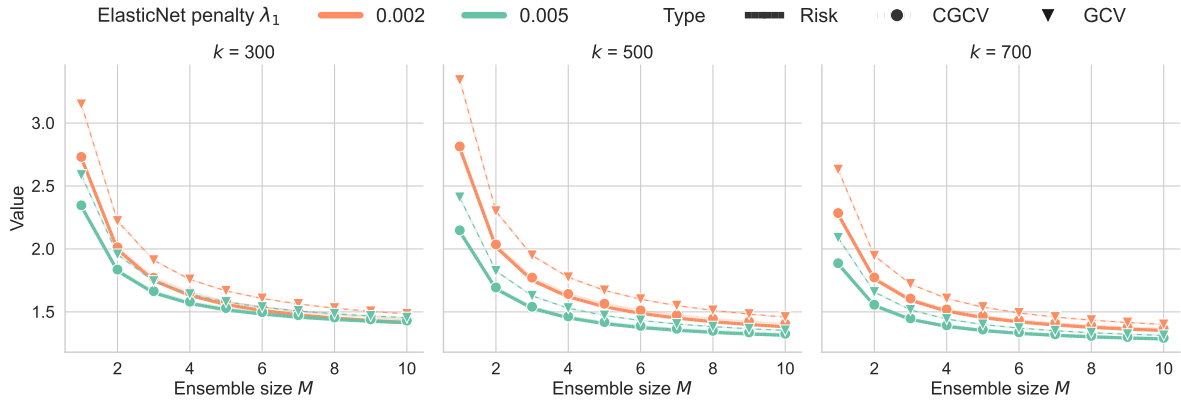
Figure 18: **GCV gets closer to the risk as ensemble size increases across different subsample sizes.** Plot of risks versus the ensemble size M for ridge ensemble with different λ and k.
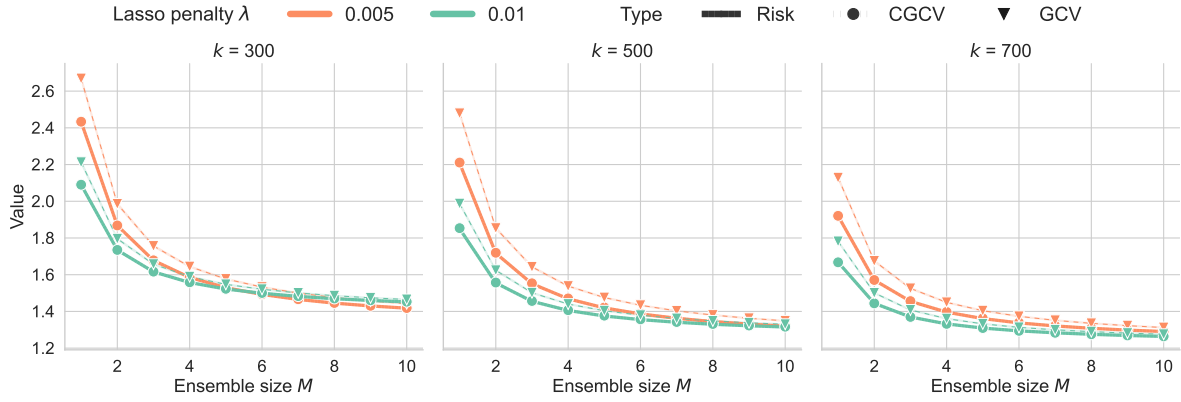
Lasso:

Figure 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
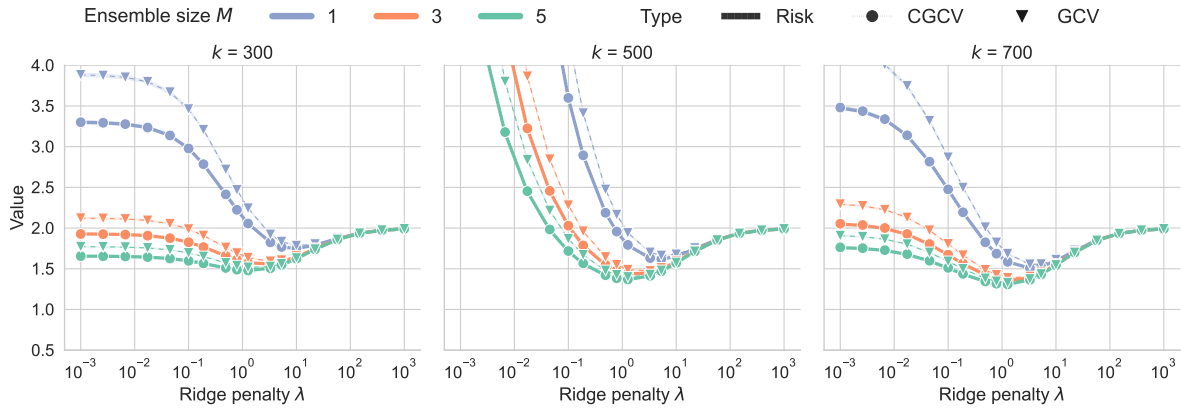
Ridge:

Figure 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.
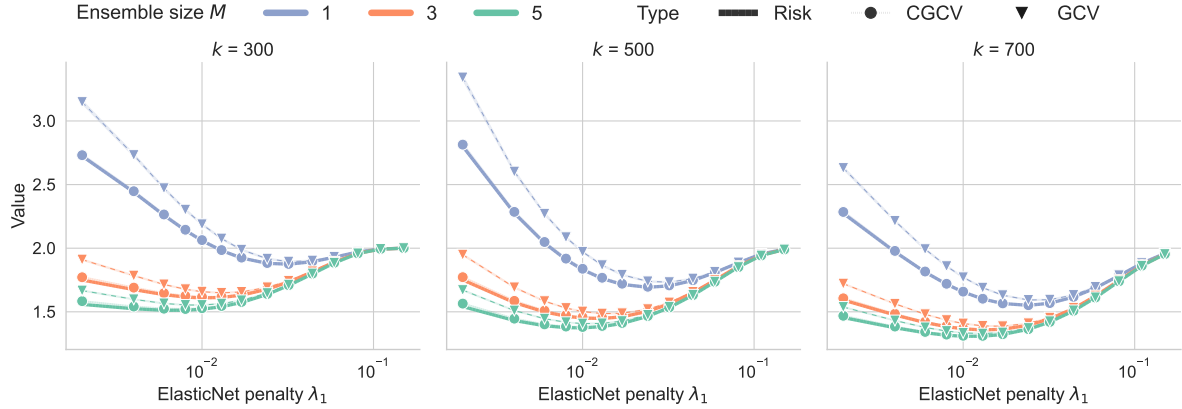
Elastic Net:

Figure 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:

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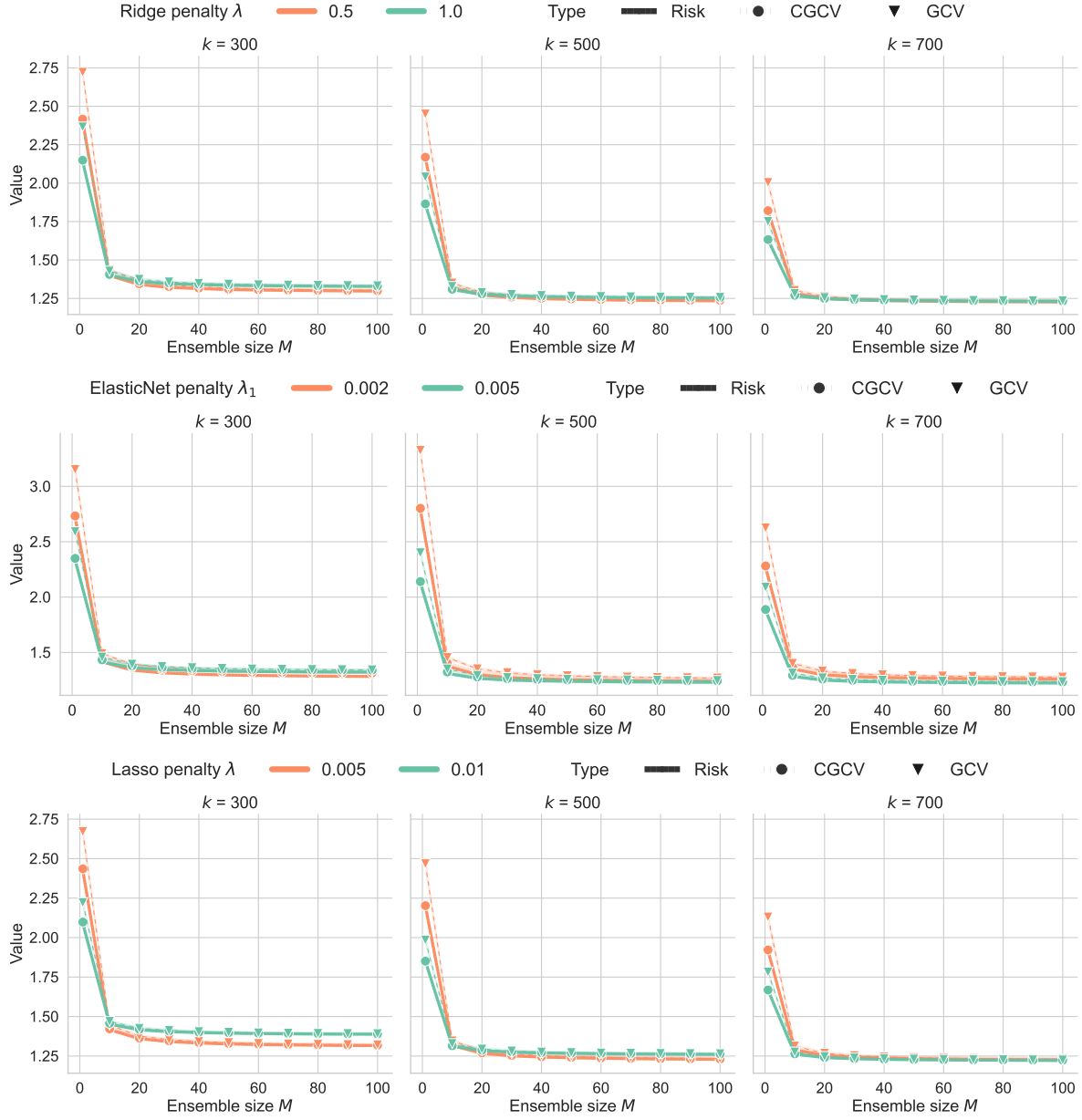
Figure 23: **CGCV efficiently estimates the actual risk for all values of M. GCV gets closer to the actual risk as M increases.** Top row: Ridge; Middle row: Elastic Net; Bottom row: Lasso. The data generating process is the same as in Figure 1.

C.7 Illustration of inconsistency for GCV variant
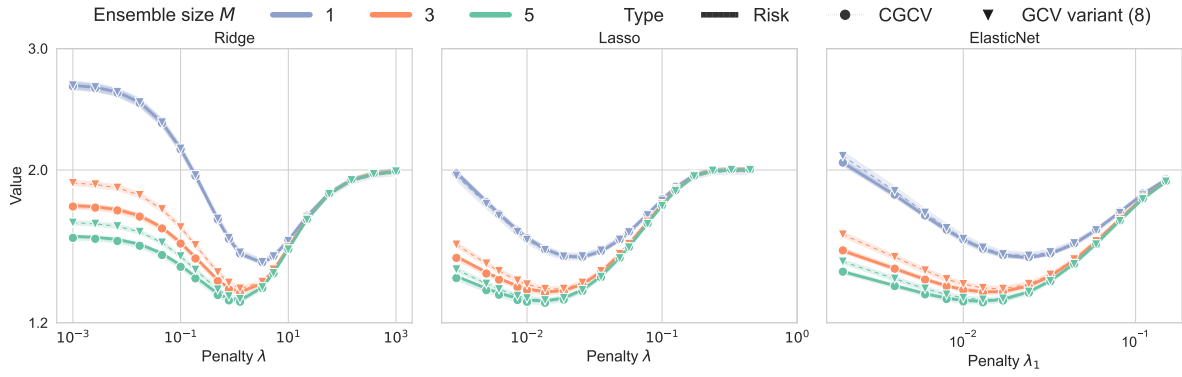
Figure 24: **The GCV variant** (8) **is not consistent for finite** $M > 1$**.** Plot of risks versus the regularization parameter for ridge, lasso, and elastic net ensembles, under different M and fixed $k = 800$. The data generating process is the same as in Figure 1.

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
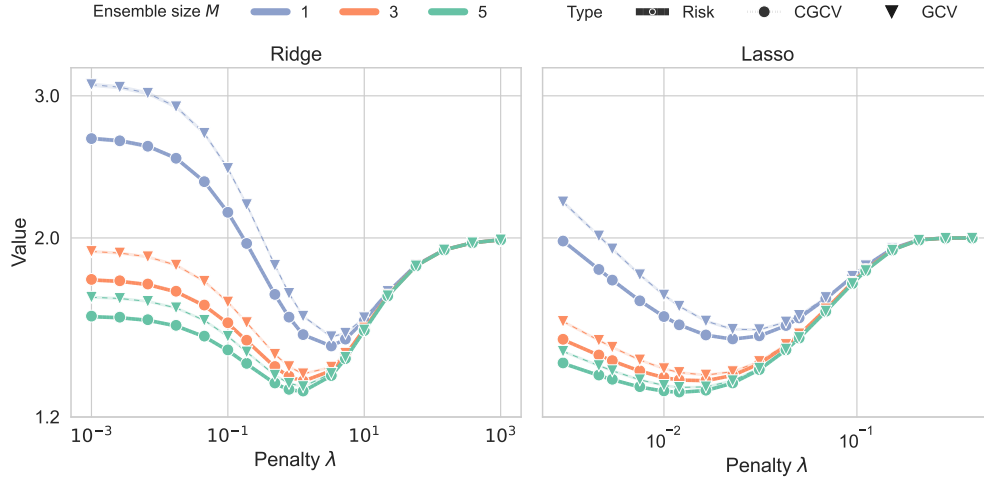
Figure 25: **CGCV is consistent for finite ensembles of penalized estimators while GCV is not.** The simulation setting is the same as Figure 1, except that here $\boldsymbol{X}$ is Rademacher distributed.
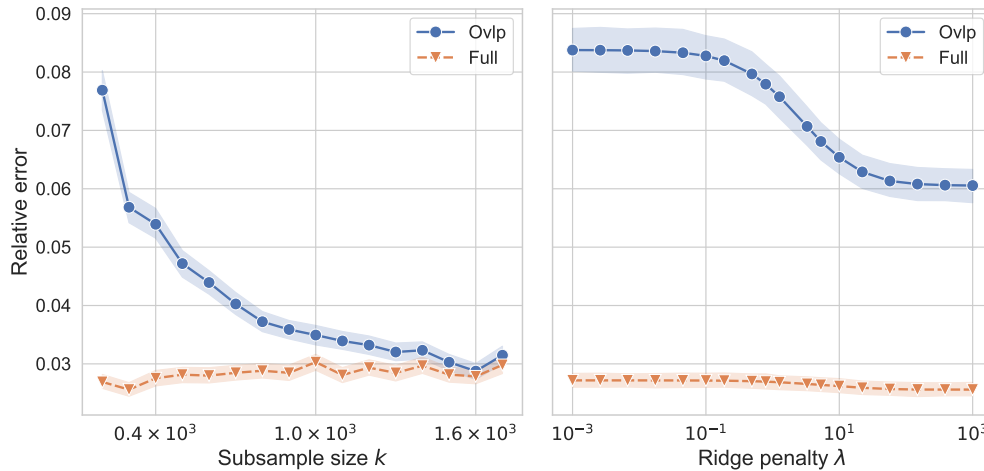
Figure 26: **Full-estimator $\widetilde{R}_M^{\mathbf{full}}$ performs better than ovlp-estimator $\widetilde{R}_M^{\mathbf{ovlp}}$ across different regularization and for small subsample size k.** The simulation setting is the same as Figure 3, except that here $\boldsymbol{X}$ is Rademacher distributed.
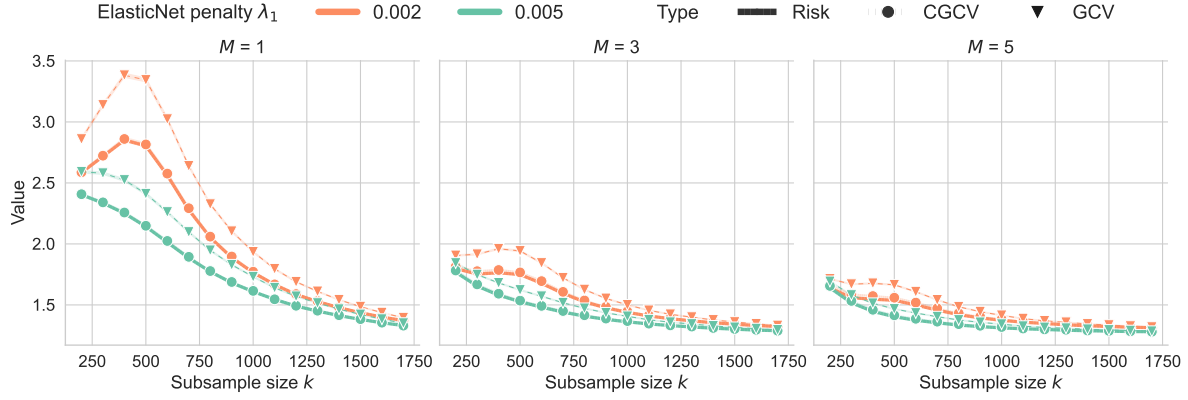
Figure 27: **CGCV is consistent for finite ensembles across different subsample sizes.** The plot of risks versus the subsample size k for Elastic Net ensemble with different λ and M and fixed $\lambda_2 = 0.01$. The simulation setting is the same as Figure 4, except that here $\boldsymbol{X}$ is Rademacher distributed.
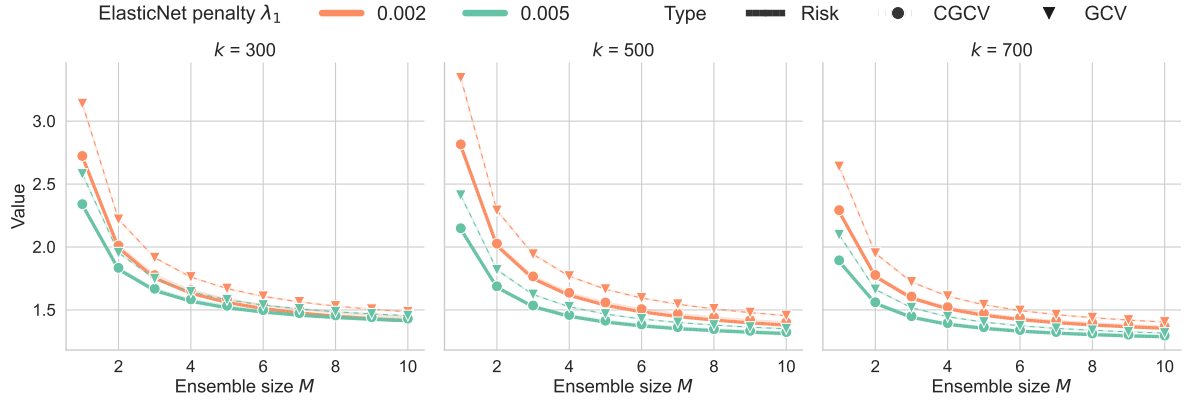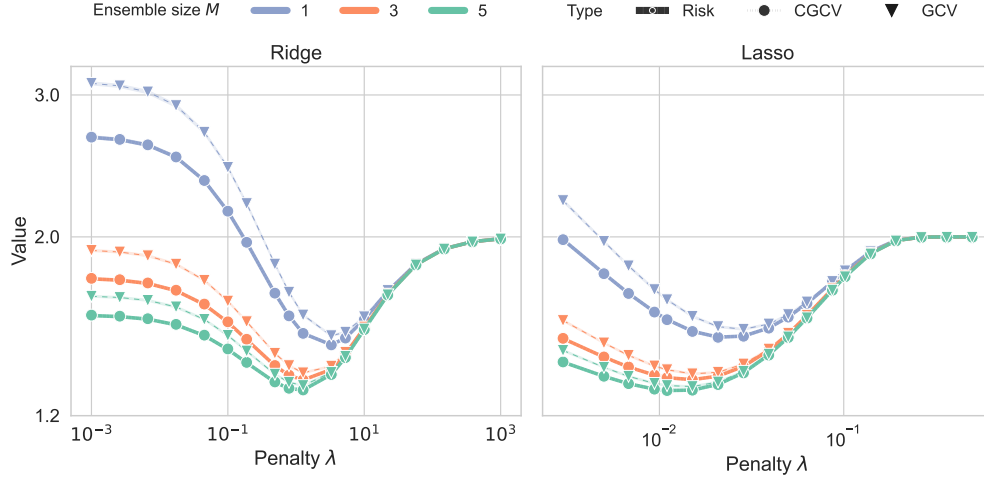

Figure 28: **GCV gets closer to the risk as ensemble size increases across different subsample sizes.** The plot of risks versus the Elastic Net ensemble size M for ridge ensemble with different λ and k and fixed $\lambda_2 = 0.01$. The simulation setting is the same as Figure 5, except that here $\boldsymbol{X}$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

D.2 Results for uniform distribution

Figure 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.
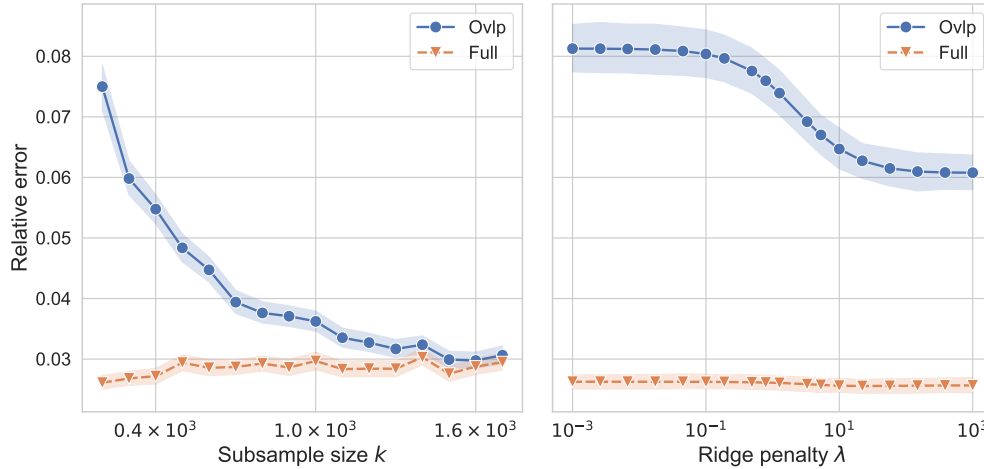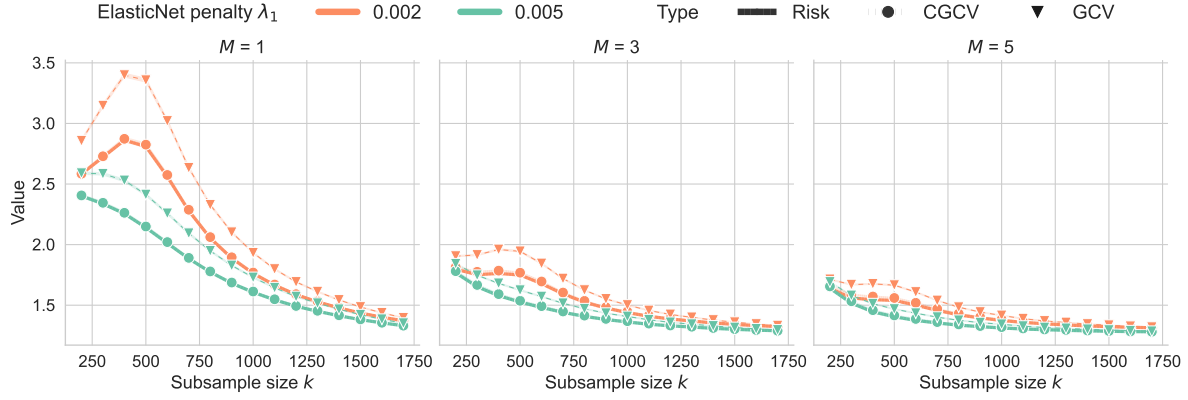

Figure 30: **Full-estimator $\widetilde{R}_M^{\mathbf{full}}$ performs better than ovlp-estimator $\widetilde{R}_M^{\mathbf{ovlp}}$ across different regularization and for small subsample size k.** The simulation setting is the same as Figure 3, except that here $\boldsymbol{X}$ is uniformly distributed.

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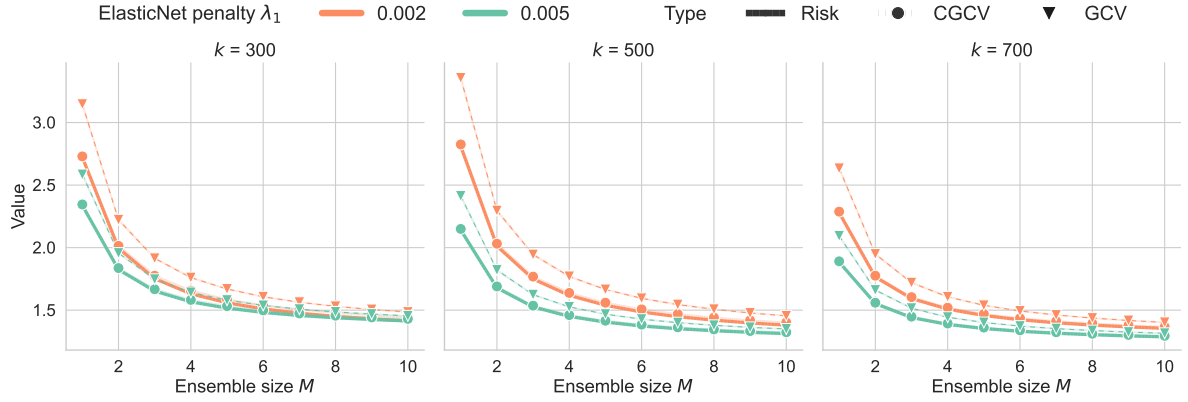
Figure 32: **GCV gets closer to the risk as ensemble size increases across different subsample sizes.** The plot of risks versus the Elastic Net ensemble size M for ridge ensemble with different λ and k and fixed $\lambda_2 = 0.01$. The simulation setting is the same as Figure 5, except that here $\boldsymbol{X}$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

91

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
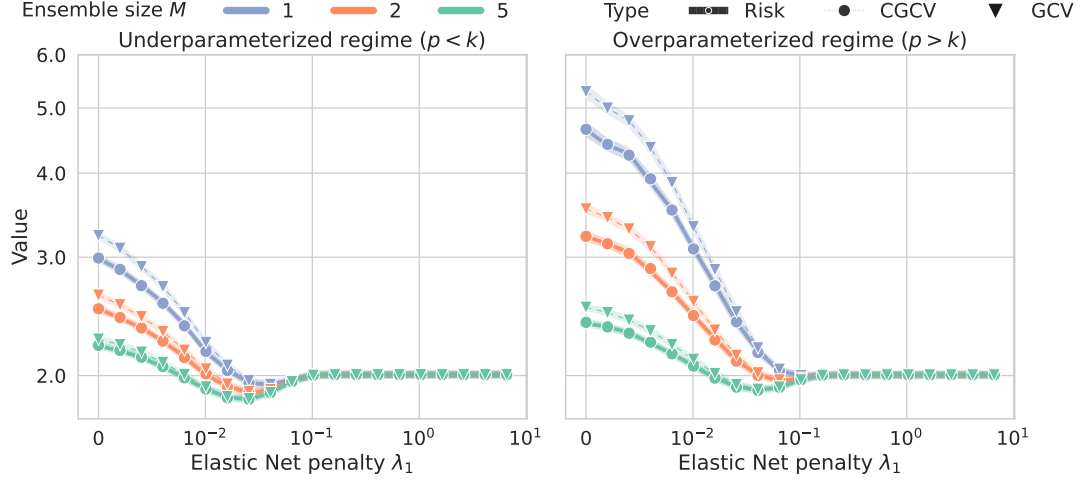
Elastic net:

Figure 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.
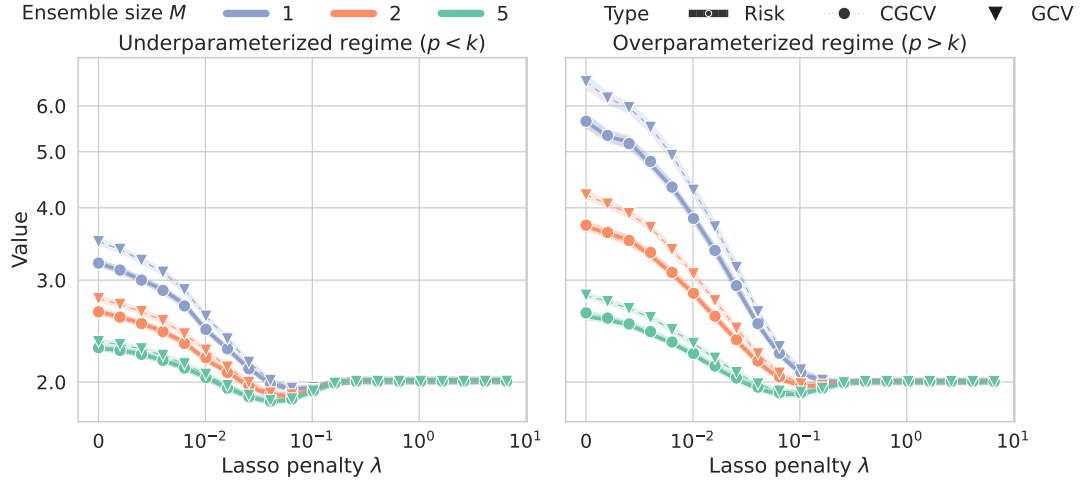
Lasso:

Figure 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.