Journal of Machine Learning Research 24 (2023) 1-113 Submitted 7/23; Revised 10/23; Published 10/23

Bagging in overparameterized learning: Risk characterization and risk monotonization

Pratik Patil† PRATIKPATIL@BERKELEY.EDU
Department of Statistics
University of California, Berkeley
Berkeley, CA 94720, USA

Jin-Hong Du† JINHONGD@ANDREW.CMU.EDU
Department of Statistics and Data Science & Machine Learning Department
Carnegie Mellon University
Pittsburgh, PA 15213, USA

Arun Kumar Kuchibhotla ARUNKU@CMU.EDU
Department of Statistics and Data Science
Carnegie Mellon University
Pittsburgh, PA 15213, USA

Editor: Daniel Hsu

Abstract

Bagging is a commonly used ensemble technique in statistics and machine learning to improve the performance of prediction procedures. In this paper, we study the prediction risk of variants of bagged predictors under the proportional asymptotics regime, in which the ratio of the number of features to the number of observations converges to a constant. Specifically, we propose a general strategy to analyze the prediction risk under squared error loss of bagged predictors using classical results on simple random sampling. Specializing the strategy, we derive the exact asymptotic risk of the bagged ridge and ridgeless predictors with an arbitrary number of bags under a well-specified linear model with arbitrary feature covariance matrices and signal vectors. Furthermore, we prescribe a generic cross-validation procedure to select the optimal subsample size for bagging and discuss its utility to eliminate the non-monotonic behavior of the limiting risk in the sample size (i.e., double or multiple descents). In demonstrating the proposed procedure for bagged ridge and ridgeless predictors, we thoroughly investigate the oracle properties of the optimal subsample size and provide an in-depth comparison between different bagging variants.

Keywords: subagging, divide-and-conquer, proportional asymptotics, ridge regression

1. Introduction

Modern machine learning models often use a large number of parameters relative to the number of observations. In this regime, several commonly used procedures exhibit a peculiar risk behavior, which is referred to as double or multiple descents in the risk profile (Belkin et al., 2019). The precise nature of the double or multiple descent behavior in the generalization error has been studied for various procedures: e.g., linear regression (Belkin

†Equal contribution.

©2023 Pratik Patil, Jin-Hong Du, and Arun Kumar Kuchibhotla.

License: CC-BY 4.0, see https://creativecommons.org/licenses/by/4.0/. Attribution requirements are provided at http://jmlr.org/papers/v24/23-0887.html.

et al., 2020; Muthukumar et al., 2020; Hastie et al., 2022), logistic regression (Deng et al., 2022), random features regression (Mei and Montanari, 2022), kernel regression (Liu et al., 2021), among others. We refer the readers to the survey papers by Bartlett et al. (2021); Belkin (2021); Dar et al. (2021) for a more comprehensive review and other related references. In these cases, the asymptotic prediction risk behavior is often studied as a function of the data aspect ratio (the ratio of the number of parameters/features to the number of observations). The double descent behavior refers to the phenomenon where the (asymptotic) risk of a sequence of predictors first increases as a function of the aspect ratio, peaks at a certain point (or diverges to infinity), and then decreases with the aspect ratio. From a traditional statistical point of view, the desirable behavior as a function of aspect ratio is not immediately obvious. We can, however, reformulate this behavior as a function of $\phi = p/n$, in terms of the observation size n with a fixed p; imagine a large but fixed p and n changing from 1 to ∞. In this reformulation, the double descent behavior translates to a pattern in which the risk first decreases as n increases, then increases, peaks at a certain point, and then decreases again with n. This is a rather counter-intuitive and sub-optimal behavior for a prediction procedure. The least one would expect from a good prediction procedure is that it yields better performance with more information (i.e., more data). However, the aforementioned works show that many commonly used predictors may not exhibit such "good" behavior. Simply put, the non-monotonicity of the asymptotic risk as a function of the number of observations or the limiting aspect ratio implies that more data may hurt generalization (Nakkiran, 2019).

Several ad hoc regularization techniques have been proposed in the literature to mitigate the double/multiple descent behaviors. Most of these methods are trial-and-error in nature in the sense that they do not directly target monotonizing the asymptotic risk but instead try a modification and check that it yields a monotonic risk. The recent work of Patil et al. (2022) introduces a generic cross-validation framework that directly addresses the problem and yields a modification of any given prediction procedure that provably monotonizes the risk. In a nutshell, the method works by training the predictor on subsets of the full data (with different subset sizes) and picking the optimal subset size based on the estimated prediction risk computed using testing data. Intuitively, it is clear that this yields a prediction procedure whose risk is a decreasing function of the observation size. In the proportional asymptotic regime, where $p/n \to \phi$ as $n, p \to \infty$, the paper proves that this strategy returns a prediction procedure whose asymptotic risk is monotonically increasing in ϕ. The paper theoretically analyzes the case where only one subset is used for each subset size and illustrates via numerical simulations that using multiple subsets of the data of the same size (i.e., subsampling) can yield better prediction performance in addition to monotonizing the risk profile. Note that averaging a predictor computed on M different subsets of the data of the same size is referred to in the literature as subagging, a variant of the classical bagging (bootstrap aggregation) proposed by Breiman (1996). The focus of the current paper is to analyze the properties of bagged predictors in two directions (in the proportional asymptotics regime): (1) what is the asymptotic predictive risk of the bagged predictors with M bags as a function of M, and (2) does the cross-validated bagged predictor provably yield improvements over the predictor computed on full data and does it have a monotone risk profile (i.e., the asymptotic risk is a monotonic function of ϕ)?

In this paper, we investigate several variants of bagging, including subagging as a special case. The second variant of bagging, which we call splagging (that stands for **spl**it-**agg**regat**ing**), is the same as the divide-and-conquer or the data-splitting approach (Rosenblatt and Nadler, 2016; Banerjee et al., 2019). The divide-and-conquer approach is widely used in distributed learning, although not commonly featured in the bagging literature (Dobriban and Sheng, 2020, 2021; Mücke et al., 2022). Formally, splagging splits the data into non-overlapping parts of equal size and averages the predictors trained on these non-overlapping parts. We refer to the equal size of each part of the data as subsample size. We use the same terminology for subagging also for the sake of simplicity. Using classical results from survey sampling and some simple lemmas about almost sure convergence, we are able to analyze the behavior of subagged and splagged predictors[1] with M bags for arbitrary prediction procedures and general $M \geq 1$. In fact, we show that the asymptotic risk of bagged predictors for general $M \geq 1$ (or simply, M-bagged predictor) can be written in terms of the asymptotic risks of bagged predictors with $M = 1$ and $M = 2$. Rather interestingly, we prove that the M-bagged predictor's finite sample predictive risk is uniformly close to its asymptotic limit over all $M \geq 1$. These results are established in a model-agnostic setting and do not require the proportional asymptotic regime. Deriving the asymptotic risk behavior of bagged predictors with $M = 1$ and $M = 2$ has to be done on a case-by-case basis, which we perform for ridge and ridgeless prediction procedures. In the context of bagging for general predictors, we further analyze the cross-validation procedure with M-bagged predictors for arbitrary $M \geq 1$ to select the "best" subsample size for both subagging and splagging. These results show that subagging and splagging for any $M \geq 1$ outperform the predictor computed on the full data. We further present conditions under which the cross-validated predictor with M-bagged predictors has an asymptotic risk monotone in the aspect ratio. Specializing these results to the ridge and ridgeless predictors leads to somewhat surprising results connecting subagging to optimal ridge regression as well as the advantages of interpolation.

Before proceeding to discuss our specific contributions, we pause to highlight the two most significant take-away messages from our work. These messages hold under a well-specified linear model, where the features possess an arbitrary covariance structure, and the response depends on an arbitrary signal vector, both of which are subject to certain bounded norm regularity constraints.

(T1) Subagging and splagging (the data-splitting approach) of the ridge and ridgeless predictors, when properly tuned, can significantly improve the prediction risks of these standalone predictors trained on the full data. This improvement is most pronounced near the interpolation threshold. Importantly, subagging always outperforms splagging. See the left panel of Figure 1 for a numerical illustration and Proposition 12 for a formal statement of this result.

(T2) A model-agnostic algorithm exists to tune the subsample size for subagging. This algorithm produces a predictor whose risk matches that of the oracle-tuned subagged predictor. Notably, the oracle-tuned subsample size for the ridgeless predictor is always

[1] A note on terminology for the paper: when referring to subagging and splagging together, we use the generic term bagging. Similarly, when referring to subagged and splagged predictors together, we simply say bagged predictors.

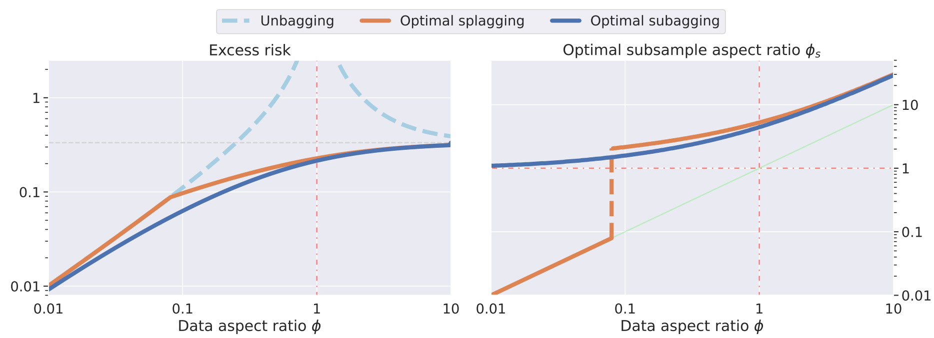

Figure 1: Overview of optimal bagging over both the subsample aspect ratio and the number of bags. (a) Optimal asymptotic excess risk curves for ridgeless predictors with and without bagging, under model (M-AR1-LI) when $\rho_{\text{ar1}} = 0.25$ and $\sigma^2 = 1$. The excess risk is the difference between the prediction risk and the noise level σ^2. The risk for the unbagged ridgeless predictor is represented by a blue dashed line, and the null risk is marked as a gray dotted line. (b) The corresponding optimal limiting subsample aspect ratio $\phi_s = p/k$ versus the data aspect ratio $\phi = p/n$ for bagged ridgeless predictors. The line $\phi_s = \phi$ is colored in green. The optimal subsample aspect ratios are larger than one (above the horizontal red dashed line). See Section 4.3 for more details on the setup and further related discussion.

smaller than the number of features. As a result, subagged ridgeless *interpolators* always outperform subagged least squares, even when the full data has more observations than the number of features. The same observation holds true for splagging whenever it provides an improvement. See the right panel of Figure 1 for numerical illustrations and Proposition 13 for formal statements of this result.

Intuitively, although bagging may induce bias due to subsampling, it can significantly reduce the prediction risk by reducing the variance for a suitably chosen subsample size that is smaller than the feature size. This tradeoff arises because of the different rates at which the bias and variance of the ridgeless predictor increase near the interpolation threshold. This advantage of *interpolation* or *overparameterization* is distinct from other benefits discussed in the literature, such as self-induced regularization (Bartlett et al., 2021).

1.1 Summary of main results

Below we provide a summary of the main results of this paper.

1. **General predictors.** In Section 2, we formulate a generic strategy for analyzing the limiting squared *data conditional risk* (expected squared error on a future data point, conditional on the full data) of general M-bagged predictors, showing that the existence of the limiting risk for $M = 1$ and $M = 2$ implies the existence of the limiting risk for every $M \geq 1$. Moreover, we show that the limiting risk of the M-bagged predictor can be written as a linear combination of the limiting risks of M-bagged predictors with $M = 1$ and $M = 2$. Interestingly, the same strategy also works for analyzing the limit of the *subsample conditional risk*, which considers conditioning on both the full data and the randomly drawn subsamples. See Theorem 5 for a formal statement. In this general framework, Theorem 5 implies that both the data conditional and subsample conditional

risks are asymptotically monotone in the number of bags M. Moreover, for general strongly convex and smooth loss functions, we can sandwich the risks between quantities of the form $\mathfrak{C}_1 + \mathfrak{C}_2/M$, for some fixed random variables $\mathfrak{C}_1$ and $\mathfrak{C}_2$ (Proposition 18).

2. **Ridge and ridgeless predictors.** In Section 3, we specialize the aforementioned general strategy to characterize the data conditional and subsample conditional risks of M-bagged ridge and ridgeless predictors. The results are formalized in Theorem 6 for subagging with and without replacement, and Theorem 8 for splagging without replacement. All these results assume a well-specified linear model, with an arbitrary covariance matrix for the features and an arbitrary signal vector. Notably, we assume neither Gaussian features nor isotropic features nor a randomly generated signal. These results reveal that for the three aforementioned bagging strategies, the bias and variance risk components are non-increasing in the number of bags M.

3. **Cross-validation.** In Section 4, we develop a generic cross-validation strategy to select the optimal subsample or split size (or equivalently, the subsample aspect ratio) and present a general result to understand the limiting risks of cross-validated predictors. Our theoretical results provide a way to verify the monotonicity of the limiting risk of the cross-validated predictor in terms of the limiting data aspect ratio ϕ (Theorem 10). In Section 4.2, we specialize in the cross-validated ridge and ridgeless predictors to obtain the optimal subsample aspect ratio for every M (Theorem 11). Moreover, when optimizing over both the subsample aspect ratio and the number of bags, we show that optimal subagging always outperforms optimal splagging (Proposition 12). Rather surprisingly, in our investigation of the oracle choice of the subsample size for optimal subagging with $M = \infty$, we find that the subsample ratio is always large than one (Proposition 13). In Section 5, we also show optimally-tuned subbaged ridgeless predictor yields the same prediction risk as the optimal ridge predictor for isotropic features (Theorem 16).

From a technical perspective, during the course of our risk analysis of the bagged ridge and ridgeless predictors, we derive novel deterministic equivalents for ridge resolvents with random Tikhonov-type regularization. We extend ideas of conditional asymptotic equivalents and related calculus, which may be of independent interest. See Section H, and in particular Section H.3.2.

1.2 Related work

The risk non-monotonicity of commonly used predictors has been well documented in the literature. For instance, a recent line of work by Belkin et al. (2019); Viering et al. (2019); Nakkiran (2019); Loog et al. (2019), among others, illustrates the non-monotonic risk behavior of several prediction procedures. See also the survey papers by Belkin (2021); Bartlett et al. (2021); Dar et al. (2021); Loog et al. (2020) for other related references. As highlighted by Loog et al. (2020), the phenomenon of multiple descents can be traced back to empirical findings in the 1990s, including earlier papers by Vallet et al. (1989); Hertz et al. (1989); Opper et al. (1990); Hansen (1993); Barber et al. (1995); Duin (1995); Opper (1995); Opper and Kinzel (1996); Raudys and Duin (1998), among others.

Since non-monotonic risk leads to suboptimal use of the data, several methods have been proposed that modify a given (class of) prediction procedure(s) to construct a new prediction procedure with a monotonic risk profile. In particular, Nakkiran et al. (2021) investigates the role of optimal tuning in the context of ridge regression and demonstrates that the optimally-tuned ℓ_2 regularization achieves monotonic generalization performance for a class of linear models under isotropic design. Mhammedi (2021) provides an algorithm to monotonize the risk profile for bounded loss functions. Patil et al. (2022) propose a general framework to monotonize the prediction risk for general predictors under both bounded and unbounded loss functions, using cross-validation. The paper also empirically shows that bagging can further improve the performance of the predictors while achieving a monotonized risk profile. In this paper, we characterize the risk behavior of bagging, which was left as an open direction in Patil et al. (2022). Below we provide a brief overview of the literature pertaining to bagging and its relation to our work.

Ensemble methods are widely used in machine learning and statistics and combine several weak predictors to produce one powerful predictor. One important class of ensemble methods is bagging (Breiman, 1996; Bühlmann and Yu, 2002), and its variants, such as subagging (Bühlmann and Yu, 2002), that operate by averaging predictors trained on independent subsamples of the data. Numerous empirical studies have demonstrated that bagging leads to significant improvements in predictive performance (Breiman, 1996; Strobl et al., 2009; Fernández-Delgado et al., 2014). However, the theoretical analysis of bagging has primarily focused on smooth predictors (predictors that are smooth functions of the empirical data distribution); see Buja and Stuetzle (2006); Friedman and Hall (2007). For some work on bagging for non-parametric estimators, see Hall and Samworth (2005); Samworth (2012); Wu et al. (2021); Bühlmann and Yu (2002); Athey et al. (2019). In addition to sample-wise bagging, bagging over linear combinations of features has also been considered in Lopes et al. (2011); Srivastava et al. (2016); Cannings and Samworth (2017). This approach broadly falls under the umbrella of feature side sketching; we refer readers to Wang et al. (2017); Derezinski et al. (2020); Lopes et al. (2018); Dereziński (2023); LeJeune et al. (2022); Patil and LeJeune (2023), among others, for related results and further references.

Bagging in the proportional asymptotic regime has also been considered in the literature. LeJeune et al. (2020) study subagging of both features and observations and derive the limiting risk of the resulting subagged predictor. Dobriban and Sheng (2020, 2021); Mücke et al. (2022) consider the divide-and-conquer approach, or splagging, and investigate their properties. These works are set in the context of distributed learning. Specifically, under proportional asymptotics, Dobriban and Sheng (2020) derive the limiting mean squared error of the distributed ridge estimator in the underparameterized regime. On the other hand, Mücke et al. (2022) provide finite-sample upper bounds on the prediction risk for ridgeless regression in the overparameterized regime.

The closest works to ours are those of LeJeune et al. (2020) and Mücke et al. (2022). LeJeune et al. (2020) investigate bagged least squares predictor obtained by subsampling both features and observations in a Gaussian isotropic design. They impose a restriction on subsampling such that the final subsampled data always has more observations than the features (so that ordinary least squares are well-defined). Consequently, they do not allow for overparameterized subsampled datasets. Similar to our work, they also study the monotonicity of the asymptotic expected squared risk with respect to the number of bags in their restricted setting. Further, they study the best subsampling ratios for optimal asymptotic risk, but do not consider the question of how to select the best subsample size. The most significant difference between their work and ours is that we subsample observations, and they effectively subsample features, which is only appropriate under isotopic covariance. On the other hand, Mücke et al. (2022) consider splagging and provide finite-sample upper bounds on the bias and variance components of the squared prediction risk under the assumption of sub-Gaussian features. In contrast, our results do not assume sub-Gaussianity for either the feature or response distributions and only impose minimal bounded moment assumptions.

1.3 Organization

The rest of the paper is organized as follows. In Section 2, we provide risk decompositions conditional on both the full dataset and subsampled datasets for different bagging variants for general predictors. Based on the form of decompositions, we provide a series of reductions and a generic strategy for analyzing the squared prediction risk of general bagged predictors. In Section 3, we give risk characterizations for bagging ridge and ridgeless predictors. We give results for both subagging with and without replacement and splagging without replacement, and show monotonicities of the bias and variance components in the number of bags. In Section 4, we prescribe a framework for monotonizing the risk profile of any given predictor based on cross-validation over subsample size. The result is then specialized to the ridge and ridgeless predictors. Furthermore, we compare the monotonized risk profiles of bagged ridgeless and ridge predictors. In Section 5, we specialize our results for isotropic features and provide explicit analytic expressions for the risks of bagged ridgeless regression. In addition, we present the analysis of the optimal subsample size and the corresponding optimal bagged risk. In Section 6, we conclude the paper by discussing related open questions.

In the supplement to this paper, we provide proof of all the results. The organization structure for the supplement is outlined in the first section of the supplement, which also presents an overview of the general notation employed throughout the paper. The source code generating all the experimental illustrations in this paper can be accessed at `https://jaydu1.github.io/overparameterized-ensembling/bagging/`.

2. Bagging general predictors

In this section, we will describe different versions of subagged predictors. But first, let us define the index sets pertinent to our study. Fix any $k \in \{1, 2, \ldots, n\}$ and any permutation $\pi : \{1, 2, \ldots, n\} \to \{1, 2, \ldots, n\}$. Define the sets $\mathcal{I}_k$ and $\mathcal{I}_k^\pi$ as follows:

$$\begin{aligned} \mathcal{I}_k &:= \{\{i_1, i_2, \ldots, i_k\} : 1 \le i_1 < i_2 < \ldots < i_k \le n\}, \\ \mathcal{I}_k^\pi &:= \left\{\{\pi((j-1)k+1), \pi((j-1)k+2), \ldots, \pi(jk)\} : 1 \le j \le \left\lfloor \frac{n}{k} \right\rfloor\right\}. \end{aligned} \tag{1}$$

Note that both the sets $\mathcal{I}_k$ and $\mathcal{I}_k^\pi$ technically need to be indexed by n, but for notation convenience, we will not explicitly indicate the dependence on n. The set $\mathcal{I}_k$ represents the set of all k subset choices from $\{1, 2, \ldots, n\}$. There are $\binom{n}{k}$ many of them. The set $\mathcal{I}_k^\pi$, on the other hand, represents the set of indices in a non-overlapping split of $\{1, 2, \ldots, n\}$ into blocks of size k. If we split $\{1, 2, \ldots, n\}$ randomly into different non-overlapping blocks each of size k, then this corresponds to choosing a permutation π randomly from the set of all permutations and splitting them in order. Finally, observe that $\mathcal{I}_k^\pi \subseteq \mathcal{I}_k$ for any permutation π and $\cup_\pi \mathcal{I}_k^\pi = \mathcal{I}_k$.

Suppose now $\mathcal{D}_n = \{(\boldsymbol{x}_1, y_1), \dots, (\boldsymbol{x}_n, y_n)\}$ represents a dataset with random vectors from $\mathbb{R}^p \times \mathbb{R}$. A prediction procedure $\widehat{f}(\cdot;\cdot)$ is defined as a map from $\mathbb{R}^p \times \mathscr{P}(\mathcal{D}_n) \to \mathbb{R}$, where $\mathscr{P}(A)$ for any set A represents the power set of A. For any $I \in \mathcal{I}_k$ (or $I \in \mathcal{I}_k^\pi$), let $\mathcal{D}_I$ and the corresponding subsampled predictor be defined as $\mathcal{D}_I = \{(\boldsymbol{x}_j, y_j) : j \in I\}$ and $\widehat{f}(\boldsymbol{x}; \mathcal{D}_I) = \widehat{f}(\boldsymbol{x}; \{(\boldsymbol{x}_j, y_j) : j \in I\})$. Given two sets of indices and two types of simple random samplings one can draw, we have four different versions of subagged predictors. When employing simple random sampling with replacement, the corresponding predictors can be expressed as follows:

$$\widetilde{f}^{\texttt{WR}}_{M,\mathcal{I}_k}(\boldsymbol{x}; \{\mathcal{D}_{I_\ell}\}_{\ell=1}^M) = \frac{1}{M}\sum_{\ell=1}^{M} \widehat{f}(\boldsymbol{x}; \mathcal{D}_{I_\ell}), \quad \text{where} \quad I_1, \dots, I_M \overset{\texttt{SRSWR}}{\sim} \begin{cases} \mathcal{I}_k & \text{for subagging} \\ \mathcal{I}_k^\pi & \text{for splagging,} \end{cases} \tag{2}$$

and the predictors $\widetilde{f}^{\texttt{WOR}}_{M,\mathcal{I}_k}$ using simple random sampling without replacement are defined analogously.

Traditionally, bagging (as in **b**ootstrap-**agg**regat**ing**) refers to computing predictors multiple times based on bootstrapped data (Breiman, 1996), which can involve repeated observations. In this paper, we do not allow for repeated observations and consider only the four versions of bagging mentioned in (2). Bühlmann and Yu (2002, Section 3.2) call $\widetilde{f}^{\texttt{WR}}_{M,\mathcal{I}_k}$ as subagging (as in **sub**sample-**agg**regat**ing**). Given that SRSWOR mean estimator has a smaller mean squared error than SRSWR mean estimator, we also consider the variant $\widetilde{f}^{\texttt{WOR}}_{M,\mathcal{I}_k}$ of subagging. Because for any fixed M, the expectation and variance of $\widetilde{f}^{\texttt{WR}}_{M,\mathcal{I}_k}$ and $\widetilde{f}^{\texttt{WOR}}_{M,\mathcal{I}_k}$ are the same as $N \to \infty$, the asymptotic risk behavior of $\widetilde{f}^{\texttt{WR}}_{M,\mathcal{I}_k}$ and $\widetilde{f}^{\texttt{WOR}}_{M,\mathcal{I}_k}$ is the same if $|\mathcal{I}_k| = \binom{n}{k} \to \infty$ (which holds, for example, if $1 \le k \le n-1$ and $n \to \infty$). Given this equivalence and the relative prevalence of subagging (i.e., $\widetilde{f}^{\texttt{WR}}_{M,\mathcal{I}_k}$), in Section 3.2, we focus our results on $\widetilde{f}^{\texttt{WR}}_{M,\mathcal{I}_k}$ although we indicate the implications for $\widetilde{f}^{\texttt{WOR}}_{M,\mathcal{I}_k}$. In what follows, we refer to $\widetilde{f}^{\texttt{WR}}_{M,\mathcal{I}_k}$ and $\widetilde{f}^{\texttt{WOR}}_{M,\mathcal{I}_k}$ as *subagging with and without replacement*, respectively.

In contrast, the predictors $\widetilde{f}^{\texttt{WR}}_{M,\mathcal{I}_k^\pi}$ and $\widetilde{f}^{\texttt{WOR}}_{M,\mathcal{I}_k^\pi}$ do not frequently appear in the bagging literature. Rather, they are more common in distributed learning, where the predictors are trained on different parts of the data and averaged to yield a final predictor. We call these versions as "splagging" (as in **spl**it-**agg**regat**ing**). Among these, the without replacement predictor $\widetilde{f}^{\texttt{WOR}}_{M,\mathcal{I}_k^\pi}$ tends to be more prevalent (Dobriban and Sheng, 2020; Mücke et al., 2022). Owing to its popularity and the fact that SRSWOR is superior to SRSWR in general, in Section 3.3, we primarily focus on $\widetilde{f}^{\texttt{WOR}}_{M,\mathcal{I}_k^\pi}$. In what follows, we refer to $\widetilde{f}^{\texttt{WR}}_{M,\mathcal{I}_k^\pi}$ and $\widetilde{f}^{\texttt{WOR}}_{M,\mathcal{I}_k^\pi}$ as *splagging with and without replacement.* For the sake of simplicity, we define $\widetilde{f}^{\texttt{WOR}}_{M,\mathcal{I}_k^\pi}$ as $\widetilde{f}^{\texttt{WOR}}_{\lfloor n/k \rfloor,\mathcal{I}_k^\pi}$ if $M > \lfloor n/k \rfloor$. In doing so, we are effectively substituting M with $\min\{M, \lfloor n/k \rfloor\}$.

The results to be discussed below are general and apply to all four versions of the bagged predictors in (2). Consider the finite population $\{\widehat{f}(\boldsymbol{x}; \mathcal{D}_I) : I \in \mathcal{I}_k\}$ or $\{\widehat{f}(\boldsymbol{x}; \mathcal{D}_I) : I \in \mathcal{I}_k^\pi\}$, but with the data $\mathcal{D}_n$ treated as fixed (non-stochastic). We know that $\widetilde{f}^{\texttt{WR}}_{M,\mathcal{I}_k}(\boldsymbol{x})$ and $\widetilde{f}^{\texttt{WOR}}_{M,\mathcal{I}_k}(\boldsymbol{x})$ has the same expectation, given by

$$\widetilde{f}_{\infty,\mathcal{I}_k}(\boldsymbol{x}) = \frac{1}{|\mathcal{I}_k|}\sum_{I \in \mathcal{I}_k} \widehat{f}(\boldsymbol{x}; \mathcal{D}_I).$$

However, the variance is smaller for $\widetilde{f}^{\texttt{WOR}}_{M,\mathcal{I}_k}(\boldsymbol{x})$. Using the bias and variance formulas from Chaudhuri (2014, Section 2.5), the following result can be derived for the subagged predictors.

Proposition 1 (Conditional risk decomposition)**.** *Without any assumptions on the data and the prediction procedure $\widehat{f}(\cdot;\cdot)$, we have for every $(\boldsymbol{x}, y) \in \mathbb{R}^p \times \mathbb{R}$,*

$$\begin{aligned} \mathbb{E}[(y - \widetilde{f}^{\texttt{WR}}_{M,\mathcal{I}_k}(\boldsymbol{x}; \{\mathcal{D}_{I_\ell}\}_{\ell=1}^M))^2 \mid \mathcal{D}_n] &= \mathscr{B}_{\mathcal{I}_k}(\boldsymbol{x}, y) + \frac{1}{M}\mathscr{V}_{\mathcal{I}_k}(\boldsymbol{x}, y), \\ \mathbb{E}[(y - \widetilde{f}^{\texttt{WOR}}_{M,\mathcal{I}_k}(\boldsymbol{x}; \{\mathcal{D}_{I_\ell}\}_{\ell=1}^M))^2 \mid \mathcal{D}_n] &= \mathscr{B}_{\mathcal{I}_k}(\boldsymbol{x}, y) + \frac{|\mathcal{I}_k| - M}{|\mathcal{I}_k| - 1}\frac{1}{M}\mathscr{V}_{\mathcal{I}_k}(\boldsymbol{x}, y), \end{aligned} \tag{3}$$

where $\mathscr{B}_{\mathcal{I}_k}(\boldsymbol{x}, y) = (y - \widetilde{f}_{\infty,\mathcal{I}_k}(\boldsymbol{x}))^2$, *and* $\mathscr{V}_{\mathcal{I}_k}(\boldsymbol{x}, y) = \frac{1}{|\mathcal{I}_k|}\sum_{I \in \mathcal{I}_k}\left(\widehat{f}(\boldsymbol{x}; \mathcal{D}_I) - \widetilde{f}_{\infty,\mathcal{I}_k}(\boldsymbol{x})\right)^2.$ (4)

The results still hold by replacing $\mathcal{I}_k$ with $\mathcal{I}_k^\pi$. Here in (3)*, the expectation is with respect to the randomness of $I_1, \ldots, I_M$ only.*

In line with traditional predictive thinking, we care about the performance of our predictors computed on $\mathcal{D}_n$ on future data from the same distribution P. As we have access to a single dataset $\mathcal{D}_n$, we consider the behavior of the predictors in terms of the conditional risk, conditional on $\mathcal{D}_n$. To be precise, for a predictor $\widehat{f}$ fitted on $\mathcal{D}_n$ and its subagged predictor $\widetilde{f}^{\texttt{WR}}_{M,\mathcal{I}_k}$ fitted on $\{\mathcal{D}_{I_\ell}\}_{\ell=1}^M$, with $I_1, \ldots, I_M$ being M samples of size k from $\mathcal{I}_k$, the conditional risks (conditional on $\mathcal{D}_n$) are defined as follows:

$$\begin{aligned} R(\widehat{f}; \mathcal{D}_n) &:= \int (y - \widehat{f}(\boldsymbol{x}; \mathcal{D}_n))^2 \,\mathrm{d}P(\boldsymbol{x}, y), \\ R(\widetilde{f}^{\texttt{WR}}_{M,\mathcal{I}_k}(\cdot; \{\mathcal{D}_{I_\ell}\}_{\ell=1}^M); \mathcal{D}_n) &:= \int \mathbb{E}\left[\left(y - \widetilde{f}^{\texttt{WR}}_{M,\mathcal{I}_k}(\boldsymbol{x}; \{\mathcal{D}_{I_\ell}\}_{\ell=1}^M)\right)^2 \,\middle|\, \mathcal{D}_n\right] \mathrm{d}P(\boldsymbol{x}, y). \end{aligned} \tag{5}$$

The conditional risk of $\widetilde{f}^{\texttt{WOR}}_{M,\mathcal{I}_k}(\cdot; \{\mathcal{D}_{I_\ell}\}_{\ell=1}^M)$ is defined similarly, and so are the conditional risks for the splagged predictors with and without replacement from $\mathcal{I}_k^\pi$ for a fixed permutation π. Observe that the conditional risk of the subagged predictor $\widetilde{f}^{\texttt{WR}}_{M,\mathcal{I}_k}(\cdot; \{\mathcal{D}_{I_\ell}\}_{\ell=1}^M)$ integrates over the randomness of the future observation $(\boldsymbol{x}, y)$ as well as the randomness due to the simple random sampling of I_ℓ, $\ell = 1, \ldots, M$. Given that only a single dataset $\mathcal{D}_n$ is observed in practice and one typically only draws one simple random sample I_ℓ, $\ell = 1, \ldots, M$, it is also insightful to consider an alternate version of the conditional risk that ignores the expectation over the simple random sample:

$$R(\widetilde{f}^{\texttt{WR}}_{M,\mathcal{I}_k}(\cdot; \{\mathcal{D}_{I_\ell}\}_{\ell=1}^M); \mathcal{D}_n, \{I_\ell\}_{\ell=1}^M) := \int \left(y - \widetilde{f}^{\texttt{WR}}_{M,\mathcal{I}_k}(\boldsymbol{x}; \{\mathcal{D}_{I_\ell}\}_{\ell=1}^M)\right)^2 \mathrm{d}P(\boldsymbol{x}, y). \tag{6}$$

We call the former type of conditional risk (conditional on $\mathcal{D}_n$) as *data conditional risk* and the latter type of conditional risk (conditional on $\mathcal{D}_n$ and $\{I_\ell\}_{\ell=1}^M$) as *subsample conditional risk*.

Proposition 1 implies that the data conditional risks of the predictors $\widetilde{f}^{\texttt{WR}}_{M,\mathcal{I}_k}(\cdot)$ and $\widetilde{f}^{\texttt{WOR}}_{M,\mathcal{I}_k}(\cdot)$ can be written as

$$
\begin{aligned}
R(\widetilde{f}_M;\mathcal{D}_n) &= \int \mathscr{B}_{\mathcal{I}_k}(\boldsymbol{x},y)\,\mathrm{d}P(\boldsymbol{x},y) + K_{|\mathcal{I}_k|,M}\frac{1}{M}\int \mathscr{V}_{\mathcal{I}_k}(\boldsymbol{x},y)\,\mathrm{d}P(\boldsymbol{x},y) \\
&= R(\widetilde{f}_\infty;\mathcal{D}_n) + \frac{K_{|\mathcal{I}_k|,M}}{M}\int \mathscr{V}_{\mathcal{I}_k}(\boldsymbol{x},y)\,\mathrm{d}P(\boldsymbol{x},y),
\end{aligned}
\tag{7}
$$

where for $N \geq 1$, $K_{N,M}$ is defined as

$$
K_{N,M} = \begin{cases} 1 & \text{if } \widetilde{f} = \widetilde{f}^{\texttt{WR}}_{M,\mathcal{I}_k} \\ (N-M)_+/(N-1) & \text{if } \widetilde{f} = \widetilde{f}^{\texttt{WOR}}_{M,\mathcal{I}_k}. \end{cases} \tag{8}
$$

The advantage of the representation (7) for the data conditional risk of $\widetilde{f}^{\texttt{WR}}_{M,\mathcal{I}_k}(\cdot)$ and $\widetilde{f}^{\texttt{WOR}}_{M,\mathcal{I}_k}(\cdot)$ is that it allows us to obtain the limiting behavior of their risks for any $M \geq 1$ by just studying their limiting risk behavior for $M = 1$ and $M = 2$. This is trivially shown by solving a system of linear equations in two variables and is formalized in the following result.

Proposition 2 (Data conditional risk for arbitrary M)**.** *Let $R(\widetilde{f}_M;\mathcal{D}_n)$ be as defined in* (5)*. For $\widetilde{f}_M \in \{\widetilde{f}^{\texttt{WR}}_{M,\mathcal{I}_k}, \widetilde{f}^{\texttt{WOR}}_{M,\mathcal{I}_k}, \widetilde{f}^{\texttt{WR}}_{M,\mathcal{I}^\pi_k}, \widetilde{f}^{\texttt{WOR}}_{M,\mathcal{I}^\pi_k}\}$, suppose there exist non-stochastic numbers a_1 and a_2 such that as $n \to \infty$,*

$$
|R(\widetilde{f}_M;\mathcal{D}_n) - a_M| \xrightarrow{\text{a.s.}} 0, \quad \textit{for} \quad M = 1, 2, \tag{9}
$$

where the almost sure convergence is with respect to the randomness of $\mathcal{D}_n$. Then, we have[2]

$$
\sup_{M\in\mathbb{N}} \left| R(\widetilde{f}_M;\mathcal{D}_n) - \left[(2a_2 - a_1) + \frac{2(a_1 - a_2)}{M}\right]\right| \xrightarrow{\text{a.s.}} 0. \tag{10}
$$

Note that according to Proposition 1, we have $a_1 \geq a_2$, irrespective of the prediction procedure. In Proposition 2, if $a_1 > a_2$ (instead of just $a_1 \geq a_2$), then the asymptotic approximations of the conditional risk $R(\widetilde{f}_M;\mathcal{D}_n)$ are strictly decreasing in M. Similarly, we can also derive the asymptotic subsample conditional risk defined in (6) of subagged predictors with an arbitrary number of bags M if we know the limiting risk for $M = 1$ and $M = 2$, as summarized in Proposition 3 below.

Proposition 3 (Subsample conditional risk for arbitrary M)**.** *Let $R(\widetilde{f}_M;\mathcal{D}_n,\{I_\ell\}_{\ell=1}^M)$ be as defined in* (6)*. For $\widetilde{f}_M \in \{\widetilde{f}^{\texttt{WR}}_{M,\mathcal{I}_k}, \widetilde{f}^{\texttt{WOR}}_{M,\mathcal{I}_k}, \widetilde{f}^{\texttt{WR}}_{M,\mathcal{I}^\pi_k}, \widetilde{f}^{\texttt{WOR}}_{M,\mathcal{I}^\pi_k}\}$, suppose there exist non-stochastic numbers b_1 and b_2 such that*

$$
|R(\widetilde{f}_1;\mathcal{D}_n, I^{(n)}) - b_1| \xrightarrow{\text{a.s.}} 0, \quad \textit{for all } I^{(n)} \in \mathcal{I}_k \textit{ or } \mathcal{I}^\pi_k, \tag{11}
$$

$$
|R(\widetilde{f}_2;\mathcal{D}_n, \{I^{(n)}_1, I^{(n)}_2\}) - b_2| \xrightarrow{\text{a.s.}} 0, \quad \textit{for random samples } I^{(n)}_1, I^{(n)}_2{}^{3}, \tag{12}
$$

[2] For SRSWOR, supremum over $M \in \mathbb{N}$ should be understood as either $M \leq |\mathcal{I}_k|$ or $M \leq |\mathcal{I}^\pi_k|$ depending on whether $\widetilde{f}_M$ is $\widetilde{f}^{\texttt{WOR}}_{M,\mathcal{I}_k}$ or $\widetilde{f}^{\texttt{WOR}}_{M,\mathcal{I}^\pi_k}$. The same convention is used for all the other results in this section.

where the almost sure convergence is with respect to the randomness of both $\mathcal{D}_n$ *and* I *(or* I_1*,* I_2*). For any* $M \in \mathbb{N}$*, suppose* $\{I_\ell\}_{\ell=1}^M$ *is a simple random sample according to the definition of* $\widetilde{f}_M$*. Then*

$$\sup_{M\in\mathbb{N}} \left| R(\widetilde{f}_M; \mathcal{D}_n, \{I_\ell\}_{\ell=1}^M) - \left[(2b_2 - b_1) + \frac{2(b_1 - b_2)}{M}\right]\right| \xrightarrow{\mathrm{p}} 0. \tag{13}$$

A few remarks regarding the assumptions of Proposition 3 are warranted. First, the requirement (11) may, on the surface, seem demanding as it necessitates almost sure convergence for all $I \in \mathcal{I}_k$. However, recall that, for any fixed $I \in \mathcal{I}_k$, $\widetilde{f}_{1,\mathcal{I}_k}(\cdot; \mathcal{D}_I)$ is the same as the prediction procedure $\widehat{f}$ computed on the subset $\mathcal{D}_I$ with cardinality k. Consequently, if the original prediction procedure exhibits almost sure convergence as the training sample size goes to ∞, then as $k \to \infty$, the requirement (11) holds for every fixed $I \in \mathcal{I}_k$. Second, in Propositions 2 and 3, we noticed that only the limiting risks for $M = 1$ and $M = 2$ matter. This is because the data conditional risk can be split as:

$$R(\widetilde{f}_{M,\mathcal{I}_k}; \mathcal{D}_n) = -\left(1 - \frac{2}{M}\right) R(\widetilde{f}_{1,\mathcal{I}_k}; \mathcal{D}_n) + 2\left(1 - \frac{1}{M}\right) R(\widetilde{f}_{2,\mathcal{I}_k}; \mathcal{D}_n).$$

The subsample conditional risk admits similar decomposition as well. See Section B for the derivations for both of them. Essentially, the interaction of subsampled datasets is only up to order two. This may not be true for other loss functions. However, a simple monotonicity property and bounds can be obtained for a large class of loss functions. See Proposition 18 in the supplement. It is also worth mentioning that while Propositions 2 and 3 are derived under the assumption that the distribution of the out-of-sample test point $(\boldsymbol{x}, y)$, $P(\boldsymbol{x}, y)$, is the same as the distribution of the training data, it is not difficult to see that the same conclusions hold for a test point sampled from any arbitrary distribution. The results are thus also applicable to out-of-distribution scenarios.

The forthcoming lemma establishes a connection between the data conditional risk and the subsample conditional risk for $M = 1, 2$. In practice, the ingredient predictor is fitted on the subsampled datasets and the subsample conditional risk is evaluated on these subsampled datasets. Thanks to Lemma 4, we are able to infer the data conditional risk based on the subsample conditional risk for the simple cases of $M = 1, 2$.

Lemma 4 (From subsample conditional to data conditional risk for $M = 1, 2$)**.** *If the conditions in Proposition 3 hold, then* (9) *holds with* $a_M = b_M$ *for* $M = 1, 2$*. As a result, the conclusions of Proposition 2 hold.*

It is worth highlighting the proof of Lemma 4 leverages the convexity of the square loss function. Therefore, analogous results can be obtained for other convex loss functions, as long as the limiting subsample conditional risks exist for $M = 1, 2$.

Finally, combining Proposition 2, Proposition 3, and Lemma 4 yields a general strategy for obtaining both limiting subsample and data conditional risks for an arbitrary number M of bags. The end-to-end result is presented in the form of Theorem 5. This theorem

[3] According to (2), $I_1^{(n)}$ and $I_2^{(n)}$ are drawn using SRSWR if $\widetilde{f}_M \in \{\widetilde{f}^{\mathtt{WR}}_{M,\mathcal{I}_k}, \widetilde{f}^{\mathtt{WR}}_{M,\mathcal{I}_k^\pi}\}$ and SRSWOR if $\widetilde{f}_M \in \{\widetilde{f}^{\mathtt{WOR}}_{M,\mathcal{I}_k}, \widetilde{f}^{\mathtt{WOR}}_{M,\mathcal{I}_k^\pi}\}$. From now on, for notational simplicity, we drop the dependence on n and simply write I_1 and I_2.

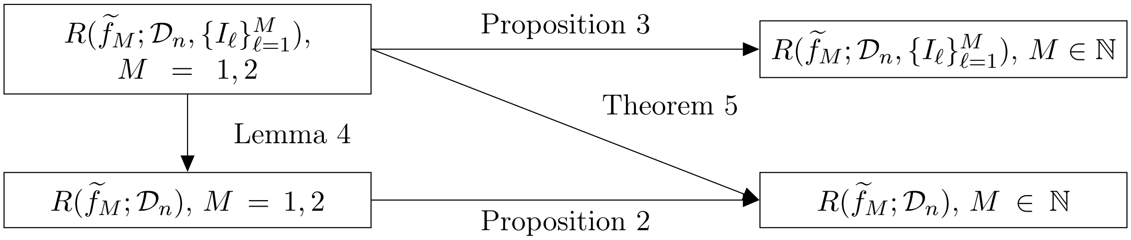

Figure 2: A general reduction strategy for obtaining limiting risks of subagged predictors with M bags.

establishes that it is sufficient to obtain the limiting subsample conditional risks for $M = 1, 2$; see Figure 2.

Theorem 5 (From subsample conditional to data conditional for general M)**.** *Suppose the conditions* (11) *and* (12) *hold, then the conclusions in Propositions 2 and 3 hold.*

For general predictors, both the data conditional risk and the subsample conditional risk for $M = 1$ (required for (11) to hold) are typically available from known results. In such cases, it remains to first derive limiting subsample conditional risk for $M = 2$ (required for (12) to hold) depending on the sampling strategies, and then verify the properties of the limiting conditional risks required in Theorem 5. In this paper, we focus on the asymptotic risk characterization for the bagged ridge and ridgeless predictors and verify the conditions (11) and (12) in the next section.

3. Bagging ridge and ridgeless predictors

In this section, we adopt the reduction strategy proposed in Section 2 to characterize the risk of subagged ridge and ridgeless predictors. The formal definitions of these predictors and data assumptions imposed for our results are given in Section 3.1. Subsequently, the risk characterizations for subagging and splagging are presented in Section 3.2 and Section 3.3, respectively.

3.1 Predictors and assumptions

Consider a dataset $\mathcal{D}_n = \{(\boldsymbol{x}_1, y_1), \dots, (\boldsymbol{x}_n, y_n)\}$ consisting of random vectors in $\mathbb{R}^p \times \mathbb{R}$. Let $\boldsymbol{X} \in \mathbb{R}^{n \times p}$ denote the corresponding feature matrix whose j-th row contains $\boldsymbol{x}_j^\top$, and let $\boldsymbol{y} \in \mathbb{R}^n$ denote the corresponding response vector whose j-th entry contains y_j. For any index set $I \subseteq \{1, 2, \dots, n\}$, let $\mathcal{D}_I = \{(\boldsymbol{x}_j, y_j) : j \in I\}$ be a subsampled dataset, and let $\boldsymbol{L} \in \mathbb{R}^{n \times n}$ denote a diagonal matrix such that $L_{jj} = 1$ if and only if $j \in I$.

Recall that the *ridge* estimator with regularization parameter $\lambda > 0$ fitted on $\mathcal{D}_I$ is defined as

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

The associated ridge predictor is given by $\widehat{f}_\lambda(\boldsymbol{x}; \mathcal{D}_I) = \boldsymbol{x}^\top \widehat{\boldsymbol{\beta}}_\lambda(\mathcal{D}_I)$. The *ridgeless* estimator is the limiting estimator $\widehat{\boldsymbol{\beta}}_\lambda(\mathcal{D}_I)$ as $\lambda \to 0^+$. When $|\mathcal{D}_I| \geq p$, and assuming that the p feature vectors are linearly independent in $\mathbb{R}^p$, it is simply the least squares estimator:

$$\widehat{\boldsymbol{\beta}}_0(\mathcal{D}_I) = (\boldsymbol{X}^\top \boldsymbol{L}\boldsymbol{X}/|\mathcal{D}_I|)^{-1}(\boldsymbol{X}^\top \boldsymbol{L}\boldsymbol{Y}/|\mathcal{D}_I|).$$

When $|\mathcal{D}_I| < p$, it is the minimum ℓ_2-norm least squares estimator:

$$\begin{aligned}\widehat{\boldsymbol{\beta}}_0(\mathcal{D}_I) &= \operatorname*{argmin}_{\boldsymbol{\beta}' \in \mathbb{R}^p} \left\{ \|\boldsymbol{\beta}'\|_2 \;\middle|\; \boldsymbol{\beta}' \in \operatorname*{argmin}_{\boldsymbol{\beta} \in \mathbb{R}^p} \sum_{j \in I} (y_j - \boldsymbol{x}_j^\top \boldsymbol{\beta})^2 \right\} \\ &= (\boldsymbol{X}^\top \boldsymbol{L}\boldsymbol{X}/|\mathcal{D}_I|)^{+}(\boldsymbol{X}^\top \boldsymbol{L}\boldsymbol{y}/|\mathcal{D}_I|).\end{aligned}$$

Here $\boldsymbol{A}^+$ denotes the Moore-Penrose inverse of matrix $\boldsymbol{A}$. Assuming that $\mathcal{D}_I$ has $|\mathcal{D}_I|$ linearly independent observation vectors in $\mathbb{R}^p$, this estimator also interpolates the data, i.e., we have $y_j = \boldsymbol{x}_j^\top \widehat{\boldsymbol{\beta}}_0(\mathcal{D}_I)$ for $j \in I$, and has the minimum ℓ_2-norm among all interpolators. The associated ridgeless predictor is again given by $\widehat{f}_0(\boldsymbol{x}; \mathcal{D}_n) = \boldsymbol{x}^\top \widehat{\boldsymbol{\beta}}_0(\mathcal{D}_n)$.

Given their relevance to the subagged predictors studied in the literature, we will primarily focus on only two of the four subagged predictors as defined in (2), although the implications for the other two can be trivially obtained. For $\lambda \geq 0$, the subagged and splagged predictors respectively are defined as

$$\begin{aligned}\widetilde{f}^{\,\mathtt{WR}}_{M,\mathcal{I}_k}(\boldsymbol{x}; \mathcal{D}_n) &= \boldsymbol{x}^\top \widetilde{\boldsymbol{\beta}}_{\lambda,M}(\{\mathcal{D}_{I_\ell}\}_{\ell=1}^M), \qquad I_1, \dots, I_M \overset{\mathtt{SRSWR}}{\sim} \mathcal{I}_k, \\ \widetilde{f}^{\,\mathtt{WOR}}_{M,\mathcal{I}_k^\pi}(\boldsymbol{x}; \mathcal{D}_n) &= \boldsymbol{x}^\top \widetilde{\boldsymbol{\beta}}_{\lambda,M}(\{\mathcal{D}_{I_\ell}\}_{\ell=1}^M), \qquad I_1, \dots, I_M \overset{\mathtt{SRSWOR}}{\sim} \mathcal{I}_k^\pi,\end{aligned} \tag{14}$$

where $\widetilde{\boldsymbol{\beta}}_{\lambda,M}(\{\mathcal{D}_{I_\ell}\}_{\ell=1}^M) = M^{-1}\sum_{\ell=1}^M \widehat{\boldsymbol{\beta}}_\lambda(\mathcal{D}_{I_\ell})$. For $M > |\mathcal{I}_k^\pi|$, the splagged predictor is defined to be the predictor with $M = |\mathcal{I}_k^\pi|$. When $\lambda = 0$, the base predictors become the ridgeless predictors.

We impose the following Assumptions 1-5 on the dataset $\mathcal{D}_n$ to characterize the risk. These assumptions are standard in the study of the ridge regression under proportional asymptotics; see, e.g., Hastie et al. (2022).

Assumption 1. *The feature vectors $\boldsymbol{x}_i \in \mathbb{R}^p$, $i = 1, \dots, n$, multiplicatively decompose as $\boldsymbol{x}_i = \boldsymbol{\Sigma}^{1/2}\boldsymbol{z}_i$, where $\boldsymbol{\Sigma} \in \mathbb{R}^{p \times p}$ is a positive semidefinite matrix and $\boldsymbol{z}_i \in \mathbb{R}^p$ is a random vector containing i.i.d. entries with mean 0, variance 1, and bounded moment of order $4 + \delta$ for some $\delta > 0$.*

Assumption 2. *The response variables $y_i \in \mathbb{R}$, $i = 1, \dots, n$, additively decompose as $y_i = \boldsymbol{x}_i^\top \boldsymbol{\beta}_0 + \epsilon_i$, where $\boldsymbol{\beta}_0 \in \mathbb{R}^p$ is an unknown signal vector and ϵ_i is an unobserved error that is assumed to be independent of $\boldsymbol{x}_i$ with mean 0, variance σ^2, and bounded moment of order $4 + \delta$ for some $\delta > 0$.*

Assumption 3. *The signal vector $\boldsymbol{\beta}_0$ has bounded limiting energy, i.e., $\lim_{p\to\infty} \|\boldsymbol{\beta}_0\|_2^2 = \rho^2 < \infty$.*

Assumption 4. *There exist real numbers $r_{\min}$ and $r_{\max}$ independent of p with $0 < r_{\min} \leq r_{\max} < \infty$ such that $r_{\min}\boldsymbol{I}_p \preceq \boldsymbol{\Sigma} \preceq r_{\max}\boldsymbol{I}_p$.*

Assumption 5. *Let $\mathbf{\Sigma} = \boldsymbol{W}\boldsymbol{R}\boldsymbol{W}^\top$ denote the eigenvalue decomposition of the covariance matrix $\mathbf{\Sigma}$, where $\boldsymbol{R} \in \mathbb{R}^{p \times p}$ is a diagonal matrix containing eigenvalues (in non-increasing order) $r_1 \geq r_2 \geq \cdots \geq r_p \geq 0$, and $\boldsymbol{W} \in \mathbb{R}^{p \times p}$ is an orthonormal matrix containing the associated eigenvectors $\boldsymbol{w}_1, \boldsymbol{w}_2, \ldots, \boldsymbol{w}_p \in \mathbb{R}^p$. Let H_p denote the empirical spectral distribution of $\mathbf{\Sigma}$ (supported on $\mathbb{R}_{>0}$) whose value at any $r \in \mathbb{R}$ is given by*

$$H_p(r) = \frac{1}{p}\sum_{i=1}^{p} \mathbb{1}_{\{r_i \leq r\}}.$$

Let G_p denote a certain distribution (supported on $\mathbb{R}_{>0}$) that encodes the components of the signal vector $\boldsymbol{\beta}_0$ in the eigenbasis of $\mathbf{\Sigma}$ via the distribution of (squared) projection of $\boldsymbol{\beta}_0$ along the eigenvectors $\boldsymbol{w}_j, 1 \leq j \leq p$, whose value at any $r \in \mathbb{R}$ is given by

$$G_p(r) = \frac{1}{\|\boldsymbol{\beta}_0\|_2^2}\sum_{i=1}^{p} (\boldsymbol{\beta}_0^\top \boldsymbol{w}_i)^2 \ \mathbb{1}_{\{r_i \leq r\}}.$$

Assume there exist fixed distributions H and G such that $H_p \xrightarrow{\text{d}} H$ and $G_p \xrightarrow{\text{d}} G$ as $p \to \infty$.

3.2 Subagging with replacement

In this section, we delve into the risk asymptotics and properties for subagging. In Section 3.2.1, we provide exact risk characterization of subagged ridge and ridgeless predictors. The monotonicity properties of the asymptotic bias and variance components of the risk are presented in Section 3.2.2.

3.2.1 Risk characterization

In preparation for our first result on the risk characterization of subagged ridge and ridgeless predictors, let us establish some notations. We will analyze the subagged predictors (with M bags) in the proportional asymptotics regime, in which the original data aspect ratio (p/n) converges to $\phi \in (0, \infty)$ as $n, p \to \infty$, and the subsample data aspect ratio (p/k) converges to ϕ_s as $k, p \to \infty$. Because $k \leq n$, ϕ_s is always no less than ϕ.

A fixed-point equation defines one of the key quantities that recurs throughout our analysis of subagged ridge predictors. Such fixed point equations have appeared in the literature before in the context of risk analysis of regularized estimators under proportional asymptotics regime. For instance, see Dobriban and Wager (2018); Hastie et al. (2022); Mei and Montanari (2022) in the context of ridge regression; and more generally, for other M-estimators, see Thrampoulidis et al. (2015, 2018), Sur et al. (2019), El Karoui (2013, 2018), Miolane and Montanari (2021), among others. For any $\lambda > 0$ and $\theta > 0$, define $v(-\lambda; \theta)$ as the unique nonnegative solution to the fixed-point equation:

$$\frac{1}{v(-\lambda;\theta)} = \lambda + \theta \int \frac{r}{1 + v(-\lambda;\theta) r} \, \mathrm{d}H(r), \tag{15}$$

and for $\lambda = 0$, $\theta > 1$, we define:

$$v(0;\theta) = \begin{cases} \lim\limits_{\lambda \to 0^+} v(-\lambda;\theta), & \text{if } \theta > 1 \\ +\infty, & \text{if } \theta \in (0, 1]. \end{cases} \tag{16}$$

The fact that the fixed-point equation (15) has a unique nonnegative solution is well known in the random matrix theory literature. See, e.g., Bai and Silverstein (2010); Couillet and Debbah (2011). For completeness, we also provide a proof in Section H.3. The existence of the limit of $v(-\lambda;\theta)$ as $\lambda \to 0^+$ is due to the fact that $v(-\lambda;\theta)$ is monotonically decreasing in $\lambda > 0$ (Patil et al., 2022, Lemma S.6.15 (4)). Additionally, we define non-negative constants $\widetilde{v}(-\lambda;\vartheta,\theta)$ and $\widetilde{c}(-\lambda;\theta)$ via the following equations:

$$\widetilde{v}(-\lambda;\vartheta,\theta) = \frac{\vartheta \int r^2(1+v(-\lambda;\theta)r)^{-2}\,\mathrm{d}H(r)}{v(-\lambda;\theta)^{-2} - \vartheta\int r^2(1+v(-\lambda;\theta)r)^{-2}\,\mathrm{d}H(r)}, \quad \widetilde{c}(-\lambda;\theta) = \int \frac{r}{(1+v(-\lambda;\theta)r)^2}\,\mathrm{d}G(r). \tag{17}$$

Theorem 6 (Risk characterization of subagged ridge and ridgeless predictors)**.** *Let $\widetilde{f}^{\mathtt{WR}}_{M,\mathcal{I}_k}$ be the predictor as defined in (14) for $\lambda \geq 0$. Suppose Assumptions 1-5 hold for the dataset $\mathcal{D}_n$. Then, as $k,n,p \to \infty$ such that $p/n \to \phi \in (0,\infty)$ and $p/k \to \phi_s \in [\phi,\infty]$ (and $\phi_s \neq 1$ if $\lambda = 0$), there exist deterministic functions $\mathscr{R}^{\mathtt{sub}}_{\lambda,M}(\phi,\phi_s)$ for $M \in \mathbb{N}$, such that for $I_1,\ldots,I_M \overset{\mathtt{SRSWR}}{\sim} \mathcal{I}_k$,*

$$\begin{aligned} \sup_{M\in\mathbb{N}} |R(\widetilde{f}^{\mathtt{WR}}_{M,\mathcal{I}_k};\mathcal{D}_n,\{I_\ell\}_{\ell=1}^M) - \mathscr{R}^{\mathtt{sub}}_{\lambda,M}(\phi,\phi_s)| &\xrightarrow{\mathrm{p}} 0,\\ \sup_{M\in\mathbb{N}} |R(\widetilde{f}^{\mathtt{WR}}_{M,\mathcal{I}_k};\mathcal{D}_n) - \mathscr{R}^{\mathtt{sub}}_{\lambda,M}(\phi,\phi_s)| &\xrightarrow{\mathrm{a.s.}} 0. \end{aligned} \tag{18}$$

The guarantee (18) also holds true if $\widetilde{f}^{\mathtt{WR}}_{M,\mathcal{I}_k}$ is replaced by $\widetilde{f}^{\mathtt{WOR}}_{M,\mathcal{I}_k}$. Furthermore, the function $\mathscr{R}^{\mathtt{sub}}_{\lambda,M}(\phi,\phi_s)$ decomposes as

$$\mathscr{R}^{\mathtt{sub}}_{\lambda,M}(\phi,\phi_s) = \sigma^2 + \mathscr{B}^{\mathtt{sub}}_{\lambda,M}(\phi,\phi_s) + \mathscr{V}^{\mathtt{sub}}_{\lambda,M}(\phi,\phi_s), \tag{19}$$

where the bias and variance terms are given by

$$\mathscr{B}^{\mathtt{sub}}_{\lambda,M}(\phi,\phi_s) = M^{-1}B_\lambda(\phi_s,\phi_s) + (1-M^{-1})B_\lambda(\phi,\phi_s), \tag{20}$$

$$\mathscr{V}^{\mathtt{sub}}_{\lambda,M}(\phi,\phi_s) = M^{-1}V_\lambda(\phi_s,\phi_s) + (1-M^{-1})V_\lambda(\phi,\phi_s), \tag{21}$$

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

$$B_\lambda(\vartheta,\theta) = \rho^2(1+\widetilde{v}(-\lambda;\vartheta,\theta))\widetilde{c}(-\lambda;\theta) \text{ and } V_\lambda(\vartheta,\theta) = \sigma^2\widetilde{v}(-\lambda;\vartheta,\theta) \text{ for } \theta\in(0,\infty],\ \vartheta\leq\theta. \tag{22}$$

Theorem 6 provides precise asymptotics for the data conditional as well as the subsample conditional risks of subagged ridge and ridgeless predictors. We have also derived the bias-variance decomposition for the asymptotic risk in (19). Interestingly, the individual bias term is a convex combination of $B_\lambda(\phi_s,\phi_s)$ and $B_\lambda(\phi,\phi_s)$, which correspond to the biases for $M = 1$ and $M = \infty$, respectively. The same conclusion also holds for the variance term. Although the risk behavior for $M = 1$ has been studied by Patil et al. (2022), the risk characterization for general (data-dependent) M is new. As we shall see later in Section 5, the risk behavior for $M = \infty$ is significantly different from that for $M = 1$.

When $\theta > 1$, the parameter $v(0;\theta)$ defined in (16) can also be seen as the unique nonnegative solution to the following fixed-point equation (Patil et al., 2022, Lemma S.6.14):

$$\frac{1}{v(0;\theta)} = \theta\int \frac{r}{1+v(0;\theta)r}\,\mathrm{d}H(r). \tag{23}$$

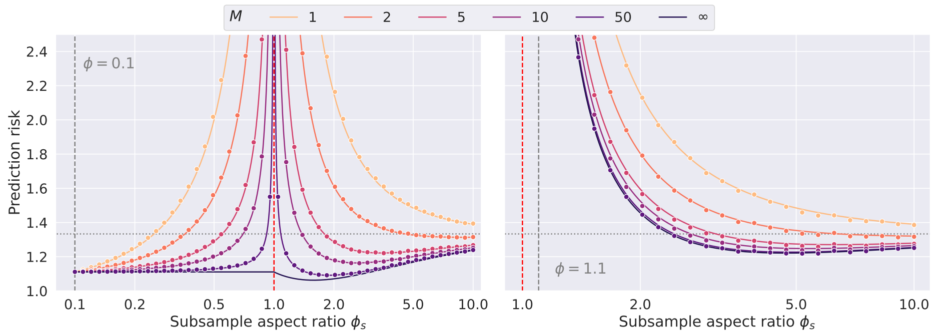

Figure 3: Asymptotic prediction risk curves in (19) for subagged ridgeless predictors ($\lambda = 0$), under model (M-AR1-LI) when $\rho_{\mathrm{ar1}} = 0.25$ and $\sigma^2 = 1$, for varying subsample sizes $k = \lfloor p/\phi_s \rfloor$ and numbers of bags M. The null risk is marked as a dotted line. For each value of M, the points denote finite-sample risks averaged over 100 dataset repetitions, with $n = \lfloor p\phi \rfloor$ and $p = 500$. The left and the right panels correspond to the cases when $p < n$ ($\phi = 0.1$) and $p > n$ ($\phi = 1.1$), respectively.

When $\theta \in (0, 1]$, since $\lim_{\lambda \to 0^+} v(-\lambda; \theta) = \infty$, we have that $\lim_{\lambda \to 0^+} \widetilde{c}(-\lambda; \theta) = 0$ and $\lim_{\lambda \to 0^+} \widetilde{v}(-\lambda; \vartheta, \theta) = \vartheta(1 - \vartheta)^{-1}$. Therefore, the bias and variance functions in (22) for $\vartheta \leq \theta$ reduce to

$$B_0(\vartheta, \theta) = \begin{cases} 0 & \theta \in (0, 1] \\ \rho^2(1 + \widetilde{v}(0; \vartheta, \theta))\widetilde{c}(0; \theta) & \theta \in (1, \infty] \end{cases}, \qquad V_0(\vartheta, \theta) = \begin{cases} \sigma^2 \dfrac{\vartheta}{1 - \vartheta} & \theta \in (0, 1) \\ \infty & \theta = 1 \\ \sigma^2 \widetilde{v}(0; \vartheta, \theta) & \theta \in (1, \infty]. \end{cases} \tag{24}$$

As a sanity check when $\vartheta = \theta$, it is easy to see that the bias and variance components collapse to that of the minimum ℓ_2-norm least squares estimator with limiting aspect ratio θ.

A few additional remarks on Theorem 6 are in order. Note that Theorem 6 shows that the data conditional risk and the subsample conditional risk both converge to the same deterministic limit. This is intuitively expected because the data conditional risk is the average subsample conditional risks over all subsamples. Lastly, Theorem 6 assumes $\lambda \geq 0$. For $\lambda < 0$, the fixed-point equation (15) may have more than one solution. However, a solution to (15) still exists with which Theorem 6 holds whenever $\lambda > -(1 - \sqrt{\phi})^2 r_{\min}$ where $r_{\min}$ is the uniform lower bound on the smallest eigenvalue of $\mathbf{\Sigma}$. For simplicity, we restrict to the case when $\lambda \geq 0$ in this paper. When $\lambda = 0$, the base predictors are ridgeless predictors. In this case, the variance function $\theta \mapsto \mathscr{V}_{0,M}(\vartheta, \theta)$ is unbounded if M is finite and $\theta \to 1$ because $V_0(\theta, \theta)$ in (24) diverges as $\theta \to 1$. This can be empirically explained by the singularity of the empirical covariance matrices with aspect ratios close to 1. However, the asymptotic risk for $M = \infty$ is always bounded.

Illustration of Theorem 6. Before we delve into the proof outline for Theorem 6, we first provide some numerical illustrations under the AR(1) data model. The covariance matrix of an auto-regressive process of order 1 (AR(1)) is denoted by $\mathbf{\Sigma}_{\mathrm{ar1}}$, where $(\mathbf{\Sigma}_{\mathrm{ar1}})_{ij} = \rho_{\mathrm{ar1}}^{|i-j|}$

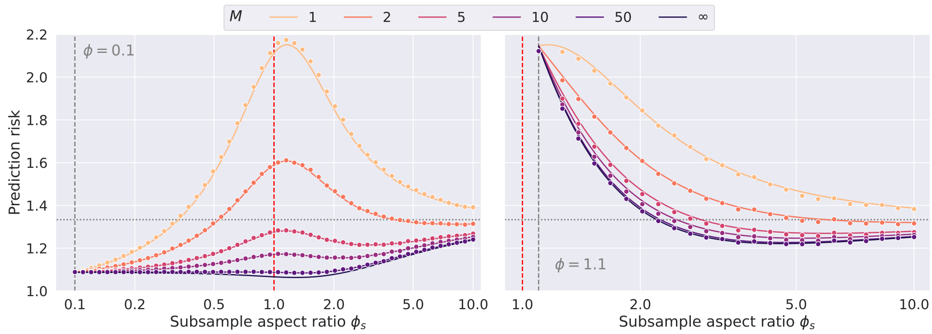

Figure 4: Asymptotic prediction risk curves in (19) for subagged ridge predictors ($\lambda = 0.1$), under model (M-AR1-LI) when $\rho_{\text{ar1}} = 0.25$ and $\sigma^2 = 1$, for varying subsample sizes $k = \lfloor p/\phi_s \rfloor$ and numbers of bags M. The null risk is marked as a dotted line. For each value of M, the points denote finite-sample risks averaged over 100 dataset repetitions, with $n = \lfloor p\phi \rfloor$ and $p = 500$. The left and the right panels correspond to the cases when $p < n$ ($\phi = 0.1$) and $p > n$ ($\phi = 1.1$), respectively.

for some parameter $\rho_{\text{ar1}} \in (0, 1)$. The AR(1) data model is defined as follows:

$$y_i = \boldsymbol{x}_i^\top \boldsymbol{\beta}_0 + \epsilon_i, \quad \boldsymbol{x}_i \sim \mathcal{N}(0, \boldsymbol{\Sigma}_{\text{ar1}}), \quad \boldsymbol{\beta}_0 = \frac{1}{5}\sum_{j=1}^{5} \boldsymbol{w}_{(j)}, \quad \epsilon_i \sim \mathcal{N}(0, \sigma^2), \qquad \text{(M-AR1-LI)}$$

where $\boldsymbol{w}_{(j)}$ is the eigenvector of $\boldsymbol{\Sigma}_{\text{ar1}}$ associated with the top jth eigenvalue $r_{(j)}$. From Grenander and Szegö (1958, pp. 69-70), the top j-th eigenvalue can be written as $r_{(j)} = (1-\rho_{\text{ar1}}^2)/(1-2\rho_{\text{ar1}}\cos\theta_{jp}+\rho_{\text{ar1}}^2)$ for some $\theta_{jp} \in ((j-1)\pi/(p+1), j\pi/(p+1))$. Under model (M-AR1-LI), the signal strength ρ^2 defined in Assumption 3 is $5^{-1}(1-\rho_{\text{ar1}}^2)/(1-\rho_{\text{ar1}})^2$, which is the limit of $25^{-1}\sum_{j=1}^{5} r_{(j)}$. The (M-AR1-LI) model is thus parameterized by two parameters ρ_{ar1} and σ^2 satisfies Assumption 1-5.

Figures 3 and 4 display the limiting risk for the subagged ridgeless predictor and subagged ridge predictor, respectively, with the number of bags M varying from 1 to ∞. In the figures, the limiting aspect ratio ϕ of the full data is fixed to be either 0.1 or 1.1, corresponding to the cases when $n > p$ and $n < p$, respectively. For each case, the limiting aspect ratio ϕ_s of each bag takes values in (ϕ, ∞). We observe that the empirical risks align with the deterministic approximations for both cases, and they are more concentrated around the deterministic approximations as M increases. This is expected as the variance of the subagged predictors reduces with M. Furthermore, for any fixed ϕ_s, the asymptotic risk decreases as M increases.

Due to the non-monotonic risk behavior of the underlying ridge and ridgeless predictors, Figures 3 and 4 show that the best subsample aspect ratio (ϕ_s) in terms of prediction risk might be strictly larger than ϕ. This holds true for any choice of $M \geq 1$. The case of $M = 1$ was already mentioned in Patil et al. (2022). This observation is intriguing as it suggests it is better to bag predictors that use even fewer observations than the original data. Similar phenomena are also observed in our simulations with varying signal-to-noise ratios; see Section J. We discuss an actionable algorithm for finding the optimal choice of ϕ_s in practice in Section 4.

Proof outline of Theorem 6. The proof of Theorem 6 employs the reduction strategy discussed in Section 3. In particular, we apply Theorem 5 (subsample conditional for $M = 1$

and $M=2$ to subsample and data subsample for any M) to prove the theorem. Below we outline the main steps:

1. The deterministic risk approximation to the subsample conditional risk for $M=1$ can be obtained from the results of Patil et al. (2022) that build on those of Hastie et al. (2022).
2. Under the linear model, to analyze the subsample conditional risk for $M=2$, we first decompose it as follows:

$$
\begin{aligned}
&R(\widetilde{f}_2;\mathcal{D}_n,\{I_1,I_2\})\\
&=\sigma^2+\frac{1}{4}\sum_{i=1}^{2}(\boldsymbol{\beta}_0-\widehat{\boldsymbol{\beta}}(\mathcal{D}_{I_i}))^\top\boldsymbol{\Sigma}(\boldsymbol{\beta}_0-\widehat{\boldsymbol{\beta}}(\mathcal{D}_{I_i}))+\frac{1}{2}(\boldsymbol{\beta}_0-\widehat{\boldsymbol{\beta}}(\mathcal{D}_{I_1}))^\top\boldsymbol{\Sigma}(\boldsymbol{\beta}_0-\widehat{\boldsymbol{\beta}}(\mathcal{D}_{I_2}))\\
&=\frac{\sigma^2}{2}+\frac{R(\widetilde{f}_1;\mathcal{D}_n,I_1)+R(\widetilde{f}_1;\mathcal{D}_n,I_2)}{4}+\frac{(\boldsymbol{\beta}_0-\widehat{\boldsymbol{\beta}}(\mathcal{D}_{I_1}))^\top\Sigma(\boldsymbol{\beta}_0-\widehat{\boldsymbol{\beta}}(\mathcal{D}_{I_2}))}{2}.
\end{aligned}
\tag{25}
$$

The first term in the display above is non-random. The asymptotic risk approximation for the second term follows from the asymptotics of the subsample conditional risk for $M=1$. The challenging part is the analysis of the final cross term $(\boldsymbol{\beta}_0-\widehat{\boldsymbol{\beta}}(\mathcal{D}_{I_1}))^\top\Sigma(\boldsymbol{\beta}_0-\widehat{\boldsymbol{\beta}}(\mathcal{D}_{I_2}))$, due to the non-trivial dependence implied by the overlap between $\mathcal{D}_{I_1}$ and $\mathcal{D}_{I_2}$. Our strategy to obtain a deterministic approximation for such a term is to write $h(\widehat{\boldsymbol{\beta}}(\mathcal{D}_{I_1}),\widehat{\boldsymbol{\beta}}(\mathcal{D}_{I_2}))=h(\widehat{\boldsymbol{\beta}}(\mathcal{D}_{I_1'}\cup\mathcal{D}_{I_0}),\widehat{\boldsymbol{\beta}}(\mathcal{D}_{I_2'}\cup\mathcal{D}_{I_0}))$ for any univariate function h. Here $I_0=I_1\cap I_2$ denotes the indices of the overlap, and $I_j'=I_j\setminus I_0$ for $j=1,2$ are the indices of non-overlapping observations. Observe that conditioning on $\mathcal{D}_{I_0}$, $\mathcal{D}_{I_1'}$ and $\mathcal{D}_{I_2'}$ are independent datasets. This conditional independence, coupled with the closed-form expression of the ridge predictor, forms a crucial piece in our argument. To carry out this program, we derive conditional deterministic equivalence results for ridge resolvents. The resulting new results here are collected in Section H.3.2.
3. To prove the results for the ridgeless predictor, we essentially take the limit as $\lambda\to0^+$ of the deterministic risk approximation for the ridge predictor with regularization λ. This process requires appealing to a uniformity argument in λ. See Section D for more details.

3.2.2 Monotonicity of bias and variance in number of bags

Monotonicity in the number of bags M for both the data conditional risk $R(\widetilde{f}^{\texttt{WR}}_{M,\mathcal{I}_k};\mathcal{D}_n)$ and the subsample conditional risk $R(\widetilde{f}^{\texttt{WR}}_{M,\mathcal{I}_k};\mathcal{D}_n,\{I_\ell\}_{\ell=1}^M)$ follow from (7). In the classical literature of bagging and subagging, however, it has been of interest to better understand the effect of aggregation on not just the risk, but also on the bias and variance. In this section, we show for the ridge and ridgeless predictors, subagging reduces both the bias and the variance. Monotonicity of the risk proved in Theorem 6, does not imply the monotonicity of asymptotic bias and variance components. Fortunately, the risk decomposition derived in Theorem 6 demonstrates that both asymptotic bias and variance components are monotonic in M, as summarized below.

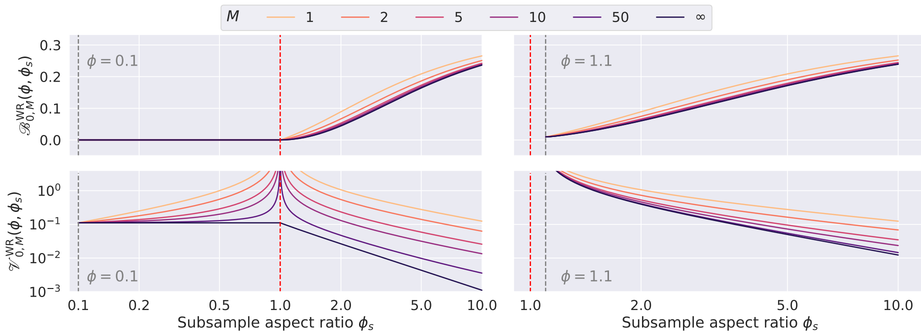

Figure 5: Asymptotic bias and variance curves in (22) for subagged ridgeless predictors ($\lambda = 0$), under model (M-AR1-LI) when $\rho_{\text{ar1}} = 0.25$ and $\sigma^2 = 0.25$, for varying subsample aspect ratio ϕ_s and numbers of bags M. The left and the right panels correspond to the cases when $p < n$ ($\phi = 0.1$) and $p > n$ ($\phi = 1.1$), respectively. The values of $\mathscr{V}^{\text{sub}}_{0,M}(\phi, \phi_s)$ are shown on a log-10 scale.

Proposition 7 (Improvement due to subagging)**.** *For all $M = 1, 2, \ldots$ and $\lambda \in [0, \infty)$, it holds that*

$$\mathscr{B}^{\text{sub}}_{\lambda,\infty}(\phi, \phi_s) \le \mathscr{B}^{\text{sub}}_{\lambda,M+1}(\phi, \phi_s) \le \mathscr{B}^{\text{sub}}_{\lambda,M}(\phi, \phi_s) \tag{26}$$

$$\mathscr{V}^{\text{sub}}_{\lambda,\infty}(\phi, \phi_s) \le \mathscr{V}^{\text{sub}}_{\lambda,M+1}(\phi, \phi_s) \le \mathscr{V}^{\text{sub}}_{\lambda,M}(\phi, \phi_s). \tag{27}$$

The inequalities in (26) are strict whenever $\rho^2 > 0$ and $\phi_s \in (\phi, \infty)$ (and $\phi_s \neq 1$ when $\lambda = 0$), while the inequalities in (27) are strict when $\sigma^2 > 0$ and $\phi_s \in (\phi, \infty)$ (and $\phi_s \neq 1$ when $\lambda = 0$). Thus, the asymptotic risk is monotonically decreasing in M, i.e., $\mathscr{R}^{\text{sub}}_{\lambda,\infty}(\phi, \phi_s) \le \mathscr{R}^{\text{sub}}_{\lambda,M+1}(\phi, \phi_s) \le \mathscr{R}^{\text{sub}}_{\lambda,M}(\phi, \phi_s)$.

The monotonicity property in Proposition 7 does not immediately follow from the decomposition of $\mathscr{B}^{\text{sub}}_{\lambda,M}(\phi, \phi_s)$ and $\mathscr{V}^{\text{sub}}_{\lambda,M}(\phi, \phi_s)$ in (20) and (21). All that is implied by (20) and (21) is that $\mathscr{B}^{\text{sub}}_{\lambda,M}(\phi, \phi_s)$ and $\mathscr{V}^{\text{sub}}_{\lambda,M}(\phi, \phi_s)$ either monotonically increase or decrease in $M \ge 1$. However, Proposition 7 confirms that they are both decreasing in M. We establish this by demonstrating that $\mathscr{B}^{\text{sub}}_{\lambda,1}(\phi, \phi_s) \ge \mathscr{B}^{\text{sub}}_{\lambda,\infty}(\phi, \phi_s)$ and $\mathscr{V}^{\text{sub}}_{\lambda,1}(\phi, \phi_s) \ge \mathscr{V}^{\text{sub}}_{\lambda,\infty}(\phi, \phi_s)$. Moreover, the proposition explicitly distinguishes the cases of non-increasing and strict decreasing of the bias and variance components.

The monotonicity properties claimed in Proposition 7 are supported by Figure 5, which shows the bias and variance components for subagged ridgeless predictors under the model (M-AR1-LI). For a similar illustration for subagged ridge predictors, see Section J.1.

3.3 Splagging without replacement

In this section, we focus on analyzing the risk asymptotics and properties for splagging. More formally, we consider the risk asymptotics of the splagged predictor obtained by averaging the predictors computed on M non-overlapping subsets of the data, each of size k. This is precisely the splagged predictor $\widetilde{f}^{\text{WOR}}_{M,\mathcal{I}^\pi_k}$. Throughout all the asymptotics below, we consider the permutation π to be fixed. Because the limiting risk below does not depend on the permutation π, the conclusions continue to hold true even when the data or subsample conditional risk is averaged over all permutations π. However, it should be emphasized that this is not the same as the data conditional risk of the splagged predictor averaged over all permutations π. In Section 3.3.1, we provide exact risk characterization of splagging without replacement for both ridge and ridgeless predictors. The monotonicity properties of asymptotic bias and variance are then established in Section 3.3.2.

3.3.1 Risk characterization

Recall our convention is defining the splagged predictor $\widetilde{f}^{\mathtt{WOR}}_{M,\mathcal{I}_k^\pi}$ as $\widetilde{f}^{\mathtt{WOR}}_{\min\{M,\lfloor n/k\rfloor\},\mathcal{I}_k^\pi}$, so that the splagged predictor is well defined for all $M \in \mathbb{N}$.

Theorem 8 (Risk characterization for splagged ridge and ridgeless predictors). *Let $\widetilde{f}^{\mathtt{WOR}}_{M,\mathcal{I}_k^\pi}$ be the predictor as defined in (14) for $\lambda \geq 0$. Suppose Assumptions 1-5 hold for the dataset $\mathcal{D}_n$. Then as $k, n, p \to \infty$, $p/n \to \phi \in (0, \infty)$, $p/k \to \phi_s \in [\phi, \infty]$ (and $\phi_s \neq 1$ for $\lambda = 0$), there exist deterministic functions $\mathscr{R}^{\mathtt{spl}}_{\lambda,M}(\phi, \phi_s)$ for all $M \in \mathbb{N}$, and $\phi_s \geq \phi$, such that for $I_1, \ldots, I_M \overset{\mathtt{SRSWOR}}{\sim} \mathcal{I}_k^\pi$,*

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

$$\sup_{M\in\mathbb{N}} |R(\widetilde{f}^{\mathtt{WOR}}_{M,\mathcal{I}_k^\pi}; \mathcal{D}_n) - \mathscr{R}^{\mathtt{spl}}_{\lambda,M}(\phi, \phi_s)| \xrightarrow{\text{a.s.}} 0.$$

Here $\mathscr{R}^{\mathtt{spl}}_{\lambda,M}(\phi, \phi_s) = \mathscr{R}^{\mathtt{spl}}_{\lambda,\lfloor\phi_s/\phi\rfloor}(\phi, \phi_s)$ for $M \geq \lfloor\phi_s/\phi\rfloor$, and for $M \leq \lfloor\phi_s/\phi\rfloor$, the function $\mathscr{R}^{\mathtt{spl}}_{\lambda,M}(\phi, \phi_s)$ decomposes as

$$\mathscr{R}^{\mathtt{spl}}_{\lambda,M}(\phi, \phi_s) = \sigma^2 + \mathscr{B}^{\mathtt{spl}}_{\lambda,M}(\phi, \phi_s) + \mathscr{V}^{\mathtt{spl}}_{\lambda,M}(\phi, \phi_s), \tag{28}$$

where $\mathscr{B}^{\mathtt{spl}}_{\lambda,M}(\phi, \phi_s) = M^{-1}B_\lambda(\phi_s, \phi_s) + (1 - M^{-1})C_\lambda(\phi_s)$, $\mathscr{V}^{\mathtt{spl}}_{\lambda,M}(\phi, \phi_s) = M^{-1}V_\lambda(\phi_s, \phi_s)$, $C_\lambda(\phi_s) = \rho^2\widetilde{c}(-\lambda; \phi_s)$, and $B_\lambda(\phi_s, \phi_s)$ and $V_\lambda(\phi_s, \phi_s)$ are quantities as defined in Theorem 6.

We next provide some remarks on Theorem 8. Firstly, for every pair (ϕ, ϕ_s) satisfying $\phi_s \geq \phi$, note that the splagged predictor and the risks are defined in a non-trivial manner only for $M = 1, \ldots, \lfloor\phi_s/\phi\rfloor$, and is defined as a constant for $M > \lfloor\phi_s/\phi\rfloor$. In particular, for a fixed pair (ϕ, ϕ_s), the sequence of risks as M varies looks like:

$$\mathscr{R}^{\mathtt{spl}}_{\lambda,1}(\phi, \phi_s),\ \mathscr{R}^{\mathtt{spl}}_{\lambda,2}(\phi, \phi_s),\ \ldots,\ \mathscr{R}^{\mathtt{spl}}_{\lambda,\lfloor\phi_s/\phi\rfloor}(\phi, \phi_s),\ \mathscr{R}^{\mathtt{spl}}_{\lambda,\lfloor\phi_s/\phi\rfloor}(\phi, \phi_s),\ \ldots.$$

Secondly, although splagging does not formally involve repeated observations like bootstrapping, we will still refer to $\phi_s = p/k$ as the subsample aspect ratio, where k is the number of observations in each split part of the full dataset. In Theorem 6 for the subagged predictor with replacement, the asymptotic risk depends on both the data aspect ratio ϕ as well as the subsample aspect ratio ϕ_s. In contrast, the asymptotic risk for the splagged predictor without replacement in Theorem 8 does not depend on the data aspect ratio ϕ. This can be seen from the expressions for $\mathscr{B}^{\mathtt{spl}}_{\lambda,M}(\phi, \phi_s)$ and $\mathscr{V}^{\mathtt{spl}}_{\lambda,M}(\phi, \phi_s)$. However, it is interesting to note that the asymptotic risk for $\widetilde{f}^{\mathtt{WR}}_{M,\mathcal{I}_k^\pi}$ depends on both ϕ and ϕ_s because $\limsup_{k,n\to\infty} |\mathcal{I}_k^\pi|$ is finite, which makes the limiting risk of $\widetilde{f}^{\mathtt{WR}}_{M,\mathcal{I}_k^\pi}$ and $\widetilde{f}^{\mathtt{WOR}}_{M,\mathcal{I}_k^\pi}$ different. Because $K_{N,M}$ defined in (8) is bounded above by 1 and $\limsup_{k,n\to\infty} K_{|\mathcal{I}_k^\pi|,M} < 1$ for any $M > 1$, $\widetilde{f}^{\mathtt{WOR}}_{M,\mathcal{I}_k^\pi}$ is a

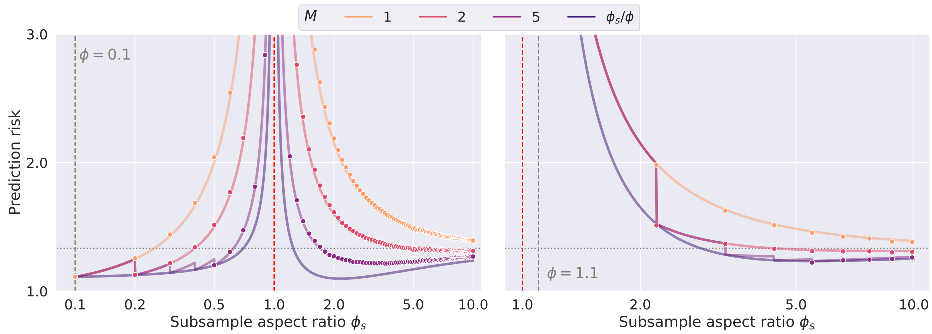

Figure 6: Asymptotic prediction risk curves in (28) for splagged ridgeless predictors ($\lambda = 0$), under model (M-AR1-LI) when $\rho_{\mathrm{ar1}} = 0.25$ and $\sigma^2 = 1$, for varying split sizes $k = \lfloor p/\phi_s \rfloor$ and numbers of bags M. The left and the right panels correspond to the cases when $p < n$ ($\phi = 0.1$) and $p > n$ ($\phi = 1.1$), respectively. The null risk is marked as a dotted line. For each value of M, the points denote finite-sample risks averaged over 100 dataset repetitions, with $p = 500$ and $n = \lfloor p\phi \rfloor$.

strictly better predictor then $\widetilde{f}^{\mathtt{WR}}_{M,\mathcal{I}_k^\pi}$ in terms of the squared risk. In other words, $\widetilde{f}^{\mathtt{WR}}_{M,\mathcal{I}_k^\pi}$ is inadmissible, even asymptotically.

Thirdly, Theorem 6 takes into account the simple average of base predictors fitted on non-overlapped samples. This is also closely related to distributed learning (Mücke et al., 2022) that leverages multiple computing devices to reduce overall training time. While Mücke et al. (2022) provide finite-sample *upper bounds* for the prediction risk of distributed ridgeless predictor, Theorem 6 gives *exact risk* characterization. The distributed ridge predictors are also studied in Dobriban and Sheng (2020). However, their goal is to obtain the optimal weight and the optimal regularization parameter and they only consider estimation risk in the underparameterized regime.

Illustration of Theorem 8. In Figures 6 and 7, we provide numerical illustrations for Theorem 8 (bagged ridgeless and ridge predictors with $\lambda = 0.1$) under the model (M-AR1-LI), with the number of bags M varying from 1 to ∞. The limiting data aspect ratio is fixed at 0.1 when $n > p$ and at 1.1 when $n < p$. We find that the empirical risks align remarkably well with the deterministic approximations, as stated in Theorem 8, for both bagged ridge and ridgeless predictors. Mirroring the findings in Figure 3, for any fixed M, the optimal ϕ_s may be strictly larger than ϕ, an implication of the non-monotonic risk behavior.

Proof outline of Theorem 8. The proof of Theorem 8 follows a similar reduction strategy as in the proof of Theorem 6, where we first analyze the subsample conditional risks for $M = 1$ and $M = 2$, and appeal to Theorem 5 to obtain the result for data conditional and subsample conditional risks for any M. Below we briefly outline the main steps:

1. The deterministic risk approximation to the subsample conditional risk for $M = 1$ splagging is exactly the same as that of subagging.
2. Under the linear model, the subsample conditional risk for $M = 2$ decomposes in a similar manner as (25), except in this case, the datasets $\mathcal{D}_{I_1}$ and $\mathcal{D}_{I_2}$ are independent of each other (conditional on I_1, I_2), which makes the analysis in this case slightly easier compared to the one for subagging. By conditioning on each of the datasets successively

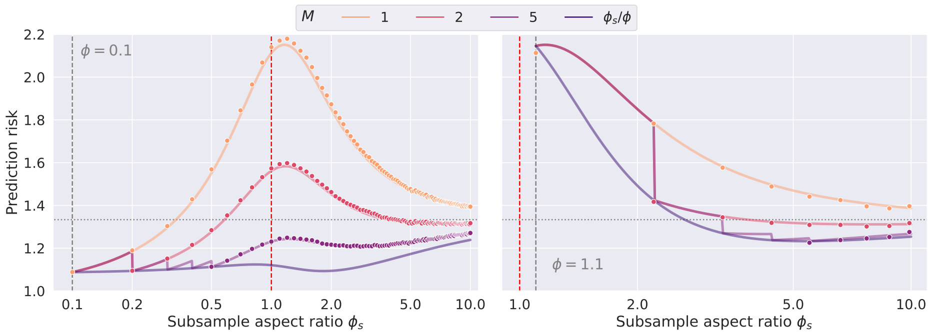

Figure 7: Asymptotic prediction risk curves in (28) for splagged ridge predictors ($\lambda = 0.1$), under model (M-AR1-LI) when $\rho_{\mathrm{ar1}} = 0.25$ and $\sigma^2 = 1$, for varying split sizes $k = \lfloor p/\phi_s \rfloor$ and numbers of bags M. The left and the right panels correspond to the cases when $p < n$ ($\phi = 0.1$) and $p > n$ ($\phi = 1.1$), respectively. The null risk is marked as a dotted line. For each value of M, the points denote finite-sample risks averaged over 100 dataset repetitions, with $p = 500$ and $n = \lfloor p\phi \rfloor$.

and utilizing the closed-form expression of the ridge estimator, we obtain the desired deterministic approximations.

3. Finally, akin to what we did for Theorem 6, we prove results for the ridgeless predictor in the form of the limiting risk approximations to the risk of the ridge predictor in the limit as $\lambda \to 0^+$, based on uniformity arguments.

3.3.2 Monotonicity of bias and variance in number of bags

Just as with subagging, the asymptotic bias and variance components of the conditional risk for splagging are also monotonically decreasing in the number of bags M. This is formalized below.

Proposition 9 (Improvement due to splagging)**.** *Fix any pair (ϕ, ϕ_s) such that $\phi_s \geq \phi$. Then for all $M \in \{1, \dots, \lfloor \phi_s/\phi \rfloor\}$,*

$$\mathscr{B}^{\mathtt{spl}}_{\lambda,\lfloor \phi_s/\phi \rfloor}(\phi, \phi_s) \leq \mathscr{B}^{\mathtt{spl}}_{\lambda,M+1}(\phi, \phi_s) \leq \mathscr{B}^{\mathtt{spl}}_{\lambda,M}(\phi, \phi_s), \tag{29}$$

$$\mathscr{V}^{\mathtt{spl}}_{\lambda,\lfloor \phi_s/\phi \rfloor}(\phi, \phi_s) \leq \mathscr{V}^{\mathtt{spl}}_{\lambda,M+1}(\phi, \phi_s) \leq \mathscr{V}^{\mathtt{spl}}_{\lambda,M}(\phi, \phi_s). \tag{30}$$

The inequalities in (29) *are strict whenever $\rho^2 > 0$ and $\phi_s \in (\phi, \infty)$ (and $\phi_s \neq 1$ when $\lambda = 0$), while the inequalities in* (30) *are strict when $\sigma^2 > 0$ and $\phi_s \in (\phi, \infty)$ (and $\phi_s \neq 1$ when $\lambda = 0$). Thus, the asymptotic risk is monotonically decreasing in M, i.e., $\mathscr{R}^{\mathtt{spl}}_{\lambda,M+1}(\phi, \phi_s) \leq \mathscr{R}^{\mathtt{spl}}_{\lambda,M}(\phi, \phi_s)$.*

As a concluding remark, because the deterministic risk approximation for splagging is defined as a constant in M for $M \geq \lfloor \phi_s/\phi \rfloor$, Proposition 9 implies that the for every fixed pair (ϕ, ϕ_s), the optimal splagged predictor utilizes $M = \lfloor \phi_s/\phi \rfloor$ bags.

4. Risk profile monotonization

The results presented in the previous sections provide risk characterizations for different variants of bagged predictors, per (2), for all possible subsample aspect ratios ϕ_s. In practice, the choice of ϕ_s is crucial for achieving optimal prediction performance. Following the cross-validation strategy discussed in Patil et al. (2022), one can apply cross-validation to choose the optimal ϕ_s in order to obtain the best possible prediction performance by subagging or splagging the base predictor across different subsample sizes. In Section 4.1, we first describe the risk monotonization results for general predictors, going back to the general setting in Section 2. In Section 4.2, we then specialize the general risk monotonization results to the bagged ridge and ridgeless predictors. In Section 4.3, we provide a comparison between the best subagged and the best splagged predictors, considering all possible choices of both ϕ_s and M, when the base predictor is either ridge or ridgeless.

4.1 Bagged general predictors

Several commonly used prediction procedures, such as min-ℓ_2-norm least squares and ridge regression, exhibit a non-monotonic risk behavior as a function of the data aspect ratio ϕ. This is referred to in the literature as double/multiple descents (Belkin et al., 2019; Hastie et al., 2022). The deterministic risk approximation, as a function of the aspect ratio ϕ, first increases, reaches a peak, and then decreases. This can be understood in the context of fixed dimension and changing sample size n as follows: the risk first decreases as the sample size increases up to a certain threshold, after which it starts increasing with a further increase in sample size. This is a counter-intuitive behavior from a conventional statistical viewpoint, as this indicates that more data may hurt performance. However, from a theoretical perspective, additional information should only lead to improved performance. The underlying issue here lies not in the theory but in the sub-optimality of the prediction procedures when applied as-is on the full data.

There are at least two ways in which one can think of improving a given predictor:

1. Obtain a new predictor whose risk is the greatest monotone minorant of the risk of the given prediction procedure. This can be achieved by computing the predictor on a smaller sample size if necessary. Such a procedure is referred to as the zero-step procedure (with $M = 1$) in Patil et al. (2022); see Algorithm 1 for details. The zero-step procedure does the bare minimum to achieve monotone risk.
2. The zero-step procedure (with $M = 1$) is not a genuine improvement of the base predictor, as it simply computes the same predictor on a smaller dataset. Building upon the positive effects of subagging or splagging mentioned in previous sections, we can further improve on the zero-step procedure by aggregating over multiple subsets of the data. This was already hinted at and illustrated in Patil et al. (2022). In this section, we delve deeper into this point.

We note from Theorem 6 and Figures 3 and 4 that for each ϕ, there are essentially infinitely many risk values possible (one for each pair of subsample aspect ratio ϕ_s and the number of bags M). The zero-step procedure (with $M = 1$) improves on the base predictor by optimizing over ϕ_s, while keeping $M = 1$ fixed. Taking a step further, based on our aforementioned results, we can consider optimizing over ϕ_s and $M \geq 1$ (or just over ϕ_s, while fixing $M \geq 1$). In the following, we present an actionable algorithm to achieve the optimum over ϕ_s for any fixed $M \geq 1$. (It is worth noting that we have already established monotonicity over $M \geq 1$, and one can always choose M to be as large as feasible in

Algorithm 1 Cross-validation for subagging or splagging

Input: A dataset $\mathcal{D}_n = \{(\boldsymbol{x}_i, y_i) \in \mathbb{R}^p \times \mathbb{R} : 1 \le i \le n\}$, a positive integer $n_{\mathrm{te}} < n$ (number of test samples), a base prediction procedure $\widehat{f}$, a real number $\nu \in (0,1)$ (bag size unit parameter), a natural number M (number of bags), a centering procedure $\mathtt{CEN} \in \{\mathtt{AVG}, \mathtt{MOM}\}$, a real number η when $\mathtt{CEN} = \mathtt{MOM}$.

1: **Data splitting:** Randomly split $\mathcal{D}_n$ into training set $\mathcal{D}_{\mathrm{tr}}$ and test set $\mathcal{D}_{\mathrm{te}}$ as:

$$\mathcal{D}_{\mathrm{tr}} = \{(\boldsymbol{x}_i, y_i) : i \in \mathcal{S}_{\mathrm{tr}}\}, \quad \text{and} \quad \mathcal{D}_{\mathrm{te}} = \{(\boldsymbol{x}_j, y_j) : j \in \mathcal{S}_{\mathrm{te}}\},$$

where $\mathcal{S}_{\mathrm{te}} \subset [n]$ with $|\mathcal{S}_{\mathrm{te}}| = n_{\mathrm{te}}$, and $\mathcal{S}_{\mathrm{tr}} = [n] \setminus \mathcal{S}_{\mathrm{te}}$.

2: **Bag sample sizes grid construction:** Let $k_0 = \lfloor n^{\nu} \rfloor$ and $\mathcal{K}_n = \{k_0, 2k_0, \dots, \lfloor n/k_0 \rfloor k_0\}$.

3: **Subagging or splagging predictors:** For each $k \in \mathcal{K}_n$, define $\widetilde{f}_{M,k}$ trained on $\mathcal{D}_{\mathrm{tr}}$ as:

- For subagging, let $\widetilde{f}_{M,k}(\cdot) = \widetilde{f}_M(\cdot; \{\mathcal{D}_{I_{k,\ell}}\}_{\ell=1}^M)$ denote the subagged predictor as in (2) with M bags. Here, $I_{k,1}, \dots, I_{k,M}$ represent a simple random sample with or without replacement from the set of all subsets of $\mathcal{S}_{\mathrm{tr}}$ of size k.
- For splagging, $\widetilde{f}_{M,k}(\cdot)$ is the same as above but now $I_{k,1}, \dots, I_{k,M}$ represent a simple random sample without replacement from a random split of $\mathcal{S}_{\mathrm{tr}}$ into $\lfloor n/k \rfloor$ parts with each part containing k elements. As explained in Section 3.1, for $M > \lfloor n/k \rfloor$, no such splitting exists. In this case, we return $\widetilde{f}_{\lfloor n/k \rfloor, k}$. Hence in general, we have $\widetilde{f}_{M,k} = \widetilde{f}_{\min\{M, \lfloor n/k \rfloor\}, k}$.

4: **Risk estimation:** For each $k \in \mathcal{K}_n$, estimate the conditional prediction risk on $\mathcal{D}_{\mathrm{te}}$ of $\widetilde{f}_{M,k}$ as:

$$\widehat{R}(\widetilde{f}_{M,k}) := \begin{cases} |\mathcal{S}_{\mathrm{te}}|^{-1} \sum_{j \in \mathcal{S}_{\mathrm{te}}} (y_j - \widetilde{f}_{M,k}(\boldsymbol{x}_j))^2, & \text{if } \mathtt{CEN}{=}\mathtt{AVG} \quad (31) \\ \mathrm{median}(\widehat{R}_1(\widetilde{f}_{M,k}), \dots, \widehat{R}_B(\widetilde{f}_{M,k})), & \text{if } \mathtt{CEN}{=}\mathtt{MOM}, \quad (32) \end{cases}$$

where $B = \lceil 8 \log(1/\eta) \rceil$, and $\widehat{R}_j(\widetilde{f}_{M,k})$, $1 \le j \le B$ is defined similarly to (31) for B random splits of the test dataset $\mathcal{D}_{\mathrm{te}}$.

5: **Cross-validation**: Set $\widehat{k} \in \mathcal{K}_n$ to be the bagging sample size that minimizes the estimated prediction risk using

$$\widehat{k} \in \operatorname*{argmin}_{k \in \mathcal{K}_n} \widehat{R}(\widetilde{f}_{M,k}). \tag{33}$$

Output: Return the predictor $\widehat{f}_M^{\mathrm{cv}}(\cdot; \mathcal{D}_n) = \widetilde{f}_{M,\widehat{k}}(\cdot) = \widetilde{f}_M(\cdot; \{\mathcal{D}_{I_{\widehat{k},\ell}}\}_{\ell=1}^M)$.

practice.) We then present Theorem 10, in which we prove that the general cross-validation attains the optimum over ϕ_s (asymptotically). Theorem 10 provides theoretical guarantees for the cross-validation procedure for general base predictors, extending the results of Patil et al. (2022) to subagging and splagging.

Theorem 10 (Risk monotonization by cross-validation)**.** *Suppose that as $n, p \to \infty$, $p/n \to \phi \in (0, \infty)$. Let $\mathcal{K}_n$ be the set of subsample sizes defined in Algorithm 1 and $\mathcal{I}_k$ be the set of subsets of $\mathcal{S}_{\rm tr}$ of size $k \in \mathcal{K}_n$ according to the sampling scheme. Suppose that for any $k \in \mathcal{K}_n$, as $n, k, p \to \infty$, and $p/k \to \phi_s \in [\phi, \infty)$, there exists a deterministic function $\mathscr{R} : (0, \infty]^2 \to [0, \infty]$ such that:*

(i) For any $I \in \mathcal{I}_k$ and $\{I_{k,1}, I_{k,2}\}$ a simple random sample from $\mathcal{I}_k$,

$$R(\widetilde{f}_1; \mathcal{D}_n, \{I\}) \xrightarrow{\text{a.s.}} \mathscr{R}(\phi_s, \phi_s), \quad \text{and} \quad R(\widetilde{f}_2; \mathcal{D}_n, \{I_{k,1}, I_{k,2}\}) \xrightarrow{\text{a.s.}} \mathscr{R}(\phi, \phi_s).$$

(ii) For any $\phi \in (0, \infty)$, $\phi_s \mapsto \mathscr{R}(\phi, \phi_s)$ is proper and lower semi-continuous over $[\phi, \infty]$, and is continuous on the set $\operatorname{argmin}_{\{\psi : \psi \geq \phi\}} \mathscr{R}(\phi, \psi)$.

Let $\widehat{f}_M^{\rm cv}$ be the cross-validated predictor returned by Algorithm 1 with base predictor $\widehat{f}$. If the estimated risk $\widehat{R}(\widetilde{f}_{M,k})$ defined in (31) or (32) is uniformly (in $k \in \mathcal{K}_n$) close to the subsample conditional risk $R(\widetilde{f}_{M,k}; \mathcal{D}_n, \{I_{k,\ell}\}_{\ell=1}^M)$ with probability converging to 1, then the following conclusions hold. For subagging with or without replacement, or splagging without replacement, for all $M \in \mathbb{N}$, we have

$$\left(R(\widehat{f}_M^{\rm cv}; \mathcal{D}_n, \{I_{\widehat{k},\ell}\}_{\ell=1}^M) - \min_{\phi_s \geq \phi} \mathscr{R}_M(\phi, \phi_s) \right)_+ \xrightarrow{\text{p}} 0,$$

where the function $\mathscr{R}_M(\phi, \phi_s)$ is defined as

$$\mathscr{R}_M(\phi, \phi_s) := (2\mathscr{R}(\phi, \phi_s) - \mathscr{R}(\phi_s, \phi_s)) + \frac{2}{M}(\mathscr{R}(\phi_s, \phi_s) - \mathscr{R}(\phi, \phi_s)).$$

Furthermore, if for any $\phi_s \in (0, \infty)$, $\phi \mapsto \mathscr{R}(\phi, \phi_s)$ is non-decreasing over $(0, \phi_s]$, then the function $\phi \mapsto \min_{\phi_s \geq \phi} \mathscr{R}_M(\phi, \phi_s)$ is monotonically increasing for every M.

Some remarks regarding Theorem 10 are worth noting. Firstly, although Theorem 10 presents a unified framework for subagging and splagging, the actual limiting risks can be (and in most cases are) different. This discrepancy arises due to the distinct expressions for assumed asymptotic risks in assumption (i) of Theorem 10.

Secondly, Theorem 10 does not exactly characterize the risk of cross-validated bagged predictor; it only states that the subsample conditional risk of $\widetilde{f}_M^{\rm cv}$ is asymptotically no larger than $\min_{\phi_s} \mathscr{R}_M(\phi, \phi_s)$. However, this is an important improvement over the results of Patil et al. (2022), who proved that the subsample conditional risk of $\widetilde{f}_M^{\rm cv}$ is asymptotically no larger than $\min_{\phi_s} \mathscr{R}_1(\phi, \phi_s)$. To precisely characterize the risk of $\widetilde{f}_M^{\rm cv}$, one can make stronger assumptions that as $n, p \to \infty$ and $p/n \to \phi$,

$$\sup_{k \leq n} |R(\widetilde{f}_1; \mathcal{D}_n, \{I_1 \overset{\texttt{SRSWR}}{\sim} \mathcal{I}_k\}) - \mathscr{R}(p/k, p/k)| \xrightarrow{\text{p}} 0, \;\; \sup_{k \leq n} |R(\widetilde{f}_2; \mathcal{D}_n, \{I_1, I_2 \overset{\texttt{SRSWR}}{\sim} \mathcal{I}_k\}) - \mathscr{R}(\phi, p/k)| \xrightarrow{\text{p}} 0,$$

which can be used to conclude

$$R(\widehat{f}_M^{\rm cv}; \mathcal{D}_n, \{I_{\widehat{k},\ell}\}_{\ell=1}^M) \xrightarrow{\text{p}} \min_{\phi_s \geq \phi} \mathscr{R}_M(\phi, \phi_s).$$

The result for bagging without replacement can be extended analogously.

Thirdly, the assumption of uniform (in $k \in \mathcal{K}_n$) closeness of the estimated risk $\widehat{R}(\widetilde{f}_{M,k})$ to the subsample conditional risk $R(\widetilde{f}_{M,k}; \mathcal{D}_n, \{I_{k,\ell}\}_{\ell=1}^M)$ is intended to represent either

$$\max_{k \in \mathcal{K}_n} |\widehat{R}(\widetilde{f}_{M,k}) - R(\widetilde{f}_{M,k}; \mathcal{D}_n, \{I_{k,\ell}\}_{\ell=1}^M)| = o_p(1), \quad \text{or} \quad \max_{k \in \mathcal{K}_n} \left| \frac{\widehat{R}(\widetilde{f}_{M,k})}{R(\widetilde{f}_{M,k}; \mathcal{D}_n, \{I_{k,\ell}\}_{\ell=1}^M)} - 1 \right| = o_p(1).$$

In Section 2 of Patil et al. (2022), the authors provide several assumptions on the data distribution and the predictors that validate this uniform closeness assumption. In Section 4.2, we will apply Theorem 10 for bagged linear predictors, which are themselves linear predictors. In this specific case, Theorem 2.22 in the aforementioned work demonstrates that uniform closeness holds true under assumptions on the data distribution alone (irrespective of the specific linear predictor used, even if they have diverging risks); see Remarks 2.19 and 2.20. We do not delve further into this uniform closeness condition here, but we note that Assumptions 1-5 imply the assumptions of Theorem 2.22 with $\texttt{CEN} = \texttt{MOM}$ (the median-of-means estimator). Additionally, sub-Gaussian features satisfy the assumptions of Theorem 2.22 with $\texttt{CEN} = \texttt{AVG}$.

4.2 Bagged ridge and ridgeless predictors

Theorem 10 provides a general result that describes the risk behavior of cross-validated bagged predictors. Building on our results in previous sections that verify condition (i) of Theorem 10 for both ridge and ridgeless predictors, we now specialize Theorem 10 to these specific predictors under Assumptions 1-5.

Theorem 11 (Risk monotonicity in aspect ratio)**.** *Suppose that the cross-validated predictor $\widehat{f}_M^{\mathrm{cv}}$ is returned by Algorithm 1 with base predictor $\widehat{f}_\lambda$ and M bags, and the conditions in Theorem 6 (or Theorem 8) hold*[4] *with $\mathscr{R}_{\lambda,M}(\phi, \phi_s)$ being the limiting risk $\mathscr{R}_{\lambda,M}^{\texttt{sub}}(\phi, \phi_s)$ (or $\mathscr{R}_{\lambda,M}^{\texttt{spl}}(\phi, \phi_s)$). Then for all $M \in \mathbb{N}$, it holds that*

$$\left(R(\widehat{f}_M^{\mathrm{cv}}; \mathcal{D}_n, \{I_{\widehat{k},\ell}\}_{\ell=1}^M) - \min_{\phi_s \geq \phi} \mathscr{R}_{\lambda,M}(\phi, \phi_s) \right)_+ \xrightarrow{\text{p}} 0. \tag{34}$$

Furthermore, $\phi \mapsto \min_{\phi_s \geq \phi} \mathscr{R}_{\lambda,M}(\phi, \phi_s)$ is a monotonically increasing function of ϕ for every M.

The monotonicity of $\phi \mapsto \min_{\phi_s \geq \phi} \mathscr{R}_{\lambda,M}$ certified by Theorem 11 implies that for every M, for the optimal bagged predictor, more data (i.e., increasing n) cannot hurt. This is illustrated in Figure 8. An observant reader may notice slight non-monotonicity of the empirical risk profile for $M = 1$. This happens because of the small sample size, which restricts the optimal cross-validated predictor from being the null predictor. To prevent this scenario, a default "null" predictor can always be included in general in the set of predictors tuned with cross-validation in Algorithm 1.

For splagging without replacement, the numerical illustrations are displayed in Figure 8(b). As expected, as the limiting aspect ratio ϕ increases, the empirical excess risks

[4] The statement as stated holds for $\texttt{CEN} = \texttt{MOM}$ in Algorithm 1. For $\texttt{CEN} = \texttt{AVG}$, we need to assume sub-Gaussian features as discussed after Theorem 10.

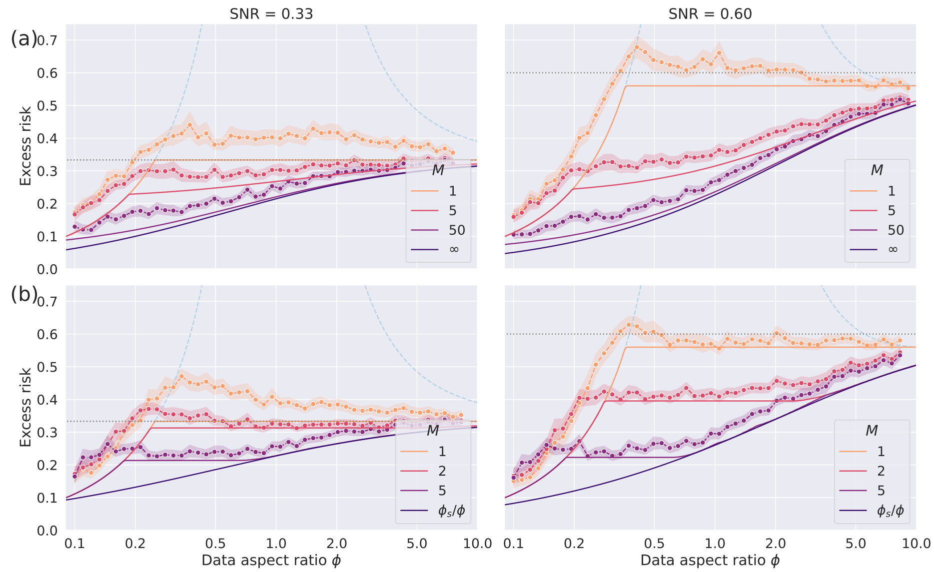

Figure 8: Asymptotic excess risk curves for cross-validated bagged ridgeless predictors ($\lambda = 0$) for (a) subagging and (b) splagging, under model (M-AR1-LI) when $\sigma^2 = 1$ for varying $\mathtt{SNR}$, subsample sizes $k = \lfloor p/\phi_s \rfloor$ and numbers of bags M. The left and the right panels correspond to the cases when $\mathtt{SNR} = 0.33$ ($\rho_{\mathrm{ar1}} = 0.25$) and 0.6 ($\rho_{\mathrm{ar1}} = 0.5$), respectively. The excess null risks and the risks for the ridgeless predictors without bagging are marked as dotted lines and dashed lines, respectively. For each value of M, the points denote finite-sample risks averaged over 100 dataset repetitions, and the shaded regions denote the values within one standard deviation, with $n = 1000$, $n_{\mathrm{te}} = 63$, and $p = \lfloor n\phi \rfloor$.

largely exhibit a monotonic increase and match with theoretical curves. Another observation from Figure 8 (splagging without replacement) is that the asymptotic risk may not be monotonically decreasing in M when ϕ is small. This is because the subsample aspect ratio ϕ_s is restricted by the number of bags M in that it cannot be below $M\phi$, and the differences in the range of ϕ_s when using different numbers of bags result in the non-monotonicity when ϕ is small. While in the overparameterized region, when ϕ is sufficiently large, the cross-validated risk for bagging without replacement is guaranteed to be monotonically decreasing in M. Furthermore, the choice of $M = \phi_s/\phi$ guarantees that the risk is always optimal compared to any other value of M.

4.3 Optimal subagging versus optimal splagging

The previously discussed cross-validated predictors yield asymptotically optimal risks over subsample aspect ratio ϕ_s for each M. Going a step further, we can obtain the optimal subagging or splagging by jointly optimizing over both ϕ_s and M. Leveraging the explicit formulas of the limiting risks for each pair of aspect ratios (ϕ, ϕ_s) and each M, we are able to compare the optimal bagged risks in the two cases.

Proposition 12 (Comparison of the optimal risk of subagging and splagging)**.** *Under Assumptions 3-5, let $\mathscr{R}^{\mathtt{sub}}_{\lambda,M}(\phi, \phi_s)$ and $\mathscr{R}^{\mathtt{spl}}_{\lambda,M}(\phi, \phi_s)$ be defined as in Theorem 6 and Theorem 8,* *respectively. Then for any $\lambda \in [0, \infty)$ and $\phi \in (0, \infty)$, the following holds:*

$$\inf_{M \in \mathbb{N}, \phi_s \in [\phi, \infty]} \mathscr{R}^{\texttt{sub}}_{\lambda, M}(\phi, \phi_s) \leq \inf_{M \in \mathbb{N}, \phi_s \in [\phi, \infty]} \mathscr{R}^{\texttt{spl}}_{\lambda, M}(\phi, \phi_s). \tag{35}$$

In words, optimal subagging is at least as good as optimal splagging (without replacement) in terms of squared loss for ridge predictors.

For any dataset with fixed aspect ratio ϕ, Proposition 12 suggests that subagging *always* achieves the optimal risk for bagged predictor across all possible choices of M and subsample aspect ratio ϕ_s. The optimal subagging and optimal splagging risks, as stated in Proposition 12, can be written as

$$\mathscr{R}^{\texttt{sub}}_{\texttt{opt}}(\phi) = \mathscr{R}^{\texttt{sub}}_{\lambda, \infty}(\phi, \phi_s^{\texttt{sub}}(\phi)), \quad \text{and} \quad \mathscr{R}^{\texttt{spl}}_{\texttt{opt}}(\phi) = \mathscr{R}^{\texttt{spl}}_{\lambda, \phi_s^{\texttt{spl}}(\phi)/\phi}(\phi, \phi_s^{\texttt{spl}}(\phi)). \tag{36}$$

Here the functions $\phi \mapsto \phi_s^{\texttt{sub}}(\phi)$ and $\phi \mapsto \phi_s^{\texttt{spl}}(\phi)$ are defined via

$$\phi_s^{\texttt{sub}}(\phi) := \operatorname*{argmin}_{\phi_s \geq \phi} \mathscr{R}^{\texttt{sub}}_{\lambda, \infty}(\phi, \phi_s), \quad \text{and} \quad \phi_s^{\texttt{spl}}(\phi) := \operatorname*{argmin}_{\phi_s \geq \phi} \mathscr{R}^{\texttt{spl}}_{\lambda, \phi_s/\phi}(\phi, \phi_s). \tag{37}$$

The fact that the optimal risks shown in Proposition 12 are the same as shown in (36) follows from the fact that the risks are monotonically decreasing in M for subagging and that the risk at $M = \phi_s/\phi$ is the best for splagging without replacement for any pair (ϕ, ϕ_s). The quantities $\phi_s^{\texttt{sub}}(\cdot)$ and $\phi_s^{\texttt{spl}}(\cdot)$ represent the best possible subsample aspect ratios for subagging and splagging (without replacement) for every data aspect ratio ϕ given. (Minimizers of lower semi-continuous functions over compact domains exist, which is true for the functions in (37) from Theorem 11.)

The theoretical optimal asymptotic risks (36) for bagged ridgeless predictors are illustrated in Figure 9. The optimal risk $\min_{\phi_s \geq \phi} \mathscr{R}^{\texttt{sub}}_{\lambda, 1}(\phi, \phi_s) = \min_{\phi_s \geq \phi} \mathscr{R}^{\texttt{spl}}_{\lambda, 1}(\phi, \phi_s)$ of the bagged ridgeless predictor with $M = 1$ is also showcased as the dashed line, which matches the monotone risk of the zero-step ridgeless predictor from Patil et al. (2022) with $M = 1$. As demonstrated in Figure 1 and Figure 9(a), the optimal risk for the subagged ridgeless predictor is always smaller than the splagged ridgeless predictor without replacement. Both strategies demonstrate an improvement over the risk of the ridgeless predictor that uses the optimal subsample aspect ratio ϕ_s with only one bag ($M = 1$).

Oracle properties of optimal subsample aspect ratios. From the preceding section, we see that optimal subagged ridge or ridgeless regression always outperforms the splagged one in terms of limiting risk. Owing to the monotonicity in the number of bags M, as established in Proposition 7, the optimal risk for subagging is always achieved at $M = \infty$ for any given subsample aspect ratio ϕ_s. This leads us to the question: what is the optimal subsample aspect ratio ϕ_s? We offer a partial answer to this question in Proposition 13, specialized for ridgeless regression.

Proposition 13 (Optimal risk for bagged ridgeless predictor)**.** *Suppose the conditions in Theorems 6 and 8 hold, and $\sigma^2, \rho^2 \geq 0$ are the noise variance and signal strength from Assumptions 2 and 3. Let $\texttt{SNR} = \rho^2/\sigma^2$. For any $\phi \in (0, \infty)$, the properties of the optimal asymptotic risks $\mathscr{R}^{\texttt{sub}}_{0, \infty}(\phi, \phi_s^{\texttt{sub}}(\phi))$ and $\mathscr{R}^{\texttt{spl}}_{0, \phi_s/\phi}(\phi, \phi_s^{\texttt{spl}}(\phi))$ in terms of $\texttt{SNR}$ and ϕ are characterized as follows:*

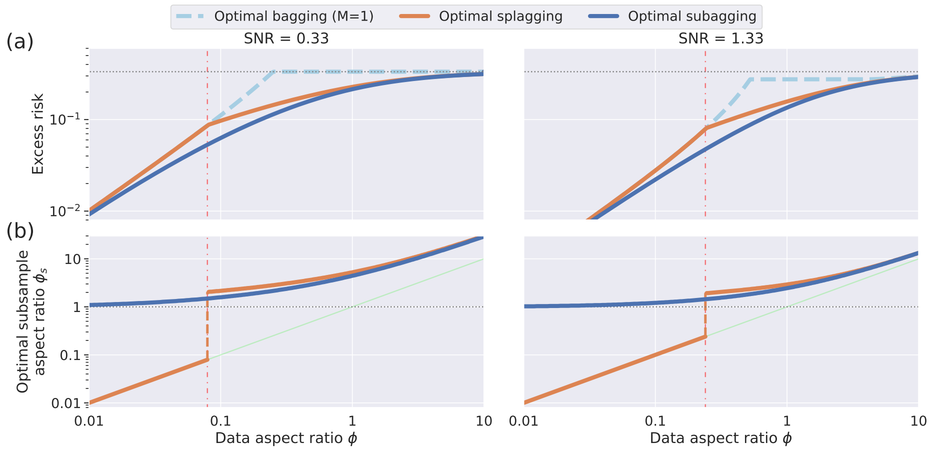

Figure 9: Comparison between optimal subagging and optimal splagging of ridgeless predictors ($\lambda = 0$) for varying limiting aspect ratios ϕ of p/n under model (M-AR1-LI) when $\sigma^2 = 1$. The left and right panels correspond to $\mathtt{SNR} = 0.33$ ($\rho_{\text{ar1}} = 0.25$) and $\mathtt{SNR} = 0.6$ ($\rho_{\text{ar1}} = 0.5$), respectively. The point of phase transition for splagging is marked as the red dash-dot line in every subplot. (a) Optimal asymptotic excess risk curves (35). The excess null risks are marked as gray dotted lines and the blue dashed lines represent the optimal risks of bagged ridgeless predictor with $M = 1$, which are the same as the risks from the zero-step procedure of Patil et al. (2022). (b) The corresponding optimal subsample aspect ratio ϕ_s as a function of data aspect ratio ϕ. For subagging, the optimal subsample aspect ratio is always larger than one (above the gray dotted line). The line $\phi_s = \phi$ is shown in green color.

(1) $\mathtt{SNR} = 0$ ($\rho^2 = 0, \sigma^2 \neq 0$)*: For all* $\phi \geq 0$*, the global minimum* σ^2 *of both* $\mathscr{R}^{\mathtt{sub}}_{0,\infty}(\phi, \phi^{\mathtt{sub}}_s(\phi))$ *and* $\mathscr{R}^{\mathtt{spl}}_{0,\phi_s/\phi}(\phi, \phi^{\mathtt{spl}}_s(\phi))$ *are obtained with* $\phi^{\mathtt{sub}}_s(\phi) = \phi^{\mathtt{spl}}_s(\phi) = \infty$.

(2) $\mathtt{SNR} > 0$*: For all* $\phi \geq 0$*, the global minimum of* $\phi_s \mapsto \mathscr{R}^{\mathtt{sub}}_{0,\infty}(\phi, \phi_s)$ *is obtained at* $\phi^{\mathtt{sub}}_s(\phi) \in (1, \infty)$*. For* $\phi \geq 1$*, the global minimum of* $\phi_s \mapsto \mathscr{R}^{\mathtt{spl}}_{0,\phi_s/\phi}(\phi, \phi_s)$ *is obtained at* $\phi^{\mathtt{spl}}_s(\phi) \in (1, \infty)$*; for* $\phi \in (0, 1)$*, the global minimum of* $\phi_s \mapsto \mathscr{R}^{\mathtt{spl}}_{0,\phi_s/\phi}(\phi, \phi_s)$ *is obtained at* $\phi^{\mathtt{spl}}_s(\phi) \in \{\phi\} \cup (1, \infty)$.

(3) $\mathtt{SNR} = \infty$ ($\rho^2 \neq 0, \sigma^2 = 0$)*: If* $\phi \in (0, 1]$*, the global minimum* $\mathscr{R}^{\mathtt{sub}}_{0,\infty}(\phi, \phi^{\mathtt{sub}}_s(\phi)) = \mathscr{R}^{\mathtt{spl}}_{0,\phi_s/\phi}(\phi, \phi^{\mathtt{spl}}_s(\phi)) = 0$ *is obtained with any* $\phi^{\mathtt{sub}}_s(\phi), \phi^{\mathtt{spl}}_s(\phi) \in [\phi, 1]$*. If* $\phi \in (1, \infty)$*, then the global minimums* $\mathscr{R}^{\mathtt{sub}}_{0,\infty}(\phi, \phi^{\mathtt{sub}}_s(\phi))$ *and* $\mathscr{R}^{\mathtt{spl}}_{0,\phi_s/\phi}(\phi, \phi^{\mathtt{spl}}_s(\phi))$ *are obtained at* $\phi^{\mathtt{sub}}_s(\phi), \phi^{\mathtt{spl}}_s(\phi) \in [\phi, \infty)$.

Proposition 13 reveals that the optimal subsample aspect ratio $\phi^{\mathtt{sub}}_s(\phi)$ for subagging always lies in the range $[1, \infty]$, i.e., within the overparameterized regime. In other words, subagging interpolators with larger aspect ratios (larger than the full data aspect ratio ϕ) can help to reduce the prediction risk, even when $\phi < 1$. For splagging, however, the minimum risk can be obtained either using the full data or splagging interpolators, depending on the data aspect ratio ϕ and the signal-to-noise ratio.

Interestingly, the optimal subsampling aspect ratio for splagging is either ϕ or falls within the overparameterized regime $(1, \infty)$. This implies that either splagging does not help, or when it helps, one has to splag interpolators. Whenever $\mathtt{SNR}$ is positive, the optimal subsample aspect ratio is finite for any ϕ. We can thus visually represent $\phi_s^{\mathtt{sub}}(\phi)$ and $\phi_s^{\mathtt{spl}}(\phi)$ in Figure 9(b). As Figure 9 illustrates, there is a point of non-differentiability of $\phi_s^{\mathtt{spl}}(\phi)$ for optimal splagging without replacement. Prior to this point of non-differentiability, $\phi_s^{\mathtt{spl}}(\phi) = \phi$, which is the same as the optimal bagged ridgeless with $M = 1$. This also coincides with the ridgeless predictor trained on the full data set. Beyond the point of non-differentiability, the optimal risk for splagging without replacement is found in the overparameterized regime, i.e., $\phi_s^{\mathtt{spl}}(\phi) > 1$. Contrasting with splagging, $\phi_s^{\mathtt{sub}}(\phi) \geq 1$ for all $\phi > 0$, implying that subagging interpolators (in the overparameterized regime) is always advantageous.

These findings suggest that when the number of bags is sufficiently large, splagging without replacement proves beneficial only when the limiting aspect ratio ϕ of the full dataset surpasses a certain threshold. However, subagging is always beneficial in reducing the prediction risk, even in the underparameterized regime.

5. Illustrations and insights

The results discussed so far are derived under Assumptions 1-5 that, in particular, allow for features with arbitrary covariance structure $\mathbf{\Sigma}$. We will shift our attention to a simpler case of isotropic features (i.e., $\mathbf{\Sigma} = \boldsymbol{I}_p$ in Assumption 1). In this case, the spectral distribution simplifies, enabling us to compute the fixed point solutions analytically. Our discussion will primarily revolve around the case of ridgeless predictors for the sake of illustration. While it is possible to obtain similar results for ridge predictors, the resulting expressions would be more involved. In Section G.3, we provide formulas for the fixed-point solutions for $\lambda > 0$. From these, one can derive the risk as well as the individual bias and variance numerically for ridge predictors (with arbitrary $\lambda > 0$). Generally speaking, these quantities can always be computed numerically for nonisotropic models.

In the case of isotropic features, the bias and variance functions presented in Theorems 6 and 8 take on relatively simple forms, as demonstrated in Corollary 14. Furthermore, the asymptotic bias and variance can be computed for all $M \in \mathbb{N}$ based on (24).

Corollary 14 (Bias-variance components for isotropic design)**.** *Assume the conditions in Theorem 6 or Theorem 8 hold with $\mathbf{\Sigma} = \boldsymbol{I}_p$. Then we have*

$$B_0(\phi,\phi_s) = \rho^2 \frac{(\phi_s - 1)^2}{\phi_s^2 - \phi} \mathbb{1}_{(1,\infty]}(\phi_s), \qquad C(\phi_s) = \rho^2 \frac{(\phi_s - 1)^2}{\phi_s^2} \mathbb{1}_{(1,\infty]}(\phi_s),$$

$$V_0(\phi,\phi_s) = \begin{cases} \sigma^2 \dfrac{\phi}{1-\phi}, & \phi_s \in (0,1) \\ \infty, & \phi_s = 1 \\ \sigma^2 \dfrac{\phi}{\phi_s^2 - \phi}, & \phi_s \in (1,\infty]. \end{cases}$$

Subagging with replacement. Based on Corollary 14, we are equipped to evaluate the closed-form asymptotic risk under model (M-ISO-LI):

$$y_i = \boldsymbol{x}_i^\top \boldsymbol{\beta}_0 + \epsilon_i, \quad \boldsymbol{x}_i \sim \mathcal{N}(0, \boldsymbol{I}_p), \quad \boldsymbol{\beta}_0 \sim \mathcal{N}(0, p^{-1}\rho^2 \boldsymbol{I}_p), \quad \epsilon_i \sim \mathcal{N}(0, \sigma^2). \qquad \text{(M-ISO-LI)}$$

Additional experimental results under model (M-ISO-LI) can be found in Section J. It is worth noting that while the Gaussianity of the noise ϵ_i in model (M-ISO-LI) simplifies numerical evaluation, it is not a requirement for Corollary 14. It suffices to have the first and second moments match as above. For $M \in \mathbb{N}$, the bias term is always increasing, while the variance term will blow up when the subsample aspect ratio ϕ_s approaches one. However, the variance for $M = \infty$ is different; it is decreasing in ϕ_s and continuous at $\phi_s = 1$. Consequently, one might be interested in the optimal subsample aspect ratio $\phi_s^{\mathtt{sub}}(\phi)$, that best trades off the bias and variance, and minimizes the risk for a given value of ϕ and $M = \infty$.

Proposition 15 (Optimal risk for subagged ridgeless predictors with isotropic features)**.** *Suppose the conditions in Corollary 14 hold, and $\sigma^2, \rho^2 \geq 0$ are the noise variance and signal strength from Assumptions 2 and 3. Let $\mathtt{SNR} = \rho^2/\sigma^2$. For any $\phi \in (0, \infty)$, the properties of the asymptotic risk $\mathscr{R}_{0,\infty}^{\mathtt{sub}}(\phi, \phi_s)$ as a function of ϕ_s are characterized as follows:*

(1) $\mathtt{SNR} = 0$ $(\rho^2 = 0, \sigma^2 \neq 0)$*: The global minimum $\mathscr{R}_{0,\infty}^{\mathtt{sub}}(\phi, \phi_s^{\mathtt{sub}}(\phi)) = \sigma^2$ is obtained at $\phi_s^{\mathtt{sub}}(\phi) = \infty$.*

(2) $\mathtt{SNR} > 0$*: The global minimum*

$$\mathscr{R}_{0,\infty}^{\mathtt{sub}}(\phi, \phi_s^{\mathtt{sub}}(\phi)) = \frac{\sigma^2}{2}\left[1 + \frac{\phi - 1}{\phi}\mathtt{SNR} + \sqrt{\left(1 - \frac{\phi - 1}{\phi}\mathtt{SNR}\right)^2 + 4\mathtt{SNR}}\right] \tag{38}$$

is obtained at $\phi_s^{\mathtt{sub}}(\phi) = A + \sqrt{A^2 - \phi} \in (1, \infty)$ where $A = (\phi + 1 + \phi/\mathtt{SNR})/2$.

(3) $\mathtt{SNR} = \infty$ $(\rho^2 \neq 0, \sigma^2 = 0)$*: If $\phi \in (0, 1]$, then the global minimum is $\mathscr{R}_{0,\infty}^{\mathtt{sub}}(\phi, \phi_s^{\mathtt{sub}}(\phi)) = 0$ is attained at any $\phi_s \in [\phi, 1]$. If $\phi \in (1, \infty)$, then the global minimum $\mathscr{R}_{0,\infty}^{\mathtt{sub}}(\phi, \phi_s^{\mathtt{sub}}(\phi)) = \sigma^2 + \rho^2(\phi - 1)/\phi$ is attained at $\phi_s^{\mathtt{sub}}(\phi) = \phi$.*

As a specific application of Proposition 13, Proposition 15 provides the analytic expression of the optimal risk attainable through optimization over all choices of the number of bags M and the subsample aspect ratio ϕ_s. Additionally, it elucidates the relationship between the optimal risk and the $\mathtt{SNR}$, which is further visualized in Figure 10. Particularly, the optimal subagged risk is monotonically decreasing in $\mathtt{SNR}$ when σ^2 is fixed, which is an intuitive behavior as one would expect a larger $\mathtt{SNR}$ results in a smaller prediction risk. In contrast, such a property is not satisfied by the ridge or ridgeless predictor computed on the full data (Hastie et al., 2022, Figure 2). It can be shown that the gap between the optimal risk, given in Proposition 15, and the underparameterized excess risk $\sigma^2\phi/(1-\phi)$, obtained with the full dataset, gets larger when $\mathtt{SNR}$ gets smaller. Most importantly, it benefits more when the $\mathtt{SNR}$ gets smaller, with a higher overparameterized aspect ratio $\phi_s^{\mathtt{sub}}(\phi)$.

Theorem 16 (Optimal subagged ridgeless risk versus optimal ridge risk)**.** *Under the conditions in Corollary 14, we have that for all $\phi \in (0, \infty)$,*

$$\min_{\phi_s \geq \phi} \mathscr{R}_{0,\infty}^{\mathtt{sub}}(\phi, \phi_s) \;=\; \min_{\lambda \geq 0} \mathscr{R}_{\lambda,1}^{\mathtt{sub}}(\phi, \phi).$$

In words, the optimal limiting risk of the subagged ridgeless predictors equals the optimal ridge predictors trained on the full data.

Theorem 16 reveals a rather surprising connection between subagging and ridge regression. This result implies that subagging a ridge predictor with $\lambda = 0$ and optimizing over

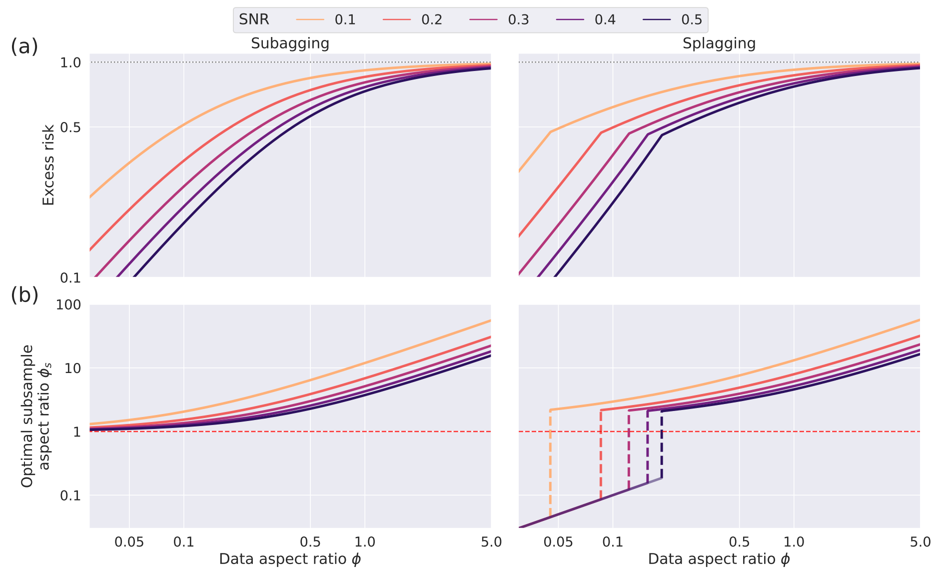

Figure 10: Properties of optimal bagged ridgeless predictors ($\lambda = 0$) under model (M-ISO-LI) when $\rho^2 = 1$, for varying signal noise ratio ($\mathtt{SNR} = \rho^2/\sigma^2$). (a) Optimal asymptotic excess risk curves of subagging (left panel) and splagging (right panel) over the number of bags M and subsample aspect ratio ϕ_s. The optimal numbers of bags are $M = \infty$ and $M = \phi_s/\phi$ for subagging and splagging, respectively. The gray dotted lines represent the excess null risk. (b) The corresponding optimal subsample aspect ratio ϕ_s as a function of data aspect ratio ϕ. For subagging, the optimal subsample aspect ratio is always larger than one (above the red dashed line).

the subsample size is "same" as using the ridge predictor with $\lambda \geq 0$ and optimizing over λ. Consequently, this suggests that subsampling and optimizing over subsample size is a form of regularization. A similar connection between subsampling features and ridge regression was made by LeJeune et al. (2020, Theorem 3.6).

Compared to Theorem 3.6 of LeJeune et al. (2020), our Theorem 16 provides the following three key improvements: (1) **Subsampling scope**. The former theorem focuses solely on the subsampling of features, whereas our theorem considers the sampling of observations. Moreover, in the approach by LeJeune et al. (2020), sampling is restricted to ensure that the final optimal ensemble comprises only least squares estimators. Specifically, they maintain the number of observations in the subsample greater than the number of features, ensuring the existence of a least squares solution for the subsampled data. In contrast, our method permits arbitrary subsample sizes, which means the optimal ensemble can encompass both subsampled least squares and ridgeless interpolators. This distinction is crucial, as there can be scenarios where the optimal subsample might contain more features than observations, a phenomenon highlighted in Proposition 15. (2) **Signal constraints.** The previous theorem limits itself to isotropic random signals $\boldsymbol{\beta}_0$. We broaden this scope to incorporate any arbitrary deterministic signals with bounded norms. (3) **Distributional assumptions.** LeJeune et al. (2020) assumes strong distributional assumptions on the features, noise, and signal, particularly requiring all of them to follow a Gaussian distribution. In comparison, our results do not require such strong distributional assumptions on either the features or the noise and accommodate any deterministic signal with bounded norms.

While Theorem 16 suggests that the two optimal limiting risks coincide under the isotropic model, it is important to note the difference in their risk monotonicity properties in the data aspect ratio ϕ. The optimal risk of the subagged ridgeless predictor is expected to remain monotonically decreasing in ϕ, as shown in Theorem 11. In contrast, it is yet to be ascertained whether the optimal ridge predictor has the same property under general models. In the isotropic case, the fixed point parameter can be explicitly solved in terms of the parameters (λ, ϕ, ϕ_s). The explicit formula enables direct analysis of the monotonicity properties of the asymptotic risk and subsequently facilitates the derivation of the optimal risks. However, in the non-isotropic case, such an explicit formula is not available. This lack of an explicit formula calls for a different strategy to extend Theorem 16 to non-isotropic features.[5]

Splagging without replacement. Unlike subagging, it is possible, though very cumbersome to obtain the optimal sub-sampling ratio $\phi_s^{\mathtt{spl}}(\phi)$ in this case. It involves solving a cubic equation (for a fixed M) or a quartic equation (for the optimal M). Consequently, we resort to numerical computation for $\phi_s^\star$ and provide a qualitative behavior for ϕ_s next. We observe that as SNR increases, the point of phase transition occurs at a larger value of ϕ. This indicates that when there are much more features than samples in the full dataset and the SNR is relatively large, then splagging does not help to reduce the prediction risk. However, when the SNR is small, splagging interpolators is beneficial, even when n is much larger than p in the full data set.

Subagging versus splagging. The comparison between subagging and splagging methods shows interesting findings in terms of prediction risks. Next we briefly summarize these findings concerning the similarities and differences between the two types of bagging strategies for ridgeless predictors. From Figure 10, we observe that for any data aspect ratio ϕ and any SNR, subagging can help to reduce the risk with a suitable subsample aspect ratio in the overparameterized regime, if we have enough bags. In contrast, splagging may not help when $\phi < 1$ and SNR is large, even if we optimize over all possible numbers of bags and subsample aspect ratios jointly. For the cases when subagging or splagging is beneficial, the maximal gain compared to the predictor computed on the full data increases as the SNR decreases. When the full data aspect ratio ϕ is near 1, both subagging and splagging substantially reduce the prediction risk; see Figures 3, 4, 6 and 7. Most surprisingly, even if the original dataset is heavily underparameterized, overparameterized subagging always helps, as shown in Figure 9(b). For example, recall in Figure 3 when $n = 5000$ and $p = 500$ (which is a favorable case in classical statistics), subagged ridgeless predictors trained on overparameterized subsampled datasets (e.g., with $n = 50$ and $p = 500$) with $M = 50$ bags have smaller prediction risk than least squares fitted on the original data.

6. Discussion

In this paper, we provide a generic reduction strategy for characterizing the prediction risk of general bagged predictors (for two bagging strategies of subagging and splagging). As

[5] Subsequent to finishing work, Theorem 16 has now been extended for non-isotropic cases in Du et al. (2023) by establishing connections between the fixed-point equations involved and utilizing their monotonicity properties.

a function of the number of bags M, we show that the asymptotic risk of the M-bagged predictor under squared error loss can be expressed as $M^{-1}\mathfrak{R}_1 + (1 - M^{-1})\mathfrak{R}_\infty$, where $\mathfrak{R}_1$ and $\mathfrak{R}_\infty$ represent the asymptotic squared risks of the M-bagged predictor with $M = 1$ and $M = \infty$, respectively. More generally, for a smooth loss function, we show that the risk of the M-bagged predictor is sandwiched between similar convex combinations. In addition, we prescribe a generic cross-validation method to tune the subsample size that aims at obtaining the best subagged predictor, which also serves to monotonize the risk profile of any given prediction procedure.

Following this general strategy, along with certain novel derivations from random matrix theory (to analyze conditional resolvents), we obtain explicit risk characterization for bagged ridge and ridgeless predictors. The risk expressions reveal bias and variance monotonicity in the number of bags. Comparing different variants of bagging for ridge and ridgeless predictors, we show that subagging (with optimal subsample size) improves upon the divide-and-conquer or the data-splitting approach of averaging the predictors computed on different non-overlapping splits of data (with optimal split size). This is especially notable in the overparameterized regime, where the latter data-splitting has been recently observed to improve upon the ridgeless predictor computed on the entire data (Mücke et al., 2022) under sub-Gaussian features.

Surprisingly, our results show that, under a well-specified linear model, subagging on properly chosen ridgeless interpolators always improves upon the ridgeless predictor trained on the complete data, even when the entire data has more observations than the number of features. Moreover, our generic and model-agnostic cross-validation procedure provably yields the best ridgeless interpolators for subagging. Further specializing to the case of isotropic features, we prove that the optimal subagged predictor has the asymptotic risk that matches the unbagged ridge predictor with optimally-tuned regularization parameter.

Several natural extensions of the current work can be considered going forward. We briefly discuss two of them below.

First, although our proposed general strategy for analyzing bagged predictors can be helpful for other prediction procedures, we have only derived the precise bagged risk expressions for the ridge and ridgeless regression. In the context of the ridge and ridgeless predictors, we had to develop new random matrix theory tools related to conditional asymptotic equivalents. It may be necessary to develop similar new tools to analyze other predictors based on our strategy. A natural prediction procedure to analyze next for bagging is the lasso or lassoless regression. An empirical investigation of the bagged lassoless predictor has already been conducted by Patil et al. (2022) (see Figure 8, for example). The traditional analysis of this predictor trained on the full data is performed via approximate message passing (AMP) techniques (Li and Wei, 2021). It would be interesting to see if our general strategy can be combined with AMP, the convex Gaussian min-max theorem, or the leave-one-out perturbation analysis to yield a more encompassing strategy for bagging analysis.

Second, we have analyzed the bagged ridge and ridgeless predictors under a well-specified linear model. It is interesting to extend the analysis to a general data-distributional setting for two main reasons: (1) to make the results more relevant for practical data analysis, and (2) to investigate whether bagging interpolators can still improve upon the ridgeless predictor trained on the full data. Regarding (1) above, the techniques developed by Bartlett et al.

(2021) may prove useful in relaxing the linear model assumptions. Regarding (2) above, we performed a simple simulation study that suggests that even in the misspecified nonlinear model, bagging properly selected interpolators can improve the unbagged ridgeless predictor. See Figure 11 for more details. Making these empirical observations more precise presents an exciting avenue for future work.

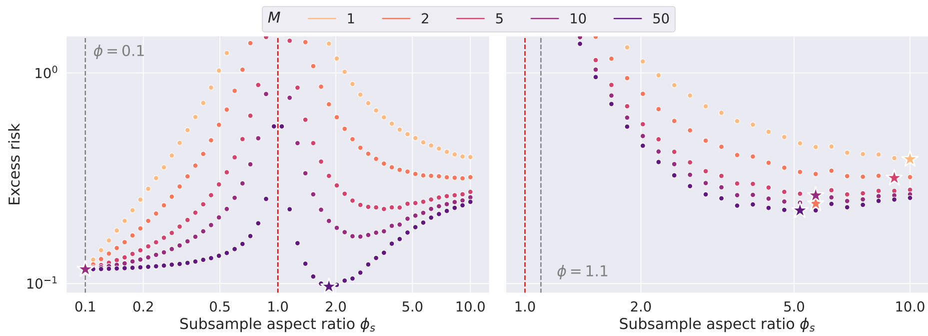

Figure 11: Finite-sample prediction risks for subagged ridgeless predictors ($\lambda = 0$) under a nonlinear model, averaged over 100 dataset repetitions, for varying bag size $k = [p/\phi_s]$ and number of bags M with replacement, with $n = [p/\phi]$ and $p = 500$. The left and the right panels correspond to the cases when $p < n$ ($\phi = 0.1$) and $p > n$ ($\phi = 1.1$), respectively. We generated data from a nonlinear model where the response y_i for $i \in [n]$ is generated from a nonlinear function of $\boldsymbol{x}_i$ with additive noise: $y_i = \boldsymbol{x}_i^\top \boldsymbol{\beta}_0 + \frac{1}{p}(\|\boldsymbol{x}_i\|_2^2 - \mathrm{tr}(\boldsymbol{\Sigma}_{\mathrm{ar1}})) + \epsilon_i$ and $\boldsymbol{\beta}_0, \boldsymbol{X}, \boldsymbol{\epsilon}$ are generated as in (M-AR1-LI) with $\rho_{\mathrm{ar1}} = 0.25$ and $\sigma^2 = 1$. We observe a similar pattern as in Figure 3 that the risk of the subagged ridgeless predictor with $M = 50$ and $\phi_s \approx 1.5$ is smaller than the risk of the ridgeless predictor fitted on the full data. Consequently, it is likely that the key results on subagging continue to hold under more general response models.

Acknowledgments

We are indebted to Ryan J. Tibshirani for his encouragement and advice throughout this work. We are also indebted to Alessandro Rinaldo, Yuting Wei, Matey Neykov, and other members of the Operational Overparameterized Statistics (OOPS) Working Group at Carnegie Mellon University, for many insightful conversations. We warmly thank the participants of the Theory of Overparameterized Learning Workshop (TOPML) 2022, in particular Daniel LeJeune, for many enlightening discussions. We sincerely thank three anonymous reviewers and the action editor for the swift review and several useful comments and suggestions that have improved the manuscript.

P. Patil was partially supported by ONR grant N00014-20-1-2787. A. K. Kuchibhotla was partially supported by NSF grant DMS-2113611.

References

Athey, S., Tibshirani, J., and Wager, S. (2019). Generalized random forests. *The Annals of Statistics*, 47(2):1148–1178.

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

Banerjee, M., Durot, C., and Sen, B. (2019). Divide and conquer in nonstandard problems and the super-efficiency phenomenon. *The Annals of Statistics*, 47(2):720–757.

Barber, D., Saad, D., and Sollich, P. (1995). Finite-size effects and optimal test set size in linear perceptrons. *Journal of Physics A: Mathematical and General*, 28(5):1325.

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

Belkin, M. (2021). Fit without fear: remarkable mathematical phenomena of deep learning through the prism of interpolation. *Acta Numerica*, 30:203–248.

Belkin, M., Hsu, D., Ma, S., and Mandal, S. (2019). Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*, 116(32):15849–15854.

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

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

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

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

Cannings, T. I. and Samworth, R. J. (2017). Random-projection ensemble classification. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(4):959–1035.

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

Couillet, R. and Debbah, M. (2011). *Random matrix methods for wireless communications*. Cambridge University Press.

Dar, Y., Muthukumar, V., and Baraniuk, R. G. (2021). A farewell to the bias-variance tradeoff? An overview of the theory of overparameterized machine learning. *arXiv preprint arXiv:2109.02355*.

Deng, Z., Kammoun, A., and Thrampoulidis, C. (2022). A model of double descent for high-dimensional binary linear classification. *Information and Inference: A Journal of the IMA*, 11(2):435–495.

Dereziński, M. (2023). Algorithmic gaussianization through sketching: Converting data into sub-gaussian random designs. In *The Annual Conference on Learning Theory*.

Derezinski, M., Bartan, B., Pilanci, M., and Mahoney, M. W. (2020). Debiasing distributed second order optimization with surrogate sketching and scaled regularization. *Advances in Neural Information Processing Systems*.

Dobriban, E. and Sheng, Y. (2020). Wonder: Weighted one-shot distributed ridge regression in high dimensions. *Journal of Machine Learning Research*, 21(66):1–52.

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

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

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

Duin, R. P. W. (1995). Small sample size generalization. In *Proceedings of the Scandinavian Conference on Image Analysis*, volume 2, pages 957–964. Citeseer.

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

Fernández-Delgado, M., Cernadas, E., Barro, S., and Amorim, D. (2014). Do we need hundreds of classifiers to solve real world classification problems? *The Journal of Machine Learning Research*, 15(1):3133–3181.

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

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

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

Hansen, L. K. (1993). Stochastic linear learning: Exact test and training error averages. *Neural Networks*, 6(3):393–396.

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.

Hertz, J. A., Krogh, A., and Thorbergsson, G. I. (1989). Phase transitions in simple learning. *Journal of Physics A: Mathematical and General*, 22(12):2133.

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

LeJeune, D., Patil, P., Javadi, H., Baraniuk, R. G., and Tibshirani, R. J. (2022). Asymptotics of the sketched pseudoinverse. *arXiv preprint arXiv:2211.03751*.

Li, Y. and Wei, Y. (2021). Minimum ℓ_1-norm interpolators: Precise asymptotics and multiple descent. *arXiv preprint arXiv:2110.09502*.

Liu, F., Liao, Z., and Suykens, J. (2021). Kernel regression in high dimensions: Refined analysis beyond double descent. In *International Conference on Artificial Intelligence and Statistics*.

Loog, M., Viering, T., and Mey, A. (2019). Minimizers of the empirical risk and risk monotonicity. *Advances in Neural Information Processing Systems*, 32.

Loog, M., Viering, T., Mey, A., Krijthe, J. H., and Tax, D. M. (2020). A brief prehistory of double descent. *Proceedings of the National Academy of Sciences*, 117(20):10625–10626.

Lopes, M., Jacob, L., and Wainwright, M. J. (2011). A more powerful two-sample test in high dimensions using random projection. *Advances in Neural Information Processing Systems*.

Lopes, M., Wang, S., and Mahoney, M. (2018). Error estimation for randomized least-squares algorithms via the bootstrap. In *International Conference on Machine Learning*.

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

Mhammedi, Z. (2021). Risk monotonicity in statistical learning. *Advances in Neural Information Processing Systems*.

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.

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*, pages 10322–10350. PMLR.

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

Nakkiran, P. (2019). More data can hurt for linear regression: Sample-wise double descent. *arXiv preprint arXiv:1912.07242*.

Nakkiran, P., Venkat, P., Kakade, S. M., and Ma, T. (2021). Optimal regularization can mitigate double descent. In *International Conference on Learning Representations*.

Opper, M. (1995). Statistical mechanics of learning: Generalization. *The handbook of brain theory and neural networks*, pages 922–925.

Opper, M. and Kinzel, W. (1996). Statistical mechanics of generalization. In *Models of Neural Networks III: Association, Generalization, and Representation*, pages 151–209. Springer.

Opper, M., Kinzel, W., Kleinz, J., and Nehl, R. (1990). On the ability of the optimal perceptron to generalise. *Journal of Physics A: Mathematical and General*, 23(11):L581.

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

Patil, P. and LeJeune, D. (2023). Asymptotically free sketched ridge ensembles: Risks, cross-validation, and tuning. *arXiv preprint arXiv:2310.04357*.

Raudys, S. and Duin, R. P. W. (1998). Expected classification error of the fisher linear classifier with pseudo-inverse covariance matrix. *Pattern recognition letters*, 19(5-6):385–392.

Rosenblatt, J. D. and Nadler, B. (2016). On the optimality of averaging in distributed statistical learning. *Information and Inference: A Journal of the IMA*, 5(4):379–404.

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

Srivastava, R., Li, P., and Ruppert, D. (2016). RAPTT: An exact two-sample test in high dimensions using random projections. *Journal of Computational and Graphical Statistics*, 25(3):954–970.

Strobl, C., Malley, J., and Tutz, G. (2009). An introduction to recursive partitioning: Rationale, application, and characteristics of classification and regression trees, bagging, and random forests. *Psychological Methods*, 14:323–48.

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

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.

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

Vallet, F., Cailton, J.-G., and Refregier, P. (1989). Linear and nonlinear extension of the pseudo-inverse solution for learning boolean functions. *Europhysics Letters*, 9(4):315.

Viering, T., Mey, A., and Loog, M. (2019). Open problem: Monotonicity of learning. In *Conference on Learning Theory*.

Wang, S., Gittens, A., and Mahoney, M. W. (2017). Sketched ridge regression: Optimization perspective, statistical perspective, and model averaging. In *International Conference on Machine Learning*.

Wu, Y.-S., Masegosa, A., Lorenzen, S., Igel, C., and Seldin, Y. (2021). Chebyshev-Cantelli PAC-Bayes-Bennett inequality for the weighted majority vote. *Advances in Neural Information Processing Systems*.

Appendix A. Notation and organization

Notation

Below we provide an overview of some general notation used in the main paper and the supplement.

We denote scalars in non-bold lower or upper case (e.g., n, λ, C), vectors in bold lower case (e.g., $\boldsymbol{x}$, $\boldsymbol{\beta}$), and matrices in bold upper case (e.g., $\boldsymbol{X}$). We denote sets using calligraphic letters (e.g., $\mathcal{D}$), and use blackboard letters to denote some special sets: $\mathbb{N}$ denotes the set of positive integers, $\mathbb{R}$ denotes the set of real numbers, $\mathbb{R}_{\geq 0}$ denotes the set of non-negative real numbers, $\mathbb{R}_{>0}$ denotes the set of positive real numbers, $\mathbb{C}$ denotes the set of complex numbers, $\mathbb{C}^{+}$ denotes the set of complex numbers with positive imaginary part, and $\mathbb{C}^{-}$ denotes the set of complex numbers with negative imaginary part. For a natural number n, we use $[n]$ to denote the set $\{1, \ldots, n\}$.

For a real number x, $(x)_{+}$ denotes its positive part, $\lfloor x \rfloor$ its floor, and $\lceil x \rceil$ its ceiling. For a vector $\boldsymbol{\beta}$, $\|\boldsymbol{\beta}\|_2$ denotes its ℓ_2 norm. For a pair of vectors $\boldsymbol{v}$ and $\boldsymbol{w}$, $\langle \boldsymbol{v}, \boldsymbol{w} \rangle$ denotes their inner product. For an event A, $\mathbb{1}_A$ denotes the associated indicator random variable. For a matrix $\boldsymbol{X} \in \mathbb{R}^{n \times p}$, $\boldsymbol{X}^{\top} \in \mathbb{R}^{p \times n}$ denotes its transpose, and $\boldsymbol{X}^{+} \in \mathbb{R}^{p \times n}$ denote its Moore-Penrose inverse. For a square matrix $\boldsymbol{A} \in \mathbb{R}^{p \times p}$, $\mathrm{tr}[\boldsymbol{A}]$ denotes its trace, and $\boldsymbol{A}^{-1} \in \mathbb{R}^{p \times p}$ denotes its inverse, provided it is invertible. For a positive semidefinite matrix $\boldsymbol{\Sigma}$, $\boldsymbol{\Sigma}^{1/2}$ denotes its principal square root. A $p \times p$ identity matrix is denoted $\boldsymbol{I}_p$, or simply by $\boldsymbol{I}$, when it is clear from the context.

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

For symmetric matrices $\boldsymbol{A}$ and $\boldsymbol{B}$, $\boldsymbol{A} \preceq \boldsymbol{B}$ denotes the Loewner ordering. For sequences of matrices $\boldsymbol{A}_n$ and $\boldsymbol{B}_n$, $\boldsymbol{A}_n \simeq \boldsymbol{B}_n$ denotes a certain notion of asymptotic equivalence (see Definitions 30 and 31). We use O_p and o_p to denote probabilistic big-O and little-o notation, respectively. We denote convergence in probability by "$\xrightarrow{\mathrm{p}}$", almost sure convergence by "$\xrightarrow{\mathrm{a.s.}}$", and convergence in distribution by "$\xrightarrow{\mathrm{d}}$".

Organization

Below we outline the structure of the rest of the supplement.

- In Section B, we present proofs of results related to general subagged predictors from Section 2.
- In Sections C and D, we present proof of Theorem 6 related to subagging from Section 3.2 for ridge and ridgeless predictors, respectively. The proofs for the two cases are separated due to length. However, the proof architecture for the two is similar.
- In Section E, we present proof of Theorem 8 related to splagging from Section 3.3 for ridge and ridgeless predictors. Because some of this proof builds on that of Theorem 6,

we can combine the two cases of the ridge and ridgeless predictors, unlike the split cases for Theorem 6.

- In Section F, we present proofs of results related to the bias-variance component monotonicity properties in Propositions 7 and 9 for subagging and splagging, respectively. In this section, we also provide proofs of results related to cross-validation and profile monotonicity and those related to oracle properties of optimized bagging from Section 4.
- In Section G, we present proofs of specialized results related to subagging and splagging under isotopic features from Section 5.
- In Section H, we formalize several calculus rules for a certain notion of conditional asymptotic equivalence of sequences of matrices that are used in the proofs of constituent lemmas in Sections C to E.
- In Section I, we collect various technical helper lemmas related to concentrations and convergences along with their proofs that are used in proofs in Sections B to E.
- In Section J, we present additional numerical illustrations for Theorems 6, 8 and 11, and for specialized isotropic results from Section 5.

Appendix B. Proofs in Section 2 (general bagged predictors)

B.1 Proof of Proposition 2 (asymptotic data conditional risk, squared loss)

Proof The key idea in the proof is to use the conditional risk decomposition from Proposition 1. Below we present the proof for sampling from $\mathcal{I}_k$. The proof for sampling from $\mathcal{I}_k^\pi$ is analogous.

SRSWR. We will do the case of SRSWR from $\mathcal{I}_k$ first. From Proposition 1, we have

$$\begin{aligned}
R(\widetilde{f}_M; \mathcal{D}_n) &= \mathbb{E}_{(\boldsymbol{x},y)}[\mathbb{E}[(\widetilde{f}_M - y)^2 \mid \mathcal{D}_n, (\boldsymbol{x}, y)]] \\
&= \mathbb{E}_{(\boldsymbol{x},y)}\left[\mathscr{B}_{\mathcal{I}_k}(\boldsymbol{x}, y) \,|\, \mathcal{D}_n\right] + \frac{1}{M}\mathbb{E}_{(\boldsymbol{x},y)}\left[\mathscr{V}_{\mathcal{I}_k}(\boldsymbol{x}, y) \,|\, \mathcal{D}_n\right] \\
&= R(\widetilde{f}_\infty; \mathcal{D}_n) + \frac{1}{M} C_n, \qquad (39)
\end{aligned}$$

where $C_n = \mathbb{E}_{(\boldsymbol{x},y)}\left[\frac{1}{|\mathcal{I}_k|}\sum_{I\in\mathcal{I}_k}\left(\widehat{f}(\boldsymbol{x};\mathcal{D}_I) - \widetilde{f}_{\infty,\mathcal{I}_k}(\boldsymbol{x})\right)^2 \,\middle|\, \mathcal{D}_n\right]$.

Since for $M = 1$ and $M = 2$, we have

$$\begin{aligned}
R(\widetilde{f}_1; \mathcal{D}_n) &= R(\widetilde{f}_\infty; \mathcal{D}_n) + C_n, \\
R(\widetilde{f}_2; \mathcal{D}_n) &= R(\widetilde{f}_\infty; \mathcal{D}_n) + \frac{C_n}{2}.
\end{aligned}$$

We can thus write $R(\widetilde{f}_\infty; \mathcal{D}_n)$ and C_n in terms of $R(\widetilde{f}^{\mathtt{WR}}_{1,\mathcal{I}_k}; \mathcal{D}_n)$ and $R(\widetilde{f}^{\mathtt{WR}}_{2,\mathcal{I}_k}; \mathcal{D}_n)$ as

$$\begin{aligned}
R(\widetilde{f}_\infty; \mathcal{D}_n) &= 2R(\widetilde{f}_2; \mathcal{D}_n) - R(\widetilde{f}_1; \mathcal{D}_n), \\
C_n &= 2R(\widetilde{f}_1; \mathcal{D}_n) - 2R(\widetilde{f}_2; \mathcal{D}_n).
\end{aligned}$$

Substituting in (39), we obtain

$$\begin{aligned} R(\widetilde{f}_M;\mathcal{D}_n) &= 2R(\widetilde{f}_2;\mathcal{D}_n) - R(\widetilde{f}_1;\mathcal{D}_n) + \frac{1}{M}\Big(2R(\widetilde{f}_1;\mathcal{D}_n) - 2R(\widetilde{f}_2;\mathcal{D}_n)\Big) \\ &= -\left(1-\frac{2}{M}\right)R(\widetilde{f}_1;\mathcal{D}_n) + \left(2-\frac{2}{M}\right)R(\widetilde{f}_2;\mathcal{D}_n). \end{aligned}$$

Thus, subtracting the desired target in (10) for with replacement from both sides, we get

$$R(\widetilde{f}_M;\mathcal{D}_n) - \left[(2a_2-a_1) + \frac{2(a_1-a_2)}{M}\right] = -\left(1-\frac{2}{M}\right)\Big(R(\widetilde{f}_1;\mathcal{D}_n) - a_1\Big) + \left(2-\frac{2}{M}\right)\Big(R(\widetilde{f}_2;\mathcal{D}_n) - a_2\Big).$$

Taking absolute values on both sides and using triangle inequality yields

$$\left|R(\widetilde{f}_M;\mathcal{D}_n) - \left[(2a_2-a_1) + \frac{2(a_1-a_2)}{M}\right]\right| \le \left|1-\frac{2}{M}\right|\left|R(\widetilde{f}_1;\mathcal{D}_n) - a_1\right| + \left(2-\frac{2}{M}\right)\left|R(\widetilde{f}_2;\mathcal{D}_n) - a_2\right|.$$

Taking supremum over M, we have

$$\sup_{M\in\mathbb{N}}\left|R(\widetilde{f}_M;\mathcal{D}_n) - \left[(2a_2-a_1) + \frac{2(a_1-a_2)}{M}\right]\right| \le \left|R(\widetilde{f}_1;\mathcal{D}_n) - a_1\right| + 2\left|R(\widetilde{f}_2;\mathcal{D}_n) - a_2\right|.$$

Finally, since we have

$$R(\widetilde{f}_1;\mathcal{D}_n) \xrightarrow{\text{a.s.}} a_1, \qquad R(\widetilde{f}_2;\mathcal{D}_n) \xrightarrow{\text{a.s.}} a_2,$$

the desired claim in (10) for with replacement follows.

SRSWOR. For SRSWOR from $\mathcal{I}_k$, similarly we have

$$\begin{aligned} R(\widetilde{f}_M;\mathcal{D}_n) &= \mathbb{E}_{(\boldsymbol{x},y)}[\mathbb{E}[(\widetilde{f}_M - y)^2|\mathcal{D}_n,(\boldsymbol{x},y)]] \\ &= \mathbb{E}_{(\boldsymbol{x},y)}\left[\mathscr{B}_{\mathcal{I}_k}(\boldsymbol{x},y)\mid\mathcal{D}_n\right] + \frac{|\mathcal{I}_k|-M}{|\mathcal{I}_k|-1}\frac{1}{M}\mathbb{E}_{(\boldsymbol{x},y)}\left[\mathscr{V}_{\mathcal{I}_k}(\boldsymbol{x},y)\mid\mathcal{D}_n\right] \\ &= R(\widetilde{f}_\infty;\mathcal{D}_n) + \frac{|\mathcal{I}_k|-M}{|\mathcal{I}_k|-1}\frac{1}{M}C_n \\ &= R(\widetilde{f}_\infty;\mathcal{D}_n) - \frac{C_n}{|\mathcal{I}_k|-1} + \frac{1}{M}\cdot\frac{|\mathcal{I}_k|C_n}{|\mathcal{I}_k|-1}, \end{aligned} \tag{40}$$

where $C_n = \mathbb{E}_{(\boldsymbol{x},y)}\left[\frac{1}{|\mathcal{I}_k|}\sum_{I\in\mathcal{I}_k}\left(\widehat{f}(\boldsymbol{x};\mathcal{D}_I) - \widetilde{f}_{\infty,\mathcal{I}_k}(\boldsymbol{x})\right)^2\,\middle|\,\mathcal{D}_n\right]$. Since for $M=1$ and $M=2$,

$$\begin{aligned} R(\widetilde{f}_1;\mathcal{D}_n) &= R(\widetilde{f}_\infty;\mathcal{D}_n) - \frac{C_n}{|\mathcal{I}_k|-1} + \frac{|\mathcal{I}_k|C_n}{|\mathcal{I}_k|-1}, \\ R(\widetilde{f}_2;\mathcal{D}_n) &= R(\widetilde{f}_\infty;\mathcal{D}_n) - \frac{C_n}{|\mathcal{I}_k|-1} + \frac{1}{2}\cdot\frac{|\mathcal{I}_k|C_n}{|\mathcal{I}_k|-1}. \end{aligned}$$

We can thus write $R(\widetilde{f}_\infty;\mathcal{D}_n) - C_n/(|\mathcal{I}_k|-1)$ and $|\mathcal{I}_k|C_n/(|\mathcal{I}_k|-1)$ in terms of $R(\widetilde{f}^{\texttt{WR}}_{1,\mathcal{I}_k};\mathcal{D}_n)$ and $R(\widetilde{f}^{\texttt{WR}}_{2,\mathcal{I}_k};\mathcal{D}_n)$ as

$$R(\widetilde{f}_\infty;\mathcal{D}_n) - \frac{C_n}{|\mathcal{I}_k|-1} = 2R(\widetilde{f}_2;\mathcal{D}_n) - R(\widetilde{f}_1;\mathcal{D}_n),$$

$$\frac{|\mathcal{I}_k| C_n}{|\mathcal{I}_k| - 1} = 2(R(\widetilde{f}_1; \mathcal{D}_n) - R(\widetilde{f}_2; \mathcal{D}_n)).$$

Substituting in (40), we obtain

$$\begin{aligned} R(\widetilde{f}_M; \mathcal{D}_n) &= 2R(\widetilde{f}_2; \mathcal{D}_n) - R(\widetilde{f}_1; \mathcal{D}_n) + \frac{1}{M} \cdot 2(R(\widetilde{f}_1; \mathcal{D}_n) - R(\widetilde{f}_2; \mathcal{D}_n)) \\ &= -\left(1 - \frac{2}{M}\right) R(\widetilde{f}_1; \mathcal{D}_n) + 2\left(1 - \frac{1}{M}\right) R(\widetilde{f}_2; \mathcal{D}_n). \end{aligned}$$

Thus, subtracting the desired target in (10) for with replacement from both sides, we get

$$R(\widetilde{f}_M; \mathcal{D}_n) - \left[(2a_2 - a_1) + \frac{2(a_1 - a_2)}{M}\right] = -\left(1 - \frac{2}{M}\right)\left(R(\widetilde{f}_1; \mathcal{D}_n) - a_1\right) + \left(2 - \frac{2}{M}\right)\left(R(\widetilde{f}_2; \mathcal{D}_n) - a_2\right).$$

Taking absolute values on both sides and using triangle inequality yields

$$\left|R(\widetilde{f}_M; \mathcal{D}_n) - \left[(2a_2 - a_1) + \frac{2(a_1 - a_2)}{M}\right]\right| \le \left|1 - \frac{2}{M}\right| \left|R(\widetilde{f}_1; \mathcal{D}_n) - a_1\right| + \left(2 - \frac{2}{M}\right)\left|R(\widetilde{f}_2; \mathcal{D}_n) - a_2\right|.$$

Taking supremum over M, we have

$$\sup_{M \in \mathbb{N}} \left|R(\widetilde{f}_M; \mathcal{D}_n) - \left[(2a_2 - a_1) + \frac{2(a_1 - a_2)}{M}\right]\right| \le \left|R(\widetilde{f}_1; \mathcal{D}_n) - a_1\right| + 2\left|R(\widetilde{f}_2; \mathcal{D}_n) - a_2\right|.$$

Finally, since we have $R(\widetilde{f}_1; \mathcal{D}_n) \xrightarrow{\text{a.s.}} a_1$ and $R(\widetilde{f}_2; \mathcal{D}_n) \xrightarrow{\text{a.s.}} a_2$, the desired claim in (10) for the case of sampling without replacement follows. ■

B.2 Proof of Proposition 3 (asymptotic subsample conditional risk, squared loss)

Before we present the proof for Proposition 3, we first show the upper bound of the squared subsample conditional risk for general M.

Lemma 17 (Bounding the squared subsample conditional risk)**.** *The subsample conditional prediction risk defined in* (5) *for the bagged predictor* $\widehat{f}_{M,\mathcal{I}_k}$ *can be bounded as:*

$$\begin{aligned} &\left|R(\widetilde{f}_{M,\mathcal{I}_k}; \mathcal{D}_n, \{I_\ell\}_{\ell=1}^M) - \left\{(2b_2 - b_1) + \frac{2(b_1 - b_2)}{M}\right\}\right| \\ &\le \left|\frac{1}{M}\sum_{\ell=1}^M R(\widetilde{f}_{1,\mathcal{I}_k}; \mathcal{D}_n, \{I_\ell\}) - b_1\right| + 2\left|\frac{1}{M(M-1)} \sum_{i,j \in [M], i \neq j} R(\widetilde{f}_{2,\mathcal{I}_k}; \mathcal{D}_n, \{I_i, I_j\}) - b_2\right|. \end{aligned} \tag{41}$$

Proof We start by expanding the squared risk as:

$R(\widetilde{f}_{M,\mathcal{I}_k};\, \mathcal{D}_n, \{I_\ell\}_{\ell=1}^M)$

$$
\begin{aligned}
&= \int \left(y - \frac{1}{M} \sum_{\ell=1}^{M} \widehat{f}(\boldsymbol{x}; \mathcal{D}_{I_\ell}) \right)^2 \, \mathrm{d}P(\boldsymbol{x}, y) \\
&= \int \left(\frac{1}{M} \sum_{\ell=1}^{M} \big(y - \widehat{f}(\boldsymbol{x}; \mathcal{D}_{I_\ell}) \big) \right)^2 \, \mathrm{d}P(\boldsymbol{x}, y) \\
&= \frac{1}{M^2} \sum_{\ell=1}^{M} \int \big(y - \widehat{f}(\boldsymbol{x}; \mathcal{D}_{I_\ell}) \big)^2 \, \mathrm{d}P(\boldsymbol{x}, y) + \frac{1}{M^2} \sum_{i=1}^{M} \sum_{\substack{j=1 \\ j \neq i}}^{M} \int \big(y - \widehat{f}(\boldsymbol{x}; \mathcal{D}_{I_i}) \big) \big(y - \widehat{f}(\boldsymbol{x}; \mathcal{D}_{I_j}) \big) \, \mathrm{d}P(\boldsymbol{x}, y) \\
&= \frac{1}{M^2} \sum_{\ell=1}^{M} R(\widetilde{f}_{1,\mathcal{I}_k}; \mathcal{D}_n, I_\ell) + \frac{1}{M^2} \sum_{i=1}^{M} \sum_{\substack{j=1 \\ j \neq i}}^{M} \int (y - \widehat{f}(\boldsymbol{x}; \mathcal{D}_{I_i}))(y - \widehat{f}(\boldsymbol{x}; \mathcal{D}_{I_j})) \, \mathrm{d}P(\boldsymbol{x}, y) \\
&\overset{(i)}{=} \frac{1}{M^2} \sum_{\ell=1}^{M} R(\widetilde{f}_{1,\mathcal{I}_k}; \mathcal{D}_n, I_\ell) \\
&\qquad + \frac{1}{M^2} \sum_{i=1}^{M} \sum_{\substack{j=1 \\ j \neq i}}^{M} \int \frac{1}{2} \left\{ 4 \Big(y - \frac{1}{2} \big(\widehat{f}(\boldsymbol{x}; \mathcal{D}_{I_i}) + \widehat{f}(\boldsymbol{x}; \mathcal{D}_{I_j}) \big) \Big)^2 - \big(y - \widehat{f}(\boldsymbol{x}; \mathcal{D}_{I_i}) \big)^2 - \big(y - \widehat{f}(\boldsymbol{x}; \mathcal{D}_{I_j}) \big)^2 \right\} \mathrm{d}P(\boldsymbol{x}, y) \\
&= \frac{1}{M^2} \sum_{\ell=1}^{M} R(\widetilde{f}_{1,\mathcal{I}_k}; \mathcal{D}_n, I_\ell) \\
&\qquad + \frac{1}{M^2} \sum_{i=1}^{M} \sum_{\substack{j=1 \\ j \neq i}}^{M} \frac{1}{2} \left\{ 4R(\widehat{f}_{2,\mathcal{I}_k}; \mathcal{D}_n; I_i, I_j) - R(\widetilde{f}_{1,\mathcal{I}_k}; \mathcal{D}_n; I_i) - R(\widetilde{f}_{1,\mathcal{I}_k}; \mathcal{D}_n; I_j) \right\} \\
&= \frac{1}{M^2} \sum_{\ell=1}^{M} R(\widetilde{f}_{1,\mathcal{I}_k}; \mathcal{D}_n, I_\ell) - \frac{1}{2M^2} \sum_{i=1}^{M} \sum_{\substack{j=1 \\ j \neq i}}^{M} R(\widetilde{f}_{1,\mathcal{I}_k}; I_i) - \frac{1}{2M^2} \sum_{i=1}^{M} \sum_{\substack{j=1 \\ j \neq i}}^{M} R(\widetilde{f}_{1,\mathcal{I}_k}; I_j) + \frac{1}{M^2} \sum_{i=1}^{M} \sum_{\substack{j=1 \\ j \neq i}}^{M} 2R(\widehat{f}_{2,\mathcal{I}_k}; \mathcal{D}_n; I_i, \\
&= \frac{1}{M^2} \sum_{\ell=1}^{M} R(\widetilde{f}_{1,\mathcal{I}_k}; \mathcal{D}_n; I_\ell) - \frac{1}{2M^2} \cdot 2 \cdot (M-1) \sum_{\ell=1}^{M} R(\widetilde{f}_{1,\mathcal{I}_k}; I_\ell) + \frac{2}{M^2} \sum_{\substack{i,j \in [M] \\ i \neq j}} R(\widehat{f}_{2,\mathcal{I}_k}; \mathcal{D}_n; I_i, I_j) \\
&= \left(\frac{1}{M^2} - \frac{(M-1)}{M^2} \right) \sum_{\ell=1}^{M} R(\widetilde{f}_{1,\mathcal{I}_k}; \mathcal{D}_n; I_\ell) + \frac{2}{M^2} \sum_{\substack{i,j \in [M] \\ i \neq j}} R(\widehat{f}_{2,\mathcal{I}_k}; \mathcal{D}_n; I_i, I_j) \\
&= -\left(\frac{1}{M} - \frac{2}{M^2} \right) \sum_{\ell=1}^{M} R(\widetilde{f}_{1,\mathcal{I}_k};\, \mathcal{D}_n, \{I_\ell\}) + \frac{2}{M^2} \sum_{\substack{i,j \in [M] \\ i \neq j}} R(\widetilde{f}_{2,\mathcal{I}_k};\, \mathcal{D}_n, \{I_i, I_j\}).
\end{aligned}
$$

In the expansion above, for equality (i), we used the fact that $ab = \{4(a/2+b/2)^2 - a^2 - b^2\}/2$.

Now, subtracting the desired limit on both sides yields

$$
\left| R(\widetilde{f}_{M,\mathcal{I}_k};\, \mathcal{D}_n, \{I_\ell\}_{\ell=1}^{M}) - \left\{ (2b_2 - b_1) + \frac{2(b_1 - b_2)}{M} \right\} \right|
$$

$$
\begin{aligned}
&= \left| -\left(\frac{1}{M} - \frac{2}{M^2}\right) \sum_{\ell=1}^{M} (R(\widetilde{f}_{1,\mathcal{I}_k}; \mathcal{D}_n, \{I_\ell\}) - b_1) + \frac{2}{M^2} \sum_{\substack{i,j \in [M] \\ i \neq j}} (R(\widetilde{f}_{2,\mathcal{I}_k}; \mathcal{D}_n, \{I_i, I_j\}) - b_2) \right| \\
&\leq \left| 1 - \frac{2}{M} \right| \cdot \left| \frac{1}{M} \sum_{\ell=1}^{M} R(\widetilde{f}_{1,\mathcal{I}_k}; \mathcal{D}_n, \{I_\ell\}) - b_1 \right| + \frac{2(M-1)}{M} \left| \frac{1}{M(M-1)} \sum_{\substack{i,j \in [M] \\ i \neq j}} R(\widetilde{f}_{2,\mathcal{I}_k}; \mathcal{D}_n, \{I_i, I_j\}) - b_2 \right| \\
&\leq \left| \frac{1}{M} \sum_{\ell=1}^{M} R(\widetilde{f}_{1,\mathcal{I}_k}; \mathcal{D}_n, \{I_\ell\}) - b_1 \right| + 2 \left| \frac{1}{M(M-1)} \sum_{\substack{i,j \in [M] \\ i \neq j}} R(\widetilde{f}_{2,\mathcal{I}_k}; \mathcal{D}_n, \{I_i, I_j\}) - b_2 \right|.
\end{aligned}
$$

This completes the proof of the upper bound. ■

Next, we present the proof of Proposition 3.

Proof Lemma 4 implies the asymptotics for the data conditional risk. Now, consider the asymptotics for the subsample conditional risk of the bagged predictors. From (41) of Lemma 17, it holds that

$$
\begin{aligned}
&\left| R(\widetilde{f}_{M,\mathcal{I}_k}; \mathcal{D}_n, \{I_\ell\}_{\ell=1}^M) - \left\{ (2b_2 - b_1) + \frac{2(b_1 - b_2)}{M} \right\} \right| \\
&\leq \left| \frac{1}{M} \sum_{\ell=1}^{M} R(\widetilde{f}_{1,\mathcal{I}_k}; \mathcal{D}_n, \{I_\ell\}) - b_1 \right| + 2 \left| \frac{1}{M(M-1)} \sum_{i,j \in [M], i \neq j} R(\widetilde{f}_{2,\mathcal{I}_k}; \mathcal{D}_n, \{I_i, I_j\}) - b_2 \right|.
\end{aligned} \tag{42}
$$

This implies that

$$
\begin{aligned}
&\sup_{M \in \mathbb{N}} \left| R(\widetilde{f}_{M,\mathcal{I}_k}; \mathcal{D}_n, \{I_\ell\}_{\ell=1}^M) - \left\{ (2b_2 - b_1) + \frac{2(b_1 - b_2)}{M} \right\} \right| \\
&\leq \sup_{I \in \mathcal{I}_k} |R(\widetilde{f}_{1,\mathcal{I}_k}; \mathcal{D}_n, \{I\}) - b_1| + 2 \sup_{M \geq 2} \left| \frac{1}{M(M-1)} \sum_{i,j \in [M], i \neq j} R(\widetilde{f}_{2,\mathcal{I}_k}; \mathcal{D}_n, \{I_i, I_j\}) - b_2 \right|.
\end{aligned}
$$

The first term on the right-hand side converges almost surely to zero by Lemma 50 (2), since the conditional risk for $M = 1$ converges for any sequence of indices. To prove that the second term converges to zero, we start by noting that

$$
U_M = \frac{1}{M(M-1)} \sum_{i,j \in [M], i \neq j} \left\{ R(\widetilde{f}_{2,\mathcal{I}_k}; \mathcal{D}_n, \{I_i, I_j\}) - b_2 \right\},
$$

is a U-statistics based on either an SRSWR or an SRSWOR sample $I_1, \ldots, I_M$ conditional on $\mathcal{D}_n$. Theorem 2 in Section 3.4.2 of Lee (1990) implies that $\{U_M\}_{M \geq 2}$ is a reverse martingale conditional on $\mathcal{D}_n$ with respect to some filtration when we have an SRSWR sample (which is same as an i.i.d. sample). Lemma 2.1 of Sen (1970) proves the same result when we have an SRSWOR sample. This, combined with Theorem 3 (maximal inequality for reverse martingales) in Section 3.4.1 of Lee (1990) (for $r = 1$[‖]) yields

$$\mathbb{P}\left(\sup_{M\geq 2}|U_M| \geq \delta \;\middle|\; \mathcal{D}_n\right) \leq \frac{1}{\delta}\mathbb{E}\left[|U_2| \;\middle|\; \mathcal{D}_n\right] = \frac{1}{\delta}\mathbb{E}\left[|R(\widetilde{f}_{2,\mathcal{I}_k};\mathcal{D}_n,\{I_1,I_2\}) - b_2| \;\middle|\; \mathcal{D}_n\right].$$

The right-hand side, we know, converges to zero almost surely. To see this, we first write as before the right-hand side as $\mathbb{E}[|R(\widetilde{f}_{2,\mathcal{I}_k};\mathcal{D}_n,\{I_1,I_2\}) - b_2| \mid \mathcal{D}_n = \mathcal{D}_n(\omega)] = \mathbb{E}[|R(\widetilde{f}_{2,\mathcal{I}_k};\mathcal{D}_n(\omega),\{I_1,I_2\}) - b_2|]$. We know that for all $\omega \in \mathcal{A}$, $R(\widetilde{f}_{2,\mathcal{I}_k};\mathcal{D}_n(\omega),\{I_1,I_2\}) \xrightarrow{\text{a.s.}} b_2$ as $n \to \infty$ (from the given assumption). Also, we know (45) and that the right-hand side of (45) converges in L_1 to its probability limit. Hence, Vitali's theorem (Bogachev, 2007, Theorem 4.5.4) implies that $\mathbb{E}[|R(\widetilde{f}_{2,\mathcal{I}_k};\mathcal{D}_n,\{I_1,I_2\}) - b_2| \mid \mathcal{D}_n = \mathcal{D}_n(\omega)]$ converges to zero for all $\omega \in \mathcal{A}$ as $n \to \infty$. Therefore, as $n \to \infty$, for all $\omega \in \mathcal{A}$,

$$\mathbb{P}\left(\sup_{M\geq 2}|U_M| \geq \delta \;\middle|\; \mathcal{D}_n = \mathcal{D}_n(\omega)\right) \to 0.$$

Because probabilities are bounded by one, dominated convergence theorem implies that

$$\mathbb{P}\left(\sup_{M\geq 2}|U_M| \geq \delta\right) \to 0, \quad \text{as} \quad n \to \infty.$$

Therefore,

$$\sup_{M\in\mathbb{N}}\left|R(\widetilde{f}_{M,\mathcal{I}_k};\mathcal{D}_n,\{I_\ell\}_{\ell=1}^M) - \left\{(2b_2 - b_1) + \frac{2(b_1 - b_2)}{M}\right\}\right| \xrightarrow{\text{p}} 0.$$

■

B.3 Proof of Lemma 4 (from subsample conditional to data conditional risk, $M = 1, 2$)

Proof Let us first prove the result when sampling with/without replacement from $\mathcal{I}_k$. The proof for $\mathcal{I}_k^\pi$ would be analogous. Note that $R(\widetilde{f}_1;\mathcal{D}_n) = \mathbb{E}[R(\widetilde{f}_{1,\mathcal{I}_k};\mathcal{D}_n,\{I_1\}) \mid \mathcal{D}_n]$ where the expectation is taken over a random draw I_1 from $\mathcal{I}_k$. We are given that $R(\widetilde{f}_{1,\mathcal{I}_k};\mathcal{D}_n,\{I\}) - b_1 \xrightarrow{\text{a.s.}} 0$ for every $I \in \mathcal{I}_k$. Although not explicitly highlighted, for clarity it is worth reminding that I is a sequence implicitly indexed by n. Under this condition, let us note that

$$\begin{aligned}\left|\mathbb{E}[R(\widetilde{f}_{1,\mathcal{I}_k};\mathcal{D}_n,\{I_1\}) \mid \mathcal{D}_n] - b_1\right| &= \left|\frac{1}{|\mathcal{I}_k|}\sum_{I\in\mathcal{I}_k}R(\widetilde{f}_{1,\mathcal{I}_k};\mathcal{D}_n,\{I\}) - b_1\right| \\ &\leq \frac{1}{|\mathcal{I}_k|}\sum_{I\in\mathcal{I}_k}|R(\widetilde{f}_{1,\mathcal{I}_k};\mathcal{D}_n,\{I\}) - b_1|\end{aligned}$$

[‖]Theorem 3 of Section 3.4.1 is only stated with $r > 1$, but from the proof, it is clear that $r = 1$ is a valid choice.

$$
\begin{aligned}
&\leq \max_{I \in \mathcal{I}_k} |R(\widetilde{f}_{1,\mathcal{I}_k}; \mathcal{D}_n, \{I\}) - b_1| \\
&\xrightarrow{\text{a.s.}} 0,
\end{aligned}
$$

by Lemma 50 (1), since again the conditional risk for $M = 1$ converges for any sequences of indices. To be explicit, the underlying triangular array invoked in Lemma 50 (1) is as described follows: For each n, recall $\mathcal{I}_k = \{I^{(1)}, \dots, I^{(N_n)}\}$ where $N_n = |\mathcal{I}_k| = \binom{n}{k}$. Note that the random quantity $R_{n,\ell} := R(\widetilde{f}_{1,\mathcal{I}_k}; \mathcal{D}_n, I^{(\ell)})$ is indexed by n and $\ell \in [N_n]$. For I drawn from $\mathcal{I}_k$, let p_n be the index such that $I = I^{(p_n)}$. The convergence then follows from applying Lemma 50 (1) to the triangular array R_{n,p_n}. Hence, we have proved that

$$
R(\widetilde{f}_{1,\mathcal{I}_k}; \mathcal{D}_n) \xrightarrow{\text{a.s.}} b_1, \quad \text{as} \quad n \to \infty. \tag{43}
$$

Now, observe that

$$
R(\widetilde{f}_{2,\mathcal{I}_k}; \mathcal{D}_n, \{I_\ell\}_{\ell=1}^2) \leq \frac{1}{2} R(\widetilde{f}_{1,\mathcal{I}_k}; \mathcal{D}_n, \{I_1\}) + \frac{1}{2} R(\widetilde{f}_{1,\mathcal{I}_k}; \mathcal{D}_n, \{I_2\}). \tag{44}
$$

We will now apply a strengthened version of dominated convergence theorem, called Pratt's lemma (see, e.g., Theorem 5.5 of Gut (2005) or Chapter 5 Exercise 30 of Resnick (2019)) to prove almost sure convergence of $\mathbb{E}[R(\widetilde{f}_{2,\mathcal{I}_k}; \mathcal{D}_n, \{I_\ell\}_{\ell=1}^2) \mid \mathcal{D}_n]$. Usually, Pratt's lemma is applied unconditionally, and here we apply it conditional on $\mathcal{D}_n$. For an easier understanding of the proof, let us write $\mathcal{D}_n(\omega)$ instead of $\mathcal{D}_n$ to make it clear that we are conditioning on $\mathcal{D}_n$. Recall that $\mathcal{D}_n$ is independent of the subsamples $\{I_\ell\}_{\ell=1}^M$ for any $M \geq 1$. In this notation, inequality (44) becomes

$$
0 \leq R(\widetilde{f}_{2,\mathcal{I}_k}; \mathcal{D}_n(\omega), \{I_\ell\}_{\ell=1}^2) \leq \frac{1}{2} R(\widetilde{f}_{1,\mathcal{I}_k}; \mathcal{D}_n(\omega), \{I_1\}) + \frac{1}{2} R(\widetilde{f}_{1,\mathcal{I}_k}; \mathcal{D}_n(\omega), \{I_2\}). \tag{45}
$$

Because $R(\widetilde{f}_{1,\mathcal{I}_k}; \mathcal{D}_n, \{I\}) \xrightarrow{\text{a.s.}} b_1$ for every $I \in \mathcal{I}_k$, there exists a set $\mathcal{A} \subseteq \Omega$ such that $\mathbb{P}(\mathcal{A}) = 1$ and for all $\omega \in \mathcal{A}$, $R(\widetilde{f}_{1,\mathcal{I}_k}; \mathcal{D}_n(\omega), \{I\}) \xrightarrow{\text{a.s.}} b_1$ for every $I \in \mathcal{I}_k$. Applying Pratt's lemma for every $\omega \in \mathcal{A}$, as $n \to \infty$ and using the fact (43) as well as the assumption $R(\widetilde{f}_{2,\mathcal{I}_k}; \mathcal{D}_n(\omega), \{I_\ell\}_{\ell=1}^2) \xrightarrow{\text{a.s.}} b_2$, we get that

$$
\mathbb{E}[R(\widetilde{f}_{2,\mathcal{I}_k}; \mathcal{D}_n(\omega), \{I_\ell\}_{\ell=1}^2)] \ \to \ b_2, \quad \text{for all} \quad \omega \in \mathcal{A}.
$$

Note that $R(\widetilde{f}_{2,\mathcal{I}_2}; \mathcal{D}_n(\omega)) = \mathbb{E}[R(\widetilde{f}_{2,\mathcal{I}_k}; \mathcal{D}_n, \{I_\ell\}_{\ell=1}^2) \mid \mathcal{D}_n = \mathcal{D}_n(\omega)]$. Therefore, we conclude

$$
R(\widetilde{f}_{2,\mathcal{I}_2}; \mathcal{D}_n) \xrightarrow{\text{a.s.}} b_2, \quad \text{as} \quad n \to \infty. \tag{46}
$$

Therefore, (9) applies to yield asymptotics for the data conditional risk uniformly over $M \in \mathbb{N}$. ■

B.4 Proof of Theorem 5 (from subsample conditional to data conditional risk, general M)

Proof The proof follows by combining Propositions 2 and 3, and Lemma 4. ■

B.5 Extension to general loss functions

Proposition 18 (Convex, strongly-convex, and smooth loss functions)**.** *For any loss function* $L : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$*, every* $(\boldsymbol{x}, y) \in \mathbb{R}^p \times \mathbb{R}$*, and for* $\widetilde{f}_M \in \{\widetilde{f}^{\texttt{WR}}_{M,\mathcal{I}_k}, \widetilde{f}^{\texttt{WOR}}_{M,\mathcal{I}_k}, \widetilde{f}^{\texttt{WR}}_{M,\mathcal{I}_k^\pi}, \widetilde{f}^{\texttt{WOR}}_{M,\mathcal{I}_k^\pi}\}$*, define*

$$R(\widetilde{f}_M; \mathcal{D}_n) = \int \mathbb{E}[L(y, \widetilde{f}_M(\boldsymbol{x}; \{\mathcal{D}_{I_\ell}\}_{\ell=1}^M)) \mid \mathcal{D}_n] \, \mathrm{d}P(\boldsymbol{x}, y).$$

If $L : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ *is convex in the second argument*[**]*, then* $R(\widetilde{f}_M, \mathcal{D}_n)$ *is non-increasing in* $M \geq 1$*, i.e.,* $R(\widetilde{f}_{M+1}; \mathcal{D}_n) \leq R(\widetilde{f}_M; \mathcal{D}_n)$*. Alternatively, if there exists* $\underline{m}, \overline{m} \in \mathbb{R}$ *such that* $L(\cdot, \cdot)$ *is* $\underline{m}$*-strongly convex and* $\overline{m}$*-smooth in the second argument*[††]*, then for* $\widetilde{f}_M \in \{\widetilde{f}^{\texttt{WR}}_{M,\mathcal{I}_k}, \widetilde{f}^{\texttt{WOR}}_{M,\mathcal{I}_k}\}$*,*

$$\frac{\underline{m} K_{|\mathcal{I}_k|,M}}{2M} \int \mathscr{V}_{\mathcal{I}_k}(\boldsymbol{x}, y) \, \mathrm{d}P(\boldsymbol{x}, y) \;\leq\; R(\widetilde{f}_M; \mathcal{D}_n) - R(\widetilde{f}_\infty; \mathcal{D}_n) \;\leq\; \frac{\overline{m} K_{|\mathcal{I}_k|,M}}{2M} \int \mathscr{V}_{\mathcal{I}_k}(\boldsymbol{x}, y) \, \mathrm{d}P(\boldsymbol{x}, y) \tag{47}$$

with $K_{N,M}$ *defined in* (8)*. The inequalities in* (47) *continue to hold for* $\widetilde{f}_M \in \{\widetilde{f}^{\texttt{WR}}_{M,\mathcal{I}_k^\pi}, \widetilde{f}^{\texttt{WOR}}_{M,\mathcal{I}_k^\pi}\}$*, with* $K_{|\mathcal{I}_k|,M}$ *and* $\mathscr{V}_{\mathcal{I}_k}$ *replaced with* $K_{|\mathcal{I}_k^\pi|,M}$ *and* $\mathscr{V}_{\mathcal{I}_k^\pi}$*, respectively.*

Proof We split the proof into two parts, depending on the assumption imposed on the loss function L.

Part (1). For any loss function $L : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ convex in the second argument, one can trivially obtain

$$\begin{aligned} R(\widetilde{f}_{M,\mathcal{I}_k}; \mathcal{D}_n) &= \mathbb{E}[L(y, \widetilde{f}_{M,\mathcal{I}_k}(\boldsymbol{x})) \mid \mathcal{D}_n] \\ &= \mathbb{E}[\mathbb{E}[L(y, \widetilde{f}_{M,\mathcal{I}_k}(\boldsymbol{x})) \mid \{I_\ell\}_{\ell=1}^M] \mid \mathcal{D}_n] \\ &\geq \mathbb{E}[L(y, \mathbb{E}[\widetilde{f}_{M,\mathcal{I}_k}(\boldsymbol{x}) \mid \{I_\ell\}_{\ell=1}^M]) \mid \mathcal{D}_n]. \end{aligned} \tag{48}$$

Here the last inequality follows from Jensen's inequality. Because $\mathbb{E}[\widetilde{f}_{M,\mathcal{I}_k}(\boldsymbol{x}) \mid \{I_\ell\}_{\ell=1}^M] = \widetilde{f}_{\infty,\mathcal{I}_k}(\boldsymbol{x})$, we get for any $M \geq 1$,

$$R(\widetilde{f}_{M,\mathcal{I}_k}; \mathcal{D}_n) \geq R(\widetilde{f}_{\infty,\mathcal{I}_k}; \mathcal{D}_n).$$

On the other hand, we have by Jensen's inequality

$$R(\widetilde{f}_{M,\mathcal{I}_k}; \mathcal{D}_n) = \mathbb{E}\left[L\left(y, \frac{1}{M}\sum_{\ell=1}^M \widetilde{f}(\boldsymbol{x}; \mathcal{D}_{I_\ell})\right) \,\Big|\, \mathcal{D}_n\right] \leq \mathbb{E}\left[\frac{1}{M}\sum_{\ell=1}^M L(y, \widetilde{f}(\boldsymbol{x}; \mathcal{D}_{I_\ell})) \,\Big|\, \mathcal{D}_n\right] = R(\widetilde{f}_{1,\mathcal{I}_k}; \mathcal{D}_n).$$

Summarizing, we get that for any $M \geq 1$,

$$R(\widetilde{f}_{1,\mathcal{I}_k}; \mathcal{D}_n) \geq R(\widetilde{f}_{M,\mathcal{I}_k}; \mathcal{D}_n) \geq R(\widetilde{f}_{\infty,\mathcal{I}_k}; \mathcal{D}_n).$$

[**]Recall that a function $f : \mathbb{R} \to \mathbb{R}$ is convex if $f(tx_1 + (1-t)x_2) \leq tf(x_1) + (1-t)f(x_2)$ for all $x_1, x_2 \in \mathbb{R}$ and $t \in [0, 1]$.

[††]A function $g : \mathbb{R} \to \mathbb{R}$ is said to be λ_1-strongly convex if $x \mapsto f(x) - \lambda_1/2x^2$ is convex. It is called a λ_2-smooth function if the derivative of f is λ_2-Lipschitz (i.e., $|f'(x_1) - f'(x_2)| \leq \lambda_2|x_1 - x_2|$ for all x_1, x_2).

One can further obtain the monotonicity property by noting that for any $M \geq 1$,

$$\widetilde{f}_{M+1,\mathcal{I}_k}(\boldsymbol{x}, \{\mathcal{D}_{I_\ell}\}_{\ell=1}^{M+1}) = \frac{1}{M+1}\sum_{\ell=1}^{M+1} \widetilde{f}(\boldsymbol{x}; \mathcal{D}_{I_\ell}) = \frac{1}{(M+1)!}\sum_{\pi'} \left(\frac{1}{M}\sum_{\ell=1}^{M} \widetilde{f}(\boldsymbol{x}; \mathcal{D}_{I_{\pi'(\ell)}})\right),$$

where π' represents a permutation of $\{1, 2, \dots, M+1\}$. Therefore, for any loss function $L : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ that is convex in the second argument, we get

$$L(y, \widetilde{f}_{M+1,\mathcal{I}_k}(\boldsymbol{x}; \{\mathcal{D}_{I_\ell}\}_{\ell=1}^{M+1})) \leq \frac{1}{(M+1)!}\sum_{\pi'} L\left(y, \widetilde{f}(\boldsymbol{x}; \{\mathcal{D}_{I_{\pi'(\ell)}}\}_{\ell=1}^{M})\right).$$

Because any (non-random) subset of a simple random sample with/without replacement is itself a simple random sample with/without replacement, taking conditional expectation on both sides conditional on $\mathcal{D}_n$ yields

$$R(\widetilde{f}_{M+1,\mathcal{I}_k}; \mathcal{D}_n) \leq R(\widetilde{f}_{M,\mathcal{I}_k}; \mathcal{D}_n).$$

This, in particular, implies that $R(\widetilde{f}_{\infty,\mathcal{I}_k}; \mathcal{D}_n) \leq R(\widetilde{f}_{M,\mathcal{I}_k}; \mathcal{D}_n) \leq R(\widetilde{f}_{1,\mathcal{I}_k}; \mathcal{D}_n)$ for any $M \geq 1$. This finishes the proof of the first part of the statement.

Part (2). If we assume that the loss function is strongly convex and differentiable in the second argument, then we can improve the lower bound of Part 1 in terms of $\widetilde{f}_\infty$. Formally, if $L : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ is $\underline{m}$-strongly convex, i.e., $L(a, b) - \underline{m}/2b^2$ is convex in b (for every a), then

$$L(y, \widetilde{f}_{M,\mathcal{I}_k}(\boldsymbol{x})) \geq L(y, \widetilde{f}_{\infty,\mathcal{I}_k}(\boldsymbol{x})) + \frac{\partial L(y, \widetilde{f}_{\infty,\mathcal{I}_k}(\boldsymbol{x}))}{\partial b}(\widetilde{f}_{M,\mathcal{I}_k}(\boldsymbol{x}) - \widetilde{f}_{\infty,\mathcal{I}_k}(\boldsymbol{x})) + \frac{\underline{m}}{2}(\widetilde{f}_{M,\mathcal{I}_k}(\boldsymbol{x}) - \widetilde{f}_{\infty,\mathcal{I}_k}(\boldsymbol{x}))^2.$$

Applying Proposition 1 and taking the expectation $(\boldsymbol{x}, y)$ conditional on $\mathcal{D}_n$, we obtain

$$R(\widetilde{f}_{M,\mathcal{I}_k}; \mathcal{D}_n) \geq R(\widetilde{f}_{\infty,\mathcal{I}_k}; \mathcal{D}_n) + \frac{\underline{m}}{2}\frac{1}{M}\int \frac{1}{|\mathcal{I}_k|}\sum_{I \in \mathcal{I}_k} (\widehat{f}(\boldsymbol{x}; \mathcal{D}_I) - \widetilde{f}_{\infty,\mathcal{I}_k}(\boldsymbol{x}))^2 \, \mathrm{d}P(\boldsymbol{x}, y). \tag{49}$$

On the other hand, if we assume that the loss function $L : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ is $\overline{m}$ smooth in the second argument, then

$$L(a, b) \leq L(a, b') + \frac{\partial L(a, b')}{\partial b}(b - b') + \frac{\overline{m}}{2}(b - b')^2.$$

It follows that

$$R(\widetilde{f}_{M,\mathcal{I}_k}; \mathcal{D}_n) \leq R(\widetilde{f}_{\infty,\mathcal{I}_k}; \mathcal{D}_n) + \frac{\overline{m}}{2}\frac{K_{|\mathcal{I}_k|,M}}{M}\int \sum_{I \in \mathcal{I}_k} (\widehat{f}(\boldsymbol{x}; \mathcal{D}_I) - \widetilde{f}_{\infty,\mathcal{I}_k}(\boldsymbol{x}))^2 \, \mathrm{d}P(\boldsymbol{x}, y). \tag{50}$$

Combining the lower bound from (49) and the upper bound from (50) finishes the proof of the second part of the statement. ■

Appendix C. Proof of Theorem 6 (subagging ridge with replacement)

For $\widetilde{f}^{\texttt{WR}}_{M,\mathcal{I}_k}$ defined in Theorem 6, we present the proof for ridge and ridgeless predictors in Theorems 19 and 23. For $\widetilde{f}^{\texttt{WOR}}_{M,\mathcal{I}_k}$ defined in Theorem 6, the conclusion still holds since the limits of the proportions of intersection between two SRSWR and SRSWOR draws from $\mathcal{I}_k$ are the same from Lemma 48. For proving the asymptotic conditional risks, we will treat $\mathcal{I}_k$ as fixed and use $\widetilde{f}^{\texttt{WR}}_{\lambda,M}$ to denote the ingredient predictor associated with regularization parameter λ.

C.1 Proof assembly

Before we present the proof, recall the nonnegative constants defined in (15) and (16): $v(-\lambda;\theta)\geq 0$ is the unique solution to the fixed-point equation

$$v(-\lambda;\theta)^{-1} = \lambda + \theta\int r(1+v(-\lambda;\theta)r)^{-1}\,\mathrm{d}H(r), \tag{51}$$

and the nonnegative constants $\widetilde{v}(-\lambda;\vartheta,\theta)$, and $\widetilde{c}(-\lambda;\theta)$ are defined via the following equations

$$\widetilde{v}(-\lambda,\vartheta,\theta) = \frac{\vartheta\int r^2(1+v(-\lambda;\theta)r)^{-2}\,\mathrm{d}H(r)}{v(-\lambda;\theta)^{-2}-\vartheta\int r^2(1+v(-\lambda;\theta)r)^{-2}\,\mathrm{d}H(r)},\ \widetilde{c}(-\lambda;\theta) = \int r(1+v(-\lambda;\theta))r)^{-2}\,\mathrm{d}G(r). \tag{52}$$

It helps to first slightly rewrite the statement of Theorem 6 for $\lambda>0$ as follows. Though it suffices to analyze the case $M=2$ according to Theorem 5, below we will do the risk decomposition for general M.

Theorem 19 (Risk characterization of subagged ridge predictor)**.** *Let $\widetilde{f}^{\texttt{WR}}_{\lambda,M}$ be the ingredient predictor as defined in* (14) *for $\lambda>0$. Suppose that Assumptions 1-5 hold, then for $M=\{1,2,3,\ldots\}$, as $k,n,p\to\infty$, $p/n\to\phi\in[0,\infty)$ and $p/k\to\phi_s\in[\phi,\infty]$, there exists a deterministic function $\mathscr{R}^{\texttt{sub}}_{\lambda,M}(\phi,\phi_s)$ such that for $I_1,\ldots,I_M \overset{\texttt{SRSWR}}{\sim} \mathcal{I}_k$,*

$$\sup_{M\in\mathbb{N}} |R(\widetilde{f}^{\texttt{WR}}_{\lambda,M};\mathcal{D}_n,\{I_\ell\}_{\ell=1}^M) - \mathscr{R}^{\texttt{sub}}_{\lambda,M}(\phi,\phi_s)| \xrightarrow{\mathrm{p}} 0,$$

and

$$\sup_{M\in\mathbb{N}} |R(\widetilde{f}^{\texttt{WR}}_{\lambda,M};\mathcal{D}_n) - \mathscr{R}^{\texttt{sub}}_{\lambda,M}(\phi,\phi_s)| \xrightarrow{\mathrm{a.s.}} 0.$$

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

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

where $\mathscr{B}^{\texttt{sub}}_{\lambda,M}(\phi,\phi_s) = M^{-1}B_\lambda(\phi,\phi_s)+(1-M^{-1})B_\lambda(\phi,\phi_s)$, and $\mathscr{V}^{\texttt{sub}}_{\lambda,M}(\phi,\phi_s) = M^{-1}V_\lambda(\phi_s,\phi_s)+(1-M^{-1})V_\lambda(\phi,\phi_s)$ with

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

where $\widetilde{v}(-\lambda;\vartheta,\theta)$ and $\widetilde{c}(-\lambda;\theta)$ are as defined in (52)*.*

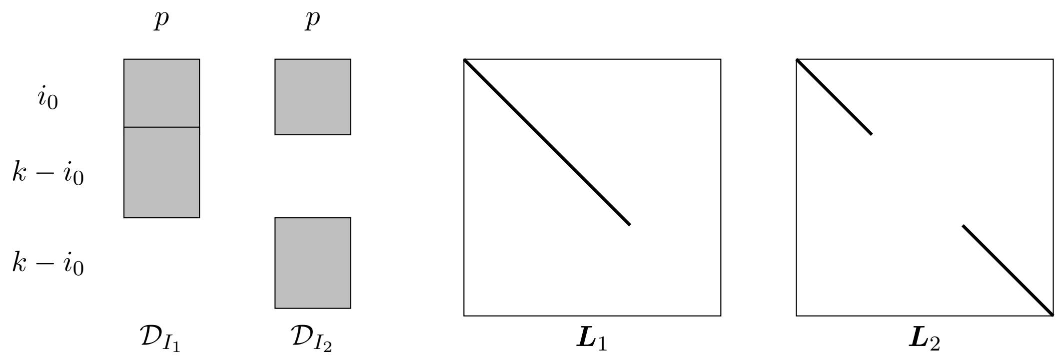

Figure 12: Illustration of subsampled datasets $\mathcal{D}_{I_1}$ and $\mathcal{D}_{I_2}$ from $\mathcal{D}_n$. The design matrix of each of them can be represented as $\boldsymbol{L}_j\boldsymbol{X}$ $(j = 1, 2)$, where $\boldsymbol{X} \in \mathbb{R}^{n\times p}$ is the full design matrix.

Proof In what follows, we will prove the results for n, k, p being a sequence of integers $\{n_m\}_{m=1}^{\infty}$, $\{k_m\}_{m=1}^{\infty}$, $\{p_m\}_{m=1}^{\infty}$. For simplicity, we drop the subscript when it is clear from the context.

For any $m \in [M]$, let I_m be a sample from $\mathcal{I}_k$, and $\boldsymbol{L}_m \in \mathbb{R}^{n\times n}$ be a diagonal matrix with $(\boldsymbol{L}_m)_{ll} = 1$ if $l \in I_m$ and 0 otherwise. An illustration of these notations for $M = 2$ is shown in Figure 12. The proof will reduce to analyze the individual terms concerning one dataset $\mathcal{D}_{I_m}$, or the cross terms concerning $\mathcal{D}_{I_m}$ and $\mathcal{D}_{I_l}$ for $m \neq l$.

The ingredient estimator takes the form:

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

Denote $\widetilde{\boldsymbol{\beta}}_{\lambda,M}(\{\mathcal{D}_{I_\ell}\}_{\ell=1}^{M})$ by $\widetilde{\boldsymbol{\beta}}_{\lambda,M}$ for simplicity. Let $\boldsymbol{M}_m = (\boldsymbol{X}^\top\boldsymbol{L}_m\boldsymbol{X}/k + \lambda\boldsymbol{I}_p)^{-1}$ for $m \in [M]$, we have

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

which yields

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

Thus, the conditional risk is given by

$$R(\widetilde{f}_{M,\lambda}; \mathcal{D}_n, \{I_\ell\}_{\ell=1}^{M}) = \mathbb{E}_{(\boldsymbol{x}_0,y_0)}[(y_0 - \boldsymbol{x}_0^\top\widetilde{\boldsymbol{\beta}}_{\lambda,M})^2]$$

$$
\begin{aligned}
&= \sigma^2 + (\boldsymbol{\beta}_0 - \widetilde{\boldsymbol{\beta}}_{\lambda,M})^\top \boldsymbol{\Sigma} (\boldsymbol{\beta}_0 - \widetilde{\boldsymbol{\beta}}_{\lambda,M}) \\
&= \sigma^2 + T_C + T_B + T_V,
\end{aligned}
$$

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

$$
T_C = -\frac{2\lambda}{M^2} \cdot \boldsymbol{\epsilon}^\top \left(\sum_{m=1}^{M} \boldsymbol{M}_m \frac{\boldsymbol{X}^\top \boldsymbol{L}_m}{k} \right)^\top \boldsymbol{\Sigma} \left(\sum_{m=1}^{M} \boldsymbol{M}_m \right) \boldsymbol{\beta}_0, \tag{53}
$$

$$
T_B = \frac{\lambda^2}{M^2} \cdot \boldsymbol{\beta}_0^\top \left(\sum_{m=1}^{M} \boldsymbol{M}_m \right) \boldsymbol{\Sigma} \left(\sum_{m=1}^{M} \boldsymbol{M}_m \right) \boldsymbol{\beta}_0, \tag{54}
$$

$$
T_V = \frac{1}{M^2} \cdot \boldsymbol{\epsilon}^\top \left(\sum_{m=1}^{M} \boldsymbol{M}_m \frac{\boldsymbol{X}^\top \boldsymbol{L}_m}{k} \right)^\top \boldsymbol{\Sigma} \left(\sum_{m=1}^{M} \boldsymbol{M}_m \frac{\boldsymbol{X}^\top \boldsymbol{L}_m}{k} \right) \boldsymbol{\epsilon}. \tag{55}
$$

Next we analyze the three terms separately for $M \in \{1, 2\}$. From Patil et al. (2022a, Lemmas S.2.2 and S.2.3), we have that $T_C \xrightarrow{\text{a.s.}} 0$, and

$$
\begin{aligned}
T_V &= \frac{1}{M^2} \sum_{m=1}^{M} \boldsymbol{\epsilon}^\top \boldsymbol{M}_m \frac{\boldsymbol{X}^\top \boldsymbol{L}_m}{k} \boldsymbol{\Sigma} \boldsymbol{M}_m \frac{\boldsymbol{X}^\top \boldsymbol{L}_m}{k} \boldsymbol{\epsilon} + \frac{1}{M^2} \sum_{m=1}^{M} \sum_{l=1}^{M} \boldsymbol{\epsilon}^\top \boldsymbol{M}_m \frac{\boldsymbol{X}^\top \boldsymbol{L}_m}{k} \boldsymbol{\Sigma} \boldsymbol{M}_l \frac{\boldsymbol{X}^\top \boldsymbol{L}_l}{k} \boldsymbol{\epsilon} \\
&\xrightarrow{\text{a.s.}} \frac{1}{M^2} \sum_{m=1}^{M} \frac{\sigma^2}{k} \operatorname{tr}(\boldsymbol{M}_m \widehat{\boldsymbol{\Sigma}}_m \boldsymbol{M}_m \boldsymbol{\Sigma}) + \frac{1}{M^2} \sum_{m \neq l} \frac{\sigma^2}{k^2} \operatorname{tr}(\boldsymbol{M}_l \boldsymbol{X}^\top \boldsymbol{L}_l \boldsymbol{L}_m \boldsymbol{X} \boldsymbol{M}_m \boldsymbol{\Sigma}) := T_V'.
\end{aligned}
$$

Thus, it remains to obtain the deterministic equivalent for the bias term T_B and the trace term T_V'. From Lemma 20 and Lemma 21, we have that for all $I_1 \in \mathcal{I}_k$ when $M = 1$ and for all $I_m, I_l \overset{\texttt{SRSWR}}{\sim} \mathcal{I}_k$ when $M = 2$, it holds that

$$
\begin{aligned}
T_B &= \frac{\lambda^2}{M^2} \sum_{m=1}^{M} \boldsymbol{\beta}_0^\top \boldsymbol{M}_m \boldsymbol{\Sigma} \boldsymbol{M}_m \boldsymbol{\beta}_0 + \frac{\lambda^2}{M^2} \sum_{m=1}^{M} \sum_{l=1}^{M} \boldsymbol{\beta}_0^\top \boldsymbol{M}_m \boldsymbol{\Sigma} \boldsymbol{M}_l \boldsymbol{\beta}_0 \\
&\xrightarrow{\text{a.s.}} \frac{\rho^2}{M} (1 + \widetilde{v}(-\lambda; \phi_s, \phi_s)) \widetilde{c}(-\lambda; \phi_s) + \frac{\rho^2 (M-1)}{M} (1 + \widetilde{v}(-\lambda; \phi, \phi_s)) \widetilde{c}(-\lambda; \phi_s) \\
T_V' &\xrightarrow{\text{a.s.}} \frac{\sigma^2}{M} \widetilde{v}(-\lambda; \phi_s, \phi_s) + \frac{\sigma^2 (M-1)}{M} \widetilde{v}(-\lambda; \phi, \phi_s),
\end{aligned}
$$

as $n, k, p \to \infty$, $p/n \to \phi \in (0, \infty)$, and $p/k \to \phi_s \in [\phi, \infty)$, where the nonnegative constants $\widetilde{v}(-\lambda; \phi, \phi_s)$ and $\widetilde{c}(-\lambda; \phi_s)$ are as defined in (52). Therefore, we have shown that for all $I \in \mathcal{I}_k$,

$$
R(\widetilde{f}_{\lambda,1}; \mathcal{D}_n, \{I\}) \xrightarrow{\text{a.s.}} \mathscr{R}^{\texttt{sub}}_{\lambda,1}(\phi, \phi_s),
$$

and for all $I_1, I_2 \overset{\texttt{SRSWR}}{\sim} \mathcal{I}_k$,

$$
R(\widetilde{f}_{\lambda,2}; \mathcal{D}_n, \{I_\ell\}_{\ell=1}^2) \xrightarrow{\text{a.s.}} \mathscr{R}^{\texttt{sub}}_{\lambda,2}(\phi, \phi_s),
$$

where

$$
\mathscr{R}^{\texttt{sub}}_{\lambda,M}(\phi, \phi_s) = \sigma^2 + \frac{1}{M} (B_\lambda(\phi_s, \phi_s) + V_\lambda(\phi_s, \phi_s)) + \frac{M-1}{M} (B_\lambda(\phi, \phi_s) + V_\lambda(\phi, \phi_s)),
$$

and the components are:

$$B_\lambda(\phi, \phi_s) = \rho^2(1 + \widetilde{v}(-\lambda; \phi, \phi_s))\widetilde{c}(-\lambda; \phi_s), \qquad V_\lambda(\phi, \phi_s) = \sigma^2\widetilde{v}(-\lambda; \phi, \phi_s).$$

The proof for the boundary case when $\phi_s = \infty$ follows from Proposition 22. Then, we have that the function $\mathscr{R}^{\mathtt{sub}}_{\lambda,M}(\phi, \phi_s)$ is continuous on $[\phi, \infty]$.

Finally, the risk expression for general M and the uniformity claim over $M \in \mathbb{N}$ follow from Theorem 5. ∎

C.2 Component deterministic approximations

C.2.1 Deterministic approximation of the bias functional

Lemma 20 (Deterministic approximation of the bias functional)**.** *Under Assumptions 1-5, for all $m \in [M]$ and $I_m \in \mathcal{I}_k$, let $\widehat{\boldsymbol{\Sigma}}_m = \boldsymbol{X}^\top \boldsymbol{L}_m \boldsymbol{X}/k$, $\boldsymbol{L}_m \in \mathbb{R}^{n\times n}$ be a diagonal matrix with $(\boldsymbol{L}_m)_{ll} = 1$ if $l \in I_m$ and 0 otherwise, and $\boldsymbol{M}_m = (\boldsymbol{X}^\top \boldsymbol{L}_m \boldsymbol{X}/k + \lambda \boldsymbol{I}_p)^{-1}$. Then, it holds that*

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

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

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

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

as $n, k, p \to \infty$, $p/n \to \phi \in (0, \infty)$, and $p/k \to \phi_s \in [\phi, \infty)$, where $\phi_0 = \phi_s^2/\phi$, T_B is as defined in (54)*, and the nonnegative constants $\widetilde{v}(-\lambda; \phi, \phi_s)$ and $\widetilde{c}(-\lambda; \phi_s)$ are as defined in* (52).

Proof From Lemma 38 (1) we have that for $m \in [M]$,

$$\lambda^2 \boldsymbol{M}_m \boldsymbol{\Sigma} \boldsymbol{M}_m \simeq (\widetilde{v}_b(-\lambda; \phi_s) + 1) \cdot (v(-\lambda; \phi_s)\boldsymbol{\Sigma} + \boldsymbol{I}_p)^{-1} \boldsymbol{\Sigma} (v(-\lambda; \phi_s)\boldsymbol{\Sigma} + \boldsymbol{I}_p)^{-1}. \tag{56}$$

By the definition of deterministic equivalent, we have

$$\begin{aligned} \lambda^2 \boldsymbol{\beta}_0^\top \boldsymbol{M}_m \boldsymbol{\Sigma} \boldsymbol{M}_m \boldsymbol{\beta}_0 &\xrightarrow{\text{a.s.}} \lim_{p\to\infty} (1 + \widetilde{v}_b(-\lambda; \phi_s)) \sum_{i=1}^p \frac{r_i}{(1 + r_i v(-\lambda; \phi_s))^2} (\boldsymbol{\beta}_0^\top w_i)^2 \\ &= \lim_{p\to\infty} \|\boldsymbol{\beta}_0\|_2^2 (1 + \widetilde{v}_b(-\lambda; \phi_s)) \int \frac{r}{(1 + v(-\lambda; \phi_s) r)^2} \, \mathrm{d}G_p(r) \\ &= \rho^2 (1 + \widetilde{v}_b(-\lambda; \phi_s)) \int \frac{r}{(1 + v(-\lambda; \phi_s) r)^2} \, \mathrm{d}G(r), \end{aligned} \tag{57}$$

where the last equality holds since G_p and G have compact supports and Assumptions 3 and 5.

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

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

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

$$\boldsymbol{M}_j = \left(\frac{i_0}{k}\boldsymbol{N}_0^{-1} + \frac{k-i_0}{k}\boldsymbol{N}_j^{-1}\right)^{-1}, \qquad j = 1, 2, \tag{58}$$

where the equalities hold because $\boldsymbol{N}_0$ is invertible when $\lambda > 0$. Conditioning on i_0, we will show a sequence of deterministic equivalents

$$\lambda^2 \boldsymbol{M}_1 \boldsymbol{\Sigma} \boldsymbol{M}_2 \overset{(a)}{\simeq} \lambda \boldsymbol{M}^{\mathrm{det}}_{\boldsymbol{N}_0, i_0} \boldsymbol{\Sigma} \boldsymbol{M}_2 \overset{(b)}{\simeq} \boldsymbol{M}^{\mathrm{det}}_{\boldsymbol{N}_0, i_0} \boldsymbol{\Sigma} \boldsymbol{M}^{\mathrm{det}}_{\boldsymbol{N}_0, i_0} \overset{(c)}{\simeq} (\lambda^2 \boldsymbol{M}_1 \boldsymbol{\Sigma} \boldsymbol{M}_2)^{\mathrm{det}}_{i_0},$$

where in each step, we consider randomness from $\boldsymbol{N}_1$, $\boldsymbol{N}_2$, $\boldsymbol{N}_0$, respectively, since they are independent of each other conditioning on i_0. The subscript of the deterministic equivalent indicates the dependence on the corresponding random variables, and we will specify each deterministic equivalent in the following proof.

When $i_0 = k$, we have $\boldsymbol{M}_1 = \boldsymbol{M}_2$ and the above asymptotic equation reduces to (56). We next prove the case when $i_0 < k$.

Part (a). Since $\boldsymbol{N}_1$ is independent of $\boldsymbol{N}_0$ conditioning on i_0, from Definition 34 and Lemma 39 (1) we have

$$\lambda \boldsymbol{M}_1 \simeq \boldsymbol{M}^{\mathrm{det}}_{\boldsymbol{N}_0, i_0} := \frac{k}{k - i_0}\left(v_1 \boldsymbol{\Sigma} + \boldsymbol{I}_p + \boldsymbol{C}_1\right)^{-1} \,\Big|\, i_0,$$

where $v_1 = v(-\lambda; \gamma_1, \boldsymbol{\Sigma}_{\boldsymbol{C}_1})$, $\boldsymbol{\Sigma}_{\boldsymbol{C}_1} = (\boldsymbol{I}_p + \boldsymbol{C}_1)^{-\frac{1}{2}} \boldsymbol{\Sigma} (\boldsymbol{I}_p + \boldsymbol{C}_1)^{-\frac{1}{2}}$, $\boldsymbol{C}_1 = i_0(\lambda(k - i_0))^{-1}\boldsymbol{N}_0^{-1}$, and $\gamma_1 = p/(k - i_0)$. Here the subscripts of v_1 and $\boldsymbol{C}_1$ are related to the aspect ratio γ_1. Because of the sub-multiplicativity of operator norm, we have

$$\|\boldsymbol{\Sigma} \boldsymbol{M}_2\|_{\mathrm{op}} \le \|\boldsymbol{\Sigma}\|_{\mathrm{op}} \|\boldsymbol{M}_2\|_{\mathrm{op}} \le \frac{r_{\max}}{\lambda}.$$

By Proposition 35 (2), we have $\lambda^2 \boldsymbol{M}_1 \boldsymbol{\Sigma} \boldsymbol{M}_2 \simeq \lambda \boldsymbol{M}^{\mathrm{det}}_{\boldsymbol{N}_0, i_0} \boldsymbol{\Sigma} \boldsymbol{M}_2 \mid i_0$.

Part (b). Analoguously, we have

$$\begin{aligned} \lambda \boldsymbol{M}^{\mathrm{det}}_{\boldsymbol{N}_0, i_0} \boldsymbol{\Sigma} \boldsymbol{M}_2 &\simeq \boldsymbol{M}^{\mathrm{det}}_{\boldsymbol{N}_0, i_0} \boldsymbol{\Sigma} \boldsymbol{M}^{\mathrm{det}}_{\boldsymbol{N}_0, i_0} \\ &\simeq \left(\frac{k}{k - i_0}\right)^2 \left(v_1 \boldsymbol{\Sigma} + \boldsymbol{I}_p + \boldsymbol{C}_1\right)^{-1} \boldsymbol{\Sigma} \left(v_1 \boldsymbol{\Sigma} + \boldsymbol{I}_p + \boldsymbol{C}_1\right)^{-1} \mid i_0, \end{aligned}$$

as $\left\|\boldsymbol{M}^{\mathrm{det}}_{\boldsymbol{N}_0, i_0}\right\|_{\mathrm{op}} \le 1$.

Part (c). As we have symmetrized the expression, we have

$$\lambda^2 \boldsymbol{M}_1 \boldsymbol{\Sigma} \boldsymbol{M}_2 \simeq \boldsymbol{M}^{\rm det}_{\boldsymbol{N}_0,i_0} \boldsymbol{\Sigma} \boldsymbol{M}^{\rm det}_{\boldsymbol{N}_0,i_0} = \frac{k^2}{i_0^2} \lambda^2 (\boldsymbol{N}_0^{-1} + \lambda \boldsymbol{C}_0)^{-1} \boldsymbol{\Sigma} (\boldsymbol{N}_0^{-1} + \lambda \boldsymbol{C}_0)^{-1} \mid i_0,$$

where $\boldsymbol{C}_0 = (k - i_0)/i_0 \cdot (v_1 \boldsymbol{\Sigma} + \boldsymbol{I}_p)$. Define $\boldsymbol{\Sigma}_{\boldsymbol{C}_0} = (\boldsymbol{I} + \boldsymbol{C}_0)^{-\frac{1}{2}} \boldsymbol{\Sigma} (\boldsymbol{I} + \boldsymbol{C}_0)^{-\frac{1}{2}}$. Conditioning on i_0, by Lemma 39 (1), we have

$$\begin{aligned}
\operatorname{tr}[\boldsymbol{\Sigma}_{\boldsymbol{C}_1} (v_1 \boldsymbol{\Sigma}_{\boldsymbol{C}_1} + \boldsymbol{I}_p)^{-1}] &= \operatorname{tr}[\boldsymbol{\Sigma} (v_1 \boldsymbol{\Sigma} + \boldsymbol{I}_p + \boldsymbol{C}_1)^{-1}] \\
&= \frac{\lambda(k - i_0)}{i_0} \operatorname{tr}\left[\boldsymbol{\Sigma} \left(\boldsymbol{N}_0^{-1} + \frac{\lambda(k - i_0)}{i_0} (v_1 \boldsymbol{\Sigma} + \boldsymbol{I}_p)\right)^{-1}\right] \\
&\overset{a.s.}{=} \frac{k - i_0}{i_0} \operatorname{tr}\left[\boldsymbol{\Sigma} \left(v_0 \boldsymbol{\Sigma} + \boldsymbol{I}_p + \frac{k - i_0}{i_0} (v_1 \boldsymbol{\Sigma} + \boldsymbol{I}_p)\right)^{-1}\right] \\
&= \operatorname{tr}\left[\boldsymbol{\Sigma} \left(\left(\frac{i_0}{k - i_0} v_0 + v_1\right) \boldsymbol{\Sigma} + \frac{k}{k - i_0} \boldsymbol{I}_p\right)^{-1}\right],
\end{aligned}$$

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

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

almost surely. Note that the solution (v_0, v_1) to the above equations is a pair of positive numbers and does not depend on samples. If (v_0, v_1) is a solution to the above system, then (v_1, v_0) is also a solution. Thus, any solution to the above equations must be unique. On the other hand, since $v_0 = v_1 = v(-\lambda; p/k)$ satisfies the above equations, it is the unique solution. By Lemma 45, we can replace $v(-\lambda; \gamma_1, \boldsymbol{\Sigma}_{\boldsymbol{C}_1})$ by the solution $v_0 = v_1 = v(-\lambda; p/k)$ of the above system, which does not depend on samples. Thus,

$$\lambda^2 \boldsymbol{M}_1 \boldsymbol{\Sigma} \boldsymbol{M}_2 \simeq \frac{k^2}{i_0^2} \lambda^2 (\boldsymbol{N}_0^{-1} + \lambda \boldsymbol{C}^*)^{-1} \boldsymbol{\Sigma} (\boldsymbol{N}_0^{-1} + \lambda \boldsymbol{C}^*)^{-1} \mid i_0, \tag{59}$$

where $\boldsymbol{C}^* = (k - i_0)/i_0 \cdot (v(-\lambda; p/k) \boldsymbol{\Sigma} + \boldsymbol{I}_p)$. By Lemma 39 (2), we have

$$\boldsymbol{M}^{\rm det}_{\boldsymbol{N}_0,i_0} \boldsymbol{\Sigma} \boldsymbol{M}^{\rm det}_{\boldsymbol{N}_0,i_0} \simeq (\lambda^2 \boldsymbol{M}_1 \boldsymbol{\Sigma} \boldsymbol{M}_2)^{\rm det}_{i_0} := \frac{k^2}{i_0^2} (\widetilde{v}_b(-\lambda; \gamma_0, \boldsymbol{C}^*) + 1) (v(-\lambda; \gamma_0, \boldsymbol{C}^*) \boldsymbol{\Sigma} + \boldsymbol{I}_p + \boldsymbol{C}^*)^{-2} \boldsymbol{\Sigma} \mid i_0, \tag{60}$$

where $\gamma_0 = p/i_0$, and $v(-\lambda; \gamma_0, \boldsymbol{C}^*)$ and $\widetilde{v}_b(-\lambda; \gamma_0, \boldsymbol{C}^*)$ are defined through the following equations:

$$\frac{1}{v(-\lambda; \gamma_0, \boldsymbol{C}^*)} = \lambda + \gamma_0 \operatorname{tr}[\boldsymbol{\Sigma} (v(-\lambda; \gamma_0, \boldsymbol{C}^*) \boldsymbol{\Sigma} + \boldsymbol{I}_p + \boldsymbol{C}^*)^{-1}]/p$$

$$\frac{1}{\widetilde{v}_b(-\lambda;\gamma_0,\boldsymbol{C}^*)} = \frac{\gamma_0 \operatorname{tr}[\boldsymbol{\Sigma}^2(v(-\lambda;\gamma_0,\boldsymbol{C}^*)\boldsymbol{\Sigma}+\boldsymbol{I}_p+\boldsymbol{C}^*)^{-2}]/p}{v(-\lambda;\gamma_0,\boldsymbol{C}^*)^{-2} - \gamma_0 \operatorname{tr}[\boldsymbol{\Sigma}^2(v(-\lambda;\gamma_0,\boldsymbol{C}^*)\boldsymbol{\Sigma}+\boldsymbol{I}_p+\boldsymbol{C}^*)^{-2}]/p}.$$

From Parts (a) to (c), we have shown that $\lambda^2\boldsymbol{M}_1\boldsymbol{\Sigma}\boldsymbol{M}_2 \simeq (\lambda^2\boldsymbol{M}_1\boldsymbol{\Sigma}\boldsymbol{M}_2)^{\det}_{i_0} \mid i_0$ for $i_0 < k$. Note that the above equivalence also holds for $i_0 = k$. That is, this holds for all $i_0 \in \{0,1,\cdots,k\}$. By Proposition 35 (1), we can obtain the unconditioned asymptotic equivalence $\boldsymbol{M}^{\det}_{\boldsymbol{N}_0,i_0}\boldsymbol{\Sigma}\boldsymbol{M}^{\det}_{\boldsymbol{N}_0,i_0} \simeq (\lambda^2\boldsymbol{M}_1\boldsymbol{\Sigma}\boldsymbol{M}_2)^{\det}_{i_0}$.

Note that from Lemma 42, $\widetilde{v}_b(-\lambda;\gamma)$ and $v(-\lambda;\gamma)$ are continuous on γ, and from Lemma 48, $i_0/k \xrightarrow{\text{a.s.}} \phi/\phi_s$, where $\phi_s \in (0,\infty)$ is the limiting ratio such that $p/k \to \phi_s$ as $k,p\to\infty$. We have

$$(\lambda^2\boldsymbol{M}_1\boldsymbol{\Sigma}\boldsymbol{M}_2)^{\det}_{i_0} \simeq \frac{\phi_s^2}{\phi^2}(\widetilde{v}_b(-\lambda;\phi_0,\boldsymbol{\Sigma}_{\boldsymbol{C}'})+1)(v(-\lambda;\phi_0,\boldsymbol{\Sigma}_{\boldsymbol{C}'})\boldsymbol{\Sigma}+\boldsymbol{I}_p+\boldsymbol{C}')^{-2}\boldsymbol{\Sigma},$$

where $\boldsymbol{C}' = (\phi_s-\phi)/\phi\cdot(v(-\lambda;\phi_s)\boldsymbol{\Sigma}+\boldsymbol{I}_p)$ and $\phi_0 = \phi_s^2/\phi$. Note that

$$\begin{aligned}
\frac{1}{v(-\lambda;\phi_0,\boldsymbol{\Sigma}_{\boldsymbol{C}'})} &= \lambda + \phi_0\int \frac{r}{1+rv(-\lambda;\phi_0,\boldsymbol{\Sigma}_{\boldsymbol{C}'})}\,\mathrm{d}H(r;\boldsymbol{\Sigma}_{\boldsymbol{C}'})\\
&= \lambda + \phi_s \lim_{p\to\infty}\operatorname{tr}\left[\boldsymbol{\Sigma}\left(\frac{\phi}{\phi_s}(v(-\lambda;\phi_0,\boldsymbol{\Sigma}_{\boldsymbol{C}'})\boldsymbol{\Sigma}+\boldsymbol{I}_p) + \left(1-\frac{\phi}{\phi_s}\right)(v(-\lambda;\phi_s)\boldsymbol{\Sigma}+\boldsymbol{I}_p)\right)^{-1}\right]/p\\
\frac{1}{v(-\lambda;\phi_s)} &= \lambda + \phi_s\lim_{p\to\infty}\operatorname{tr}\left[\boldsymbol{\Sigma}(v(-\lambda;\phi_s)\boldsymbol{\Sigma}+\boldsymbol{I}_p)^{-1}\right]/p.
\end{aligned}$$

We have

$$v(-\lambda;\phi_0,\boldsymbol{\Sigma}_{\boldsymbol{C}'}) = v(-\lambda;\phi_s) \tag{61}$$

is a solution to the first fixed-point equation. From Lemma 41 (2), this solution is also unique. Then, we have

$$\begin{aligned}
1+\widetilde{v}_b(-\lambda;\phi_0,\boldsymbol{\Sigma}_{\boldsymbol{C}'}) &= \lim_{p\to\infty}\frac{v(-\lambda;\phi_0,\boldsymbol{C}')^{-2}}{v(-\lambda;\phi_0,\boldsymbol{C}')^{-2} - \phi_0\operatorname{tr}[\boldsymbol{\Sigma}^2(v(-\lambda;\phi_0,\boldsymbol{C}')\boldsymbol{\Sigma}+\boldsymbol{I}_p+\boldsymbol{C}')^{-2}]/p}\\
&= \lim_{p\to\infty}\frac{v(-\lambda;\phi_s)^{-2}}{v(-\lambda;\phi_s)^{-2}-\phi\operatorname{tr}[\boldsymbol{\Sigma}^2(v(-\lambda;\phi_s)\boldsymbol{\Sigma}+\boldsymbol{I}_p)^{-2}]/p}\\
&= \frac{v(-\lambda;\phi_s)^{-2}}{v(-\lambda;\phi_s)^{-2}-\phi\int\frac{r^2}{(1+v(-\lambda;\phi_s)r)^2}\,\mathrm{d}H(r)} := 1+\widetilde{v}(-\lambda;\phi,\phi_s).
\end{aligned}$$

From Lemma 41 (4), we have that $1+\widetilde{v}(-\lambda;\phi,\phi_s) > 0$. To conclude, we have shown that

$$\lambda^2\boldsymbol{M}_1\boldsymbol{\Sigma}\boldsymbol{M}_2 \simeq (1+\widetilde{v}(-\lambda;\phi,\phi_s))\left(v(-\lambda;\phi_s)\boldsymbol{\Sigma}+\boldsymbol{I}_p\right)^{-2}\boldsymbol{\Sigma}. \tag{62}$$

By the definition of deterministic equivalent, we have

$$\begin{aligned}
\lambda^2\boldsymbol{\beta}_0^\top\boldsymbol{M}_1\boldsymbol{\Sigma}\boldsymbol{M}_2\boldsymbol{\beta}_0 &\xrightarrow{\text{a.s.}} \lim_{p\to\infty}\sum_{i=1}^p\frac{(1+\widetilde{v}(-\lambda;\phi,\phi_s))r_i}{(1+v(-\lambda;\phi_s)r_i)^2}(\boldsymbol{\beta}_0^\top w_i)^2\\
&= \lim_{p\to\infty}\|\boldsymbol{\beta}_0\|_2^2\int\frac{(1+\widetilde{v}(-\lambda;\phi,\phi_s))r}{(1+v(-\lambda;\phi_s)r)^2}\,\mathrm{d}G_p(r)
\end{aligned}$$

$$= \rho^2 \int \frac{(1+\widetilde{v}(-\lambda;\phi,\phi_s))r}{(1+v(-\lambda;\phi_s)r)^2}\,\mathrm{d}G(r), \tag{63}$$

where in the last line we used the fact that G_p and G have compact supports and Assumptions 3 and 5. The conclusion follows by combining (57) and (63). ■

C.2.2 Deterministic approximation of the variance functional

Lemma 21 (Deterministic approximation of the variance functional)**.** *Under Assumptions 1-5, for all $m \in [M]$ and $I_m \in \mathcal{I}_k$, let $\widehat{\boldsymbol{\Sigma}}_m = \boldsymbol{X}^\top \boldsymbol{L}_m \boldsymbol{X}/k$, $\boldsymbol{L}_m \in \mathbb{R}^{n\times n}$ be a diagonal matrix with $(\boldsymbol{L}_m)_{ll} = 1$ if $l \in I_m$ and 0 otherwise, and $\boldsymbol{M}_m = (\boldsymbol{X}^\top \boldsymbol{L}_m \boldsymbol{X}/k + \lambda \boldsymbol{I}_p)^{-1}$. Then, it holds that*

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

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

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

$$\frac{1}{k^2}\operatorname{tr}(\boldsymbol{M}_l \boldsymbol{X}^\top \boldsymbol{L}_l \boldsymbol{L}_m \boldsymbol{X} \boldsymbol{M}_m \boldsymbol{\Sigma}) \xrightarrow{\text{a.s.}} \widetilde{v}(-\lambda;\phi,\phi_s),$$

as $n, k, p \to \infty$, $p/n \to \phi \in (0,\infty)$, and $p/k \to \phi_s \in [\phi,\infty)$, where the nonnegative constant $\widetilde{v}(\lambda;\phi,\phi_s)$ is as defined in (52).

Proof From Lemma 38 (2), we have that for $j \in [M]$,

$$\boldsymbol{M}_j \widehat{\boldsymbol{\Sigma}}_j \boldsymbol{M}_j \boldsymbol{\Sigma} \simeq \widetilde{v}_v(-\lambda;\phi_s)(v(-\lambda;\phi_s)\boldsymbol{\Sigma} + \boldsymbol{I}_p)^{-2}\boldsymbol{\Sigma}^2. \tag{64}$$

By the trace rule Lemma 33 (4) , we have

$$\begin{aligned}
\frac{1}{k}\operatorname{tr}(\boldsymbol{M}_j \widehat{\boldsymbol{\Sigma}}_j \boldsymbol{M}_j \boldsymbol{\Sigma}) &\xrightarrow{\text{a.s.}} \lim_{p\to\infty} \frac{p}{k}\cdot\frac{1}{p}\operatorname{tr}(\widetilde{v}_v(-\lambda;\phi_s)(v(-\lambda;\phi_s)\boldsymbol{\Sigma}+\boldsymbol{I}_p)^{-2}\boldsymbol{\Sigma}^2)\\
&= \phi_s \widetilde{v}_v(-\lambda;\phi_s) \lim_{p\to\infty}\frac{1}{p}\sum_{i=1}^{p}\frac{r_i^2}{(v(-\lambda;\phi_s)r_i+1)^2}\\
&= \phi_s \widetilde{v}_v(-\lambda;\phi_s) \lim_{p\to\infty}\int \frac{r^2}{(v(-\lambda;\phi_s)r+1)^2}\,\mathrm{d}H_p(r)\\
&= \phi_s \widetilde{v}_v(-\lambda;\phi_s) \int \frac{r^2}{(v(-\lambda;\phi_s)r+1)^2}\,\mathrm{d}H(r), \qquad j = 1, 2,
\end{aligned} \tag{65}$$

where in the last line we used the fact that H_p and H have compact supports and Assumption 5.

For the cross term, it suffices to derive the deterministic equivalent of $\boldsymbol{M}_1 \widehat{\boldsymbol{\Sigma}}_0 \boldsymbol{M}_2 \boldsymbol{\Sigma}$ where $\widehat{\boldsymbol{\Sigma}}_0 = \boldsymbol{X}^\top \boldsymbol{L}_1 \boldsymbol{L}_2 \boldsymbol{X}/i_0$ and $i_0 = \operatorname{tr}(\boldsymbol{L}_1 \boldsymbol{L}_2)$. We again show a sequence of deterministic equivalents as in the proof for Lemma 20:

$$\boldsymbol{M}_1 \widehat{\boldsymbol{\Sigma}}_0 \boldsymbol{M}_2 \boldsymbol{\Sigma} \overset{(a)}{\simeq} \boldsymbol{M}^{\det}_{\boldsymbol{N}_0,i_0} \widehat{\boldsymbol{\Sigma}}_0 \boldsymbol{M}_2 \boldsymbol{\Sigma} \overset{(b)}{\simeq} \boldsymbol{M}^{\det}_{\boldsymbol{N}_0,i_0} \widehat{\boldsymbol{\Sigma}}_0 \boldsymbol{M}^{\det}_{\boldsymbol{N}_0,i_0} \boldsymbol{\Sigma} \overset{(c)}{\simeq} (\boldsymbol{M}_1 \widehat{\boldsymbol{\Sigma}}_0 \boldsymbol{M}_2 \boldsymbol{\Sigma})^{\det}_{i_0} \mid i_0.$$

When $i_0 = k$, this reduces to (64). We next show the case when $i_0 < k$.

Part (a). We use Lemma 39 (1) to obtain

$$\boldsymbol{M}_1 \simeq \boldsymbol{M}^{\mathrm{det}}_{\boldsymbol{N}_0,i_0} := \frac{k}{\lambda(k-i_0)} \left(v(-\lambda;\gamma_1,\boldsymbol{\Sigma}_{\boldsymbol{C}_1})\boldsymbol{\Sigma} + \boldsymbol{I}_p + \boldsymbol{C}_1\right)^{-1} \mid i_0 \tag{66}$$

where $\boldsymbol{\Sigma}_{\boldsymbol{C}_1} = (\boldsymbol{I}_p+\boldsymbol{C}_1)^{-\frac{1}{2}}\boldsymbol{\Sigma}(\boldsymbol{I}_p+\boldsymbol{C}_1)^{-\frac{1}{2}}$, $\boldsymbol{C}_1 = i_0(\lambda(k-i_0))^{-1}\boldsymbol{N}_0^{-1}$, and $\gamma_1 = p/(k-i_0)$. Let $\gamma_0 = p/i_0$. Note that conditioning on i_0, $\limsup \left\|\widehat{\boldsymbol{\Sigma}}_0\right\|_{\mathrm{op}} \le r_{\max}(1+\sqrt{\phi_0})^2$ almost surely as $i_0, p\to\infty$ and $\gamma_0 \to \phi_0 \in (0,\infty)$ from Bai and Silverstein (2010). Then $\widehat{\boldsymbol{\Sigma}}_0\boldsymbol{M}_2\boldsymbol{\Sigma}$ has bounded operator norm and we have $\boldsymbol{M}_1\widehat{\boldsymbol{\Sigma}}_0\boldsymbol{M}_2\boldsymbol{\Sigma} \simeq \boldsymbol{M}^{\mathrm{det}}_{\boldsymbol{N}_0,i_0}\widehat{\boldsymbol{\Sigma}}_0\boldsymbol{M}_2\boldsymbol{\Sigma} \mid i_0$ by Proposition 35 (2).

Part (b). Similarly, we have $\boldsymbol{M}_2 \simeq \boldsymbol{M}^{\mathrm{det}}_{\boldsymbol{N}_0,i_0} \mid i_0$ and $\boldsymbol{M}_1\widehat{\boldsymbol{\Sigma}}_0\boldsymbol{M}_2\boldsymbol{\Sigma} \simeq \boldsymbol{M}^{\mathrm{det}}_{\boldsymbol{N}_0,i_0}\widehat{\boldsymbol{\Sigma}}_0\boldsymbol{M}^{\mathrm{det}}_{\boldsymbol{N}_0,i_0}\boldsymbol{\Sigma} \mid i_0$.

Part (c). Note that

$$\begin{aligned}
\boldsymbol{M}^{\mathrm{det}}_{\boldsymbol{N}_0,i_0}\widehat{\boldsymbol{\Sigma}}_0\boldsymbol{M}^{\mathrm{det}}_{\boldsymbol{N}_0,i_0}\boldsymbol{\Sigma} &= \frac{k^2}{\lambda^2(k-i_0)^2}\left(v(-\lambda;\gamma_1,\boldsymbol{\Sigma}_{\boldsymbol{C}_1})\boldsymbol{\Sigma}+\boldsymbol{I}_p+\boldsymbol{C}_1\right)^{-1}\widehat{\boldsymbol{\Sigma}}_0\left(v(-\lambda;\gamma_1,\boldsymbol{\Sigma}_{\boldsymbol{C}_1})\boldsymbol{\Sigma}+\boldsymbol{I}_p+\boldsymbol{C}_1\right)^{-1}\boldsymbol{\Sigma}\\
&= \frac{k^2}{i_0^2}\left(\boldsymbol{N}_0^{-1}+\lambda\boldsymbol{C}_0\right)^{-1}\widehat{\boldsymbol{\Sigma}}_0\left(\boldsymbol{N}_0^{-1}+\lambda\boldsymbol{C}_0\right)^{-1}\boldsymbol{\Sigma},
\end{aligned}$$

where $\boldsymbol{C}_0 = (k-i_0)/i_0\cdot(v(-\lambda;\gamma_1,\boldsymbol{\Sigma}_{\boldsymbol{C}_1})\boldsymbol{\Sigma}+\boldsymbol{I}_p)$. Define $\boldsymbol{\Sigma}_{\boldsymbol{C}_0} = (\boldsymbol{I}+\boldsymbol{C}_0)^{-\frac{1}{2}}\boldsymbol{\Sigma}(\boldsymbol{I}+\boldsymbol{C}_0)^{-\frac{1}{2}}$. Conditioning on i_0, by Lemma 39 (1) we have

$$\begin{aligned}
\mathrm{tr}[\boldsymbol{\Sigma}_{\boldsymbol{C}_1}(v_1\boldsymbol{\Sigma}_{\boldsymbol{C}_1}+\boldsymbol{I}_p)^{-1}] &= \mathrm{tr}[\boldsymbol{\Sigma}(v_1\boldsymbol{\Sigma}+\boldsymbol{I}_p+\boldsymbol{C}_1)^{-1}]\\
&= \frac{\lambda(k-i_0)}{i_0}\,\mathrm{tr}\left[\boldsymbol{\Sigma}\left(\boldsymbol{N}_0^{-1}+\frac{\lambda(k-i_0)}{i_0}(v_1\boldsymbol{\Sigma}+\boldsymbol{I}_p)\right)^{-1}\right]\\
&\overset{a.s.}{=} \frac{k-i_0}{i_0}\,\mathrm{tr}\left[\boldsymbol{\Sigma}\left(v_0\boldsymbol{\Sigma}+\boldsymbol{I}_p+\frac{k-i_0}{i_0}(v_1\boldsymbol{\Sigma}+\boldsymbol{I}_p)\right)^{-1}\right]\\
&= \mathrm{tr}\left[\boldsymbol{\Sigma}\left(\left(\frac{i_0}{k-i_0}v_0+v_1\right)\boldsymbol{\Sigma}+\frac{k}{k-i_0}\boldsymbol{I}_p\right)^{-1}\right]
\end{aligned}$$

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

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

almost surely. By the same argument as in the proof for Lemma 20, we have that the solution $v_0 = v_1 = v(-\lambda;p/k)$ of the above system does not depend on samples and equals to $v(-\lambda;\gamma_1,\boldsymbol{\Sigma}_{\boldsymbol{C}_1})$ or $v(-\lambda;\gamma_0,\boldsymbol{\Sigma}_{\boldsymbol{C}_0})$ almost surely. Thus, by Lemma 45,

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

where $\boldsymbol{C}^* = (k - i_0)/i_0 \cdot (v(-\lambda; p/k)\boldsymbol{\Sigma} + \boldsymbol{I}_p)$. From Lemma 39 (3), we have

$$\boldsymbol{M}^{\mathrm{det}}_{\boldsymbol{N}_0,i_0}\widehat{\boldsymbol{\Sigma}}_0\boldsymbol{M}^{\mathrm{det}}_{\boldsymbol{N}_0,i_0}\boldsymbol{\Sigma} \simeq (\boldsymbol{M}_1\widehat{\boldsymbol{\Sigma}}_0\boldsymbol{M}_2\boldsymbol{\Sigma})^{\mathrm{det}}_{i_0} := \frac{k^2}{i_0^2}\widetilde{v}_v(-\lambda;\gamma_0,\boldsymbol{\Sigma}_{\boldsymbol{C}^*})(v(-\lambda;\gamma_0,\boldsymbol{\Sigma}_{\boldsymbol{C}^*})\boldsymbol{\Sigma} + \boldsymbol{I}_p + \boldsymbol{C}^*)^{-2}\boldsymbol{\Sigma}^2 \mid i_0,$$

where $\gamma_0 = p/i_0$.

From Parts (a) to (c), we have shown that $\boldsymbol{M}_1\widehat{\boldsymbol{\Sigma}}_0\boldsymbol{M}_2\boldsymbol{\Sigma} \simeq (\boldsymbol{M}_1\widehat{\boldsymbol{\Sigma}}_0\boldsymbol{M}_2\boldsymbol{\Sigma})^{\mathrm{det}}_{i_0} \mid i_0$ for $i_0 < k$. Note that this also holds for $i_0 = k$. Then by Proposition 35, $\boldsymbol{M}_1\widehat{\boldsymbol{\Sigma}}_0\boldsymbol{M}_2\boldsymbol{\Sigma} \simeq (\boldsymbol{M}_1\widehat{\boldsymbol{\Sigma}}_0\boldsymbol{M}_2\boldsymbol{\Sigma})^{\mathrm{det}}_{i_0}$.

Note that from Lemma 42, $\widetilde{v}_b(-\lambda;\gamma)$ and $v(-\lambda;\gamma)$ are continuous on γ, and from Lemma 48, $i_0/k \xrightarrow{\text{a.s.}} \phi/\phi_s$ where $\phi_s \in (0,\infty)$ is the limiting ratio such that $p/k \to \phi_s$ as $k,p\to\infty$. We have

$$\boldsymbol{M}_1\widehat{\boldsymbol{\Sigma}}_0\boldsymbol{M}_2\boldsymbol{\Sigma} \simeq \frac{\phi_s^2}{\phi^2}\widetilde{v}_v(-\lambda;\phi_0,\boldsymbol{\Sigma}_{\boldsymbol{C}'})(v(-\lambda;\phi_0,\boldsymbol{\Sigma}_{\boldsymbol{C}'})\boldsymbol{\Sigma} + \boldsymbol{I}_p + \boldsymbol{C}')^{-2}\boldsymbol{\Sigma}^2,$$

where $\phi_0 = \phi_s^2/\phi$, $\boldsymbol{\Sigma}_{\boldsymbol{C}'} = (\boldsymbol{I}_p + \boldsymbol{C}')^{-\frac{1}{2}}\boldsymbol{\Sigma}(\boldsymbol{I}_p + \boldsymbol{C}')^{-\frac{1}{2}}$, and $\boldsymbol{C}' = (\phi_s - \phi)/\phi \cdot (v(-\lambda;\phi_s)\boldsymbol{\Sigma} + \boldsymbol{I}_p)$. From (61), we have that $v(-\lambda;\phi_0;\boldsymbol{\Sigma}_{\boldsymbol{C}'}) = v(-\lambda;\phi_s)$, and

$$\begin{aligned}
\phi\widetilde{v}_v(-\lambda;\phi_0,\boldsymbol{\Sigma}_{\boldsymbol{C}'}) &= \lim_{p\to\infty}\frac{\phi}{v(-\lambda;\phi_0,\boldsymbol{C}')^{-2} - \phi_0\operatorname{tr}[\boldsymbol{\Sigma}^2(v(-\lambda;\phi_0,\boldsymbol{C}')\boldsymbol{\Sigma} + \boldsymbol{I}_p + \boldsymbol{C}')^{-2}]/p} \\
&= \lim_{p\to\infty}\frac{\phi}{v(-\lambda;\phi_s)^{-2} - \phi\operatorname{tr}[\boldsymbol{\Sigma}^2(v(-\lambda;\phi_s)\boldsymbol{\Sigma} + \boldsymbol{I}_p)^{-2}]/p} \\
&= \frac{\phi}{v(-\lambda;\phi_s)^{-2} - \phi\int\frac{r^2}{(1+v(-\lambda;\phi_s)r)^2}\,\mathrm{d}H(r)} := v_v(-\lambda;\phi,\phi_s).
\end{aligned}$$

From Lemma 41 (4), we have that $v_v(-\lambda;\phi,\phi_s) > 0$. Then we have

$$\boldsymbol{M}_1\widehat{\boldsymbol{\Sigma}}_0\boldsymbol{M}_2\boldsymbol{\Sigma} \simeq \phi^{-1}v_v(-\lambda;\phi,\phi_s)(v(-\lambda;\phi_s)\boldsymbol{\Sigma} + \boldsymbol{I}_p)^{-2}\boldsymbol{\Sigma}^2, \tag{67}$$

and thus, we have

$$\begin{aligned}
\frac{i_0}{k^2}\operatorname{tr}(\boldsymbol{M}_1\widehat{\boldsymbol{\Sigma}}_0\boldsymbol{M}_2\boldsymbol{\Sigma})) &\xrightarrow{\text{a.s.}} \lim_{p\to\infty}\frac{i_0 p}{k^2}\frac{1}{\phi}\cdot\frac{1}{p}\operatorname{tr}(v_v(-\lambda;\phi,\phi_s)(v(-\lambda;\phi_s)\boldsymbol{\Sigma} + \boldsymbol{I}_p)^{-2}\boldsymbol{\Sigma}^2) \\
&= \lim_{p\to\infty}\frac{1}{p}\sum_{i=1}^{p}\frac{v_v(-\lambda;\phi,\phi_s)r_i^2}{(1+v(-\lambda;\phi_s)r_i)^2} \\
&= \lim_{p\to\infty}\int\frac{v_v(-\lambda;\phi,\phi_s)r^2}{(1+v(-\lambda;\phi_s)r)^2}\,\mathrm{d}H_p(r) \\
&= \int\frac{v_v(-\lambda;\phi,\phi_s)r^2}{(1+v(-\lambda;\phi_s)r)^2}\,\mathrm{d}H(r) := \widetilde{v}(-\lambda;\phi,\phi_s),
\end{aligned} \tag{68}$$

where in the last line we used the fact that H_p and H have compact supports and Assumption 5. ■

C.3 Boundary case: diverging subsample aspect ratio

Proposition 22 (Risk approximation when $\phi_s \to +\infty$). *Under Assumptions 1-5, it holds for all $M \in \mathbb{N}$*

$$R(\widetilde{f}^{\mathtt{WR}}_{\lambda,M}; \mathcal{D}_n, \{I_\ell\}_{\ell=1}^M) \xrightarrow{\text{a.s.}} \mathscr{R}^{\mathtt{sub}}_{\lambda,M}(\phi, \infty),$$

as $k, n, p \to \infty$, $p/n \to \phi \in (0, \infty)$ and $p/k \to \infty$, where

$$\mathscr{R}^{\mathtt{sub}}_{\lambda,M}(\phi, \infty) := \lim_{\phi_s \to +\infty} \mathscr{R}^{\mathtt{sub}}_{\lambda,M}(\phi, \phi_s) = \sigma^2 + \rho^2 \int r \, \mathrm{d}G(r) \tag{69}$$

and $\mathscr{R}^{\mathtt{sub}}_{\lambda,M}(\phi, \phi_s)$ is defined in Theorem 19.

Proof Note that

$$\begin{aligned}
R(\widetilde{f}^{\mathtt{WR}}_{\lambda,M}; \mathcal{D}_n, \{I_\ell\}_{\ell=1}^M) &= \mathbb{E}_{(\boldsymbol{x}_0, y_0)}[(y_0 - \boldsymbol{x}_0^\top \widetilde{\boldsymbol{\beta}}_{\lambda,M}(\{\mathcal{D}_{I_\ell}\}_{\ell=1}^M))^2] \\
&= \mathbb{E}_{(\boldsymbol{x}_0, y_0)}[(\boldsymbol{\epsilon}_0 + \boldsymbol{x}_0^\top(\boldsymbol{\beta}_0 - \widetilde{\boldsymbol{\beta}}_{\lambda,M}(\{\mathcal{D}_{I_\ell}\}_{\ell=1}^M)))^2] \\
&= \sigma^2 + \mathbb{E}_{(\boldsymbol{x}_0, y_0)}[(\boldsymbol{\beta}_0 - \widetilde{\boldsymbol{\beta}}_{\lambda,M}(\{\mathcal{D}_{I_\ell}\}_{\ell=1}^M))^\top \boldsymbol{x}_0 \boldsymbol{x}_0^\top (\boldsymbol{\beta}_0 - \widetilde{\boldsymbol{\beta}}_{\lambda,M}(\{\mathcal{D}_{I_\ell}\}_{\ell=1}^M))] \\
&= \sigma^2 + (\boldsymbol{\beta}_0 - \widetilde{\boldsymbol{\beta}}_{\lambda,M}(\{\mathcal{D}_{I_\ell}\}_{\ell=1}^M))^\top \boldsymbol{\Sigma} (\boldsymbol{\beta}_0 - \widetilde{\boldsymbol{\beta}}_{\lambda,M}(\{\mathcal{D}_{I_\ell}\}_{\ell=1}^M)).
\end{aligned}$$

Then, by the Cauchy-Schwarz inequality, we have

$$\begin{aligned}
R(\widetilde{f}^{\mathtt{WR}}_{\lambda,M}; \mathcal{D}_n, \{I_\ell\}_{\ell=1}^M) - (\boldsymbol{\beta}_0^\top \boldsymbol{\Sigma} \boldsymbol{\beta}_0 + \sigma^2) &= \|\boldsymbol{\Sigma}^{\frac{1}{2}} \widetilde{\boldsymbol{\beta}}_{\lambda,M}(\{\mathcal{D}_{I_\ell}\}_{\ell=1}^M)\|_2^2 - 2\widetilde{\boldsymbol{\beta}}_{\lambda,M}(\{\mathcal{D}_{I_\ell}\}_{\ell=1}^M)^\top \boldsymbol{\Sigma} \boldsymbol{\beta}_0 \\
&\leq \frac{1}{r_{\min}} \|\widetilde{\boldsymbol{\beta}}_{\lambda,M}(\{\mathcal{D}_{I_\ell}\}_{\ell=1}^M)\|_2^2 + 2\|\widetilde{\boldsymbol{\beta}}_{\lambda,M}(\{\mathcal{D}_{I_\ell}\}_{\ell=1}^M)\|_2 \|\boldsymbol{\Sigma}\|_2 \\
&\leq \frac{1}{r_{\min}} \|\widetilde{\boldsymbol{\beta}}_{\lambda,M}(\{\mathcal{D}_{I_\ell}\}_{\ell=1}^M)\|_2^2 + 2r_{\max}\rho \|\widetilde{\boldsymbol{\beta}}_{\lambda,M}(\{\mathcal{D}_{I_\ell}\}_{\ell=1}^M)\|_2,
\end{aligned}$$

almost surely as $k, n, p \to$ and $p/k \to \infty$. Thus, we have the following holds almost surely:

$$\begin{aligned}
\|\widetilde{\boldsymbol{\beta}}_{\lambda,M}(\{\mathcal{D}_{I_\ell}\}_{\ell=1}^M)(\{\mathcal{D}_{I_\ell}\}_{\ell=1}^M)\|_2 &\leq \frac{1}{M} \sum_{m=1}^M \|(\boldsymbol{X}^\top \boldsymbol{L}_m \boldsymbol{X}/k + \lambda \boldsymbol{I}_p)^{-1} (\boldsymbol{X}^\top \boldsymbol{L}_m \boldsymbol{y}/k)\|_2 \\
&\leq \frac{1}{M} \sum_{m=1}^M \|(\boldsymbol{X}^\top \boldsymbol{L}_m \boldsymbol{X}/k + \lambda \boldsymbol{I}_p)^{-1} \boldsymbol{X}^\top \boldsymbol{L}_m/\sqrt{k}\| \cdot \|\boldsymbol{L}_m \boldsymbol{y}/\sqrt{k}\|_2 \\
&\leq C\sqrt{\rho^2 + \sigma^2} \cdot \frac{1}{M} \sum_{m=1}^M \|(\boldsymbol{X}^\top \boldsymbol{L}_m \boldsymbol{X}/k + \lambda \boldsymbol{I}_p)^{-1} \boldsymbol{X}^\top \boldsymbol{L}_m/\sqrt{k}\|,
\end{aligned}$$

where the last inequality holds eventually almost surely since Assumptions 1-3 imply that the entries of $\boldsymbol{y}$ have bounded 4-th moment, and thus from the strong law of large numbers, $\|\boldsymbol{L}_m \boldsymbol{y}/\sqrt{k}\|_2$ is eventually almost surely bounded above by $C\sqrt{\mathbb{E}[y_1^2]} = C\sqrt{\rho^2 + \sigma^2}$ for some constant C. Observe that operator norm of the matrix $(\boldsymbol{X}^\top \boldsymbol{L}_m \boldsymbol{X}/k + \lambda \boldsymbol{I}_p)^{-1} \boldsymbol{X} \boldsymbol{L}_m/\sqrt{k}$ is upper bounded $\max_i s_i/(s_i^2 + \lambda) \leq 1/s_{\min}$ 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 in Bloemendal et al. (2016)) and therefore, $\|\widetilde{\boldsymbol{\beta}}_{\lambda,M}(\{\mathcal{D}_{I_\ell}\}_{\ell=1}^M)\|_2 \to 0$ almost surely. Thus, we have shown that

$$R(\widetilde{f}^{\mathtt{WR}}_{\lambda,M}; \mathcal{D}_n, \{I_\ell\}_{\ell=1}^M) \xrightarrow{\text{a.s.}} \sigma^2 + \boldsymbol{\beta}_0^\top \boldsymbol{\Sigma} \boldsymbol{\beta}_0.$$

or equivalently

$$R(\widetilde{f}^{\texttt{WR}}_{\lambda,M}; \mathcal{D}_n, \{I_\ell\}_{\ell=1}^M) \xrightarrow{\text{a.s.}} \sigma^2 + \rho^2 \int r \, \mathrm{d}G(r).$$

From Lemma 42, we have

$$\lim_{\phi_s \to +\infty} v(-\lambda; \phi_s) = \lim_{\phi_s \to +\infty} \widetilde{v}_b(-\lambda; \phi_s) = \lim_{\phi_s \to +\infty} \widetilde{v}_v(-\lambda; \phi_s).$$

Thus,

$$\lim_{\phi_s \to +\infty} V_\lambda(\phi, \phi_s) = \lim_{\phi_s \to +\infty} V_\lambda(\phi, \phi_s) = 0$$

and

$$\lim_{\phi_s \to +\infty} B_\lambda(\phi, \phi_s) = \lim_{\phi_s \to +\infty} B_\lambda(\phi, \phi_s) = \rho^2 \int r \, \mathrm{d}G(r).$$

Therefore, we have $\mathscr{R}^{\texttt{sub}}_{\lambda,M}(\phi, \infty) := \lim_{\phi_s \to +\infty} \mathscr{R}^{\texttt{sub}}_{\lambda,M}(\phi, \phi_s) = \sigma^2 + \rho^2 \int r \, \mathrm{d}G(r)$. Thus, $\mathscr{R}^{\texttt{sub}}_{\lambda,M}(\phi, \infty)$ is well defined and $\mathscr{R}^{\texttt{sub}}_{\lambda,M}(\phi, \phi_s)$ is right continuous at $\phi_s = +\infty$. ■

Appendix D. Proof of Theorem 6 (subagging ridgeless with replacement)

As done in Section C, for proving the asymptotic conditional risks, we will treat $\mathcal{I}_k$ or $\mathcal{I}_k^\sigma$ as fixed. We will use $\widetilde{f}^{\texttt{WR}}_{0,M}$ to denote the ingredient predictor associated with the ridge penalty $\lambda = 0$.

D.1 Proof assembly

We first explicitly write out the statement of Theorem 6 for the ridgeless case of $\lambda = 0$. As in Section C, we obtain the risk decomposition for general M though it suffices to analyze the case $M = 2$ according to Theorem 5.

For ridgeless predictors ($\lambda = 0$) and $\theta > 1$, the scalar $v(0;\theta)$ is the unique fixed-point solution to the following equation:

$$v(0;\theta)^{-1} = \theta \int r(1 + v(0;\theta) r)^{-1} \, \mathrm{d}H(r). \tag{70}$$

and the nonnegative constants $\widetilde{v}(0;\vartheta,\theta)$ and $\widetilde{c}(0;\theta)$ are defined via the following equations:

$$\widetilde{v}(0;\vartheta,\theta) = \frac{\vartheta \int r^2 (1 + v(0;\theta) r)^{-2} \, \mathrm{d}H(r)}{v(0;\theta)^{-2} - \vartheta \int r^2 (1 + v(0;\theta) r)^{-2} \, \mathrm{d}H(r)}, \qquad \widetilde{c}(0;\theta) = \int r(1 + v(0;\theta) r)^{-2} \, \mathrm{d}G(r). \tag{71}$$

When $\theta \leq 1$, the quantities defined in (70) and (71) are interpreted as $\lim_{\lambda \to 0^+} v(-\lambda;\theta) = \infty$, $\lim_{\lambda \to 0^+} \widetilde{c}(-\lambda;\theta) = 0$ and $\lim_{\lambda \to 0^+} \widetilde{v}(-\lambda;\vartheta,\theta) = \vartheta(1-\vartheta)^{-1}$.

Theorem 23 (Risk characterization of subagged ridgeless predictor). *Let $\widetilde{f}^{\texttt{WR}}_{0,M}$ be the ingredient predictor as defined in* (14)*. Suppose Assumptions 1-5 hold for the dataset $\mathcal{D}_n$. Then,* *as $k, n, p \to \infty$ such that $p/n \to \phi \in (0, \infty)$ and $p/k \to \phi_s \in [\phi, \infty]$ and $\phi_s \neq 1$, there exists a deterministic function $\mathscr{R}^{\mathtt{sub}}_{0,M}(\phi, \phi_s)$, $M \in \mathbb{N}$, such that for $I_1, \ldots, I_M \overset{\mathtt{SRSWR}}{\sim} \mathcal{I}_k$,*

$$\sup_{M\in\mathbb{N}} |R(\widetilde{f}^{\mathtt{WR}}_{0,M}; \{\mathcal{D}_{I_\ell}\}_{\ell=1}^M) - \mathscr{R}^{\mathtt{sub}}_{0,M}(\phi, \phi_s)| \xrightarrow{\text{p}} 0,$$

and

$$\sup_{M\in\mathbb{N}} |R(\widetilde{f}_{0,M}; \mathcal{D}_n) - \mathscr{R}^{\mathtt{sub}}_{0,M}(\phi, \phi_s)| \xrightarrow{\text{a.s.}} 0.$$

Furthermore, the function $\mathscr{R}^{\mathtt{sub}}_{0,M}(\phi, \phi_s)$ decomposes as $\mathscr{R}^{\mathtt{sub}}_{0,M}(\phi, \phi_s) = \sigma^2 + \mathscr{B}^{\mathtt{sub}}_{0,M}(\phi, \phi_s) + \mathscr{V}^{\mathtt{sub}}_{0,M}(\phi, \phi_s)$, where the terms are given by $\mathscr{B}^{\mathtt{sub}}_{0,M}(\phi, \phi_s) = M^{-1}B_0(\phi_s, \phi_s) + (1-M^{-1})B_0(\phi, \phi_s)$, and $\mathscr{V}^{\mathtt{sub}}_{0,M}(\phi, \phi_s) = M^{-1}V_0(\phi_s, \phi_s) + (1-M^{-1})V_0(\phi, \phi_s)$, and the functions $B_0(\cdot, \cdot)$ and $V_0(\cdot, \cdot)$ are defined as

$$B_0(\vartheta, \theta) = \begin{cases} 0 & \theta \in (0,1),\ \vartheta \le \theta \\ \rho^2(1 + \widetilde{v}(0; \vartheta, \theta)\widetilde{c}(0; \theta) & \theta \in (1, \infty],\ \vartheta \le \theta \end{cases}, \quad V_0(\vartheta, \theta) = \begin{cases} \sigma^2 \dfrac{\vartheta}{1-\vartheta} & \theta \in (0,1),\ \vartheta \le \theta \\ \sigma^2 \widetilde{v}(0; \vartheta, \theta) & \theta \in (1, \infty],\ \vartheta \le \theta \end{cases},$$

where the nonnegative constants $\widetilde{v}(0; \vartheta, \theta)$ and $\widetilde{c}(0; \theta)$ are as defined in (71).

Proof We use the same notations as in the proof for Theorem 19 and let $\widehat{\boldsymbol{\Sigma}}_m = \boldsymbol{X}^\top \boldsymbol{L}_m \boldsymbol{X}/k$ for all $m \in [M]$. Note that

$$\boldsymbol{\beta}_0 - \widetilde{\boldsymbol{\beta}}(\{\mathcal{D}_{I_\ell}\}_{\ell=1}^M) = \frac{1}{M}\sum_{m=1}^M (\boldsymbol{I}_p - \widehat{\boldsymbol{\Sigma}}_m^+ \widehat{\boldsymbol{\Sigma}}_m)\boldsymbol{\beta}_0 - \frac{1}{M}\sum_{m=1}^M \widehat{\boldsymbol{\Sigma}}_m^+ \frac{\boldsymbol{X}^\top \boldsymbol{L}_m \boldsymbol{\epsilon}}{k}.$$

We have

$$\begin{aligned} R(\widetilde{f}^{\mathtt{WR}}_{0,M}; \mathcal{D}_n, \{I_\ell\}_{\ell=1}^M) &= \sigma^2 + (\boldsymbol{\beta}_0 - \widetilde{\boldsymbol{\beta}}(\{\mathcal{D}_{I_\ell}\}_{\ell=1}^M))^\top \boldsymbol{\Sigma} (\boldsymbol{\beta}_0 - \widetilde{\boldsymbol{\beta}}(\{\mathcal{D}_{I_\ell}\}_{\ell=1}^M)) \\ &= \sigma^2 + T_B + T_V + T_C, \end{aligned}$$

where

$$T_C = -\frac{2}{M^2}\boldsymbol{\epsilon}^\top \left(\sum_{m=1}^M \widehat{\boldsymbol{\Sigma}}_m^+ \frac{\boldsymbol{X}^\top \boldsymbol{L}_m}{k}\right)^\top \boldsymbol{\Sigma} \left(\sum_{m=1}^M (\boldsymbol{I}_p - \widehat{\boldsymbol{\Sigma}}_m^+ \widehat{\boldsymbol{\Sigma}}_m)\right)\boldsymbol{\beta}_0, \tag{72}$$

$$T_B = \frac{1}{M^2}\boldsymbol{\beta}_0^\top \left(\sum_{m=1}^M (\boldsymbol{I}_p - \widehat{\boldsymbol{\Sigma}}_m^+ \widehat{\boldsymbol{\Sigma}}_m)\right) \boldsymbol{\Sigma} \left(\sum_{m=1}^M (\boldsymbol{I}_p - \widehat{\boldsymbol{\Sigma}}_m^+ \widehat{\boldsymbol{\Sigma}}_m)\right)\boldsymbol{\beta}_0, \tag{73}$$

$$T_V = \frac{1}{M^2}\boldsymbol{\epsilon}^\top \left(\sum_{m=1}^M \widehat{\boldsymbol{\Sigma}}_m^+ \frac{\boldsymbol{X}^\top \boldsymbol{L}_m}{k}\right)^\top \boldsymbol{\Sigma} \left(\sum_{m=1}^M \widehat{\boldsymbol{\Sigma}}_m^+ \frac{\boldsymbol{X}^\top \boldsymbol{L}_m}{k}\right)\boldsymbol{\epsilon}. \tag{74}$$

Next we analyze the three term separately for $M \in \{1, 2\}$. From Patil et al. (2022a, Lemma S.3.2), we have that $T_C \xrightarrow{\text{a.s.}} 0$. Further, from Lemma 24, Lemma 25, and Lemma 26, for all $I_1 \in \mathcal{I}_k$ when $M = 1$ and for all $I_m, I_l \overset{\mathtt{SRSWR}}{\sim} \mathcal{I}_k$ when $M = 2$, it holds that

$$R(\widetilde{f}_{M,\lambda}; \{\mathcal{D}_{I_\ell}\}_{\ell=1}^M) \xrightarrow{\text{a.s.}} \mathscr{R}^{\mathtt{sub}}_{0,M}(\phi, \phi_s)$$

as $n, k, p \to \infty$, $p/n \to \phi \in (0,\infty)$, and $p/k \to \phi_s \in [\phi,\infty) \setminus \{1\}$, where

$$\mathscr{R}^{\texttt{sub}}_{0,M}(\phi,\phi_s) = \sigma^2 + \frac{1}{M}(B_0(\phi_s,\phi_s) + V_0(\phi_s,\phi_s)) + \frac{M-1}{M}(B_0(\phi,\phi_s) + V_0(\phi,\phi_s)).$$

Here, the components are:

$$B_0(\phi,\phi_s) = \begin{cases} 0, & \phi_s \in (0,1) \\ \rho^2(1+\widetilde{v}(0;\phi,\phi_s))\widetilde{c}(0;\phi_s), & \phi_s \in (1,\infty) \end{cases}, \quad V_0(\phi,\phi_s) = \begin{cases} \sigma^2 \frac{\phi}{1-\phi}, & \phi_s \in (0,1) \\ \sigma^2 \widetilde{v}(0;\phi,\phi_s), & \phi_s \in (1,\infty) \end{cases},$$

and the nonnegative constants $\widetilde{v}(0;\phi,\phi_s)$ and $\widetilde{c}(0;\phi_s)$ are as defined in (71). The proof for the boundary case when $\phi_s = \infty$ follows from Proposition 27. Then, we have that the function $\mathscr{R}^{\texttt{sub}}_{0,M}(\phi,\phi_s)$ is continuous on $[\phi,\infty] \setminus \{1\}$ and lower-semi continuous on $[\phi,\infty]$.

Finally, the risk expression for general M and the uniformity claim over $M \in \mathbb{N}$ follow from Theorem 5. ■

D.2 Component deterministic approximations

D.2.1 Deterministic approximation of the bias functional

Lemma 24 (Deterministic approximation of the bias functional)**.** *Under Assumptions 1-5, for all $m \in [M]$ and $I_m \in \mathcal{I}_k$, let $\widehat{\mathbf{\Sigma}}_m = \boldsymbol{X}^\top \boldsymbol{L}_m \boldsymbol{X}/k$ and $\boldsymbol{L}_m \in \mathbb{R}^{n\times n}$ be a diagonal matrix with $(\boldsymbol{L}_m)_{ll} = 1$ if $l \in I_m$ and 0 otherwise. Then, it holds that*

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

$$\boldsymbol{\beta}_0^\top(\boldsymbol{I}_p - \widehat{\mathbf{\Sigma}}_m^+\widehat{\mathbf{\Sigma}}_m)\mathbf{\Sigma}(\boldsymbol{I}_p - \widehat{\mathbf{\Sigma}}_m^+\widehat{\mathbf{\Sigma}}_m)\boldsymbol{\beta}_0 \xrightarrow{\text{a.s.}} \begin{cases} 0 & \phi_s \in (0,1) \\ \rho^2(1+\widetilde{v}(0;\phi_s,\phi_s))\widetilde{c}(0;\phi_s) & \phi_s \in (1,\infty), \end{cases}$$

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

$$\boldsymbol{\beta}_0^\top(\boldsymbol{I}_p - \widehat{\mathbf{\Sigma}}_m^+\widehat{\mathbf{\Sigma}}_m)\mathbf{\Sigma}(\boldsymbol{I}_p - \widehat{\mathbf{\Sigma}}_l^+\widehat{\mathbf{\Sigma}}_l)\boldsymbol{\beta}_0 \xrightarrow{\text{a.s.}} \begin{cases} 0 & \phi_s \in (0,1) \\ \rho^2(1+\widetilde{v}(0;\phi,\phi_s))\widetilde{c}(0;\phi_s) & \phi_s \in (1,\infty), \end{cases}$$

as $n, k, p \to \infty$, $p/n \to \phi \in (0,\infty)$, and $p/k \to \phi_s \in [\phi,\infty) \setminus \{1\}$, where the nonnegative constants $\widetilde{v}(0;\phi,\phi_s)$ and $\widetilde{c}(0;\phi_s)$ are as defined in (71).

Proof For the first term, we have that for $m \in [M]$,

$$\boldsymbol{\beta}_0^\top(\boldsymbol{I}_p - \widehat{\mathbf{\Sigma}}_m^+\widehat{\mathbf{\Sigma}}_m)\mathbf{\Sigma}(\boldsymbol{I}_p - \widehat{\mathbf{\Sigma}}_m^+\widehat{\mathbf{\Sigma}}_m)\boldsymbol{\beta}_0 \xrightarrow{\text{a.s.}} \begin{cases} 0 & \text{if } \phi_s \in (0,1) \\ \rho^2(1+\widetilde{v}_b(0;\phi_s)) \int \frac{r}{(1+v(0;\phi_s)r)^2}\,\mathrm{d}G(r) & \text{if } \phi_s \in (1,\infty). \end{cases} \tag{75}$$

Next we analyze the second term, by considering the following two cases separately for $(m,l) = (1,2)$.

(1) $\phi_s \in (0,1)$. Since the singular values of $\widehat{\mathbf{\Sigma}}_j$'s are almost surely lower bounded away from 0, we have $\widehat{\mathbf{\Sigma}}_j^+\widehat{\mathbf{\Sigma}}_j = \boldsymbol{I}_p$ almost surely. Then $\boldsymbol{\beta}_0^\top(\boldsymbol{I}_p - \widehat{\mathbf{\Sigma}}_1^+\widehat{\mathbf{\Sigma}}_1)\mathbf{\Sigma}(\boldsymbol{I}_p - \widehat{\mathbf{\Sigma}}_2^+\widehat{\mathbf{\Sigma}}_2)\boldsymbol{\beta}_0 \xrightarrow{\text{a.s.}} 0$ when $k, p \to \infty$ and $p/k \to \phi_s \in (0,1)$.
(2) $\phi_s \in (1,\infty)$. We begin with analyzing the deterministic equivalent of $(\boldsymbol{I}_p - \widehat{\mathbf{\Sigma}}_1^+\widehat{\mathbf{\Sigma}}_1)\mathbf{\Sigma}(\boldsymbol{I}_p - \widehat{\mathbf{\Sigma}}_2^+\widehat{\mathbf{\Sigma}}_2)$. Recall that i_0 is the number of shared samples between $\mathcal{D}_{I_1}$ and $\mathcal{D}_{I_2}$, and $\widehat{\mathbf{\Sigma}}_0 = \boldsymbol{X}^\top \boldsymbol{L}_1\boldsymbol{L}_2\boldsymbol{X}^\top / i_0$ and $\widehat{\mathbf{\Sigma}}_j^{\text{ind}} = \boldsymbol{X}^\top(\boldsymbol{L}_j - \boldsymbol{L}_1\boldsymbol{L}_2)\boldsymbol{X}^\top/(k-i_0)$ are the common and individual covariance estimators of the two datasets. Also note that from (62), we have $\lambda^2\boldsymbol{M}_1\mathbf{\Sigma}\boldsymbol{M}_2 \simeq (\lambda^2\boldsymbol{M}_1\mathbf{\Sigma}\boldsymbol{M}_2)^{\det}$, where

$$(\lambda^2\boldsymbol{M}_1\mathbf{\Sigma}\boldsymbol{M}_2)^{\det} = (1+\widetilde{v}(-\lambda;\phi_s,\phi))\,(v(-\lambda;\phi_s)\mathbf{\Sigma}+\boldsymbol{I}_p)^{-2}\,\mathbf{\Sigma} > 0, \tag{76}$$

and $\widetilde{v}(-\lambda;\phi_s,\phi)$ is as defined in (71). Let $\lambda \in \Lambda = [0, \lambda_{\max}]$ where $\lambda_{\max} < \infty$. For any matrix $\boldsymbol{T} \in \mathbb{R}^{p\times p}$ with trace norm uniformly bounded by M,

$$|\operatorname{tr}[\lambda^2\boldsymbol{M}_1\mathbf{\Sigma}\boldsymbol{M}_2\boldsymbol{T}]| \le \lambda^2 \|\boldsymbol{M}_1\|_{\text{op}} \|\boldsymbol{M}_2\|_{\text{op}} \|\mathbf{\Sigma}\|_{\text{op}} |\operatorname{tr}[(\boldsymbol{T}^\top\boldsymbol{T})^{\frac{1}{2}}]| \le M r_{\max} \|\mathbf{\Sigma}\|_{\text{op}}$$

where the second inequality holds because $\|\boldsymbol{M}_1\|_{\text{op}} \le \lambda^{-1}$ and $\|\mathbf{\Sigma}\|_{\text{op}} \le r_{\max}$. Since $\phi_0 \ge \phi_s > 1$, it follows from Patil et al. (2022b, Lemma S.6.14) that, there exists $M' > 0$ such that the magnitudes of $v(-\lambda;\phi_s)$ and $v_b(\lambda,\phi_s,\phi) - 1$, and their derivatives with respect to λ are continuous and bounded by M' for all $\lambda \in \Lambda$. Thus, we get

$$\begin{aligned}\left|\operatorname{tr}\left[(\lambda^2\boldsymbol{M}_1\mathbf{\Sigma}\boldsymbol{M}_2)^{\det}\boldsymbol{T}\right]\right| &\le (1+M')\,\|v(-\lambda;\phi_s)\mathbf{\Sigma}+\boldsymbol{I}_p\|_{\text{op}}^{-2}\,\|\mathbf{\Sigma}\|_{\text{op}}\,|\operatorname{tr}[(\boldsymbol{T}^\top\boldsymbol{T})^{\frac{1}{2}}]| \\ &\le (1+M')M r_{\max}.\end{aligned}$$

Similarly, in the same interval the derivatives of $\operatorname{tr}\left[\lambda^2\boldsymbol{M}_1\mathbf{\Sigma}\boldsymbol{M}_2\boldsymbol{T}\right]$ and $\operatorname{tr}\left[(\lambda^2\boldsymbol{M}_1\mathbf{\Sigma}\boldsymbol{M}_2)^{\det}\boldsymbol{T}\right]$ with respect to λ also have bounded magnitudes for $\lambda \in \Lambda$. Therefore, the family of functions

$$\operatorname{tr}[\lambda^2\boldsymbol{M}_1\mathbf{\Sigma}\boldsymbol{M}_2\boldsymbol{T}] - \operatorname{tr}\left[(\lambda^2\boldsymbol{M}_1\mathbf{\Sigma}\boldsymbol{M}_2)^{\det}\boldsymbol{T}\right]$$

forms an equicontinuous family in λ over $\lambda \in \Lambda$. Thus, the convergence in Part 1 of Lemma 38 is uniform in λ. We can now use the Moore-Osgood theorem and the continuity property from Lemma 44 to interchange the limits to obtain

$$\begin{aligned}&\lim_{p\to\infty}\operatorname{tr}\left[(\boldsymbol{I}_p - \widehat{\mathbf{\Sigma}}_1^+\widehat{\mathbf{\Sigma}}_1)\mathbf{\Sigma}(\boldsymbol{I}_p - \widehat{\mathbf{\Sigma}}_2^+\widehat{\mathbf{\Sigma}}_2)\boldsymbol{T}\right] - \operatorname{tr}\left[((\boldsymbol{I}_p - \widehat{\mathbf{\Sigma}}_1^+\widehat{\mathbf{\Sigma}}_1)\mathbf{\Sigma}(\boldsymbol{I}_p - \widehat{\mathbf{\Sigma}}_2^+\widehat{\mathbf{\Sigma}}_2))^{\det}\boldsymbol{T}\right] \\ &= \lim_{p\to\infty}\lim_{\lambda\to 0^+}\operatorname{tr}\left[\lambda^2\boldsymbol{M}_1\mathbf{\Sigma}\boldsymbol{M}_2\boldsymbol{T}\right] - \operatorname{tr}\left[(\lambda^2\boldsymbol{M}_1\mathbf{\Sigma}\boldsymbol{M}_2)^{\det}\boldsymbol{T}\right] \\ &= \lim_{\lambda\to 0^+}\lim_{p\to\infty}\operatorname{tr}\left[\lambda^2\boldsymbol{M}_1\mathbf{\Sigma}\boldsymbol{M}_2\boldsymbol{T}\right] - \operatorname{tr}\left[(\lambda^2\boldsymbol{M}_1\mathbf{\Sigma}\boldsymbol{M}_2)^{\det}\boldsymbol{T}\right] \\ &= 0,\end{aligned}$$

where

$$((\boldsymbol{I}_p - \widehat{\mathbf{\Sigma}}_1^+\widehat{\mathbf{\Sigma}}_1)\mathbf{\Sigma}(\boldsymbol{I}_p - \widehat{\mathbf{\Sigma}}_2^+\widehat{\mathbf{\Sigma}}_2))^{\det} = (1+\widetilde{v}(0;\phi,\phi_s))(v(0;\phi_s)\mathbf{\Sigma}+\boldsymbol{I}_p)^{-2}\mathbf{\Sigma}.$$

As $p \to \infty$, replacing the empirical distribution $G_p(r)$ by limiting distribution $G(r)$ yields the desired results. ∎

D.2.2 Deterministic approximation of the variance functional

Lemma 25 (Deterministic approximation of the variance functional when $\phi_s < 1$)**.** *Under Assumptions 1-5, for all $m \in [M]$ and $I_m \in \mathcal{I}_k$, let $\widehat{\mathbf{\Sigma}}_m = \mathbf{X}^\top \mathbf{L}_m \mathbf{X}/k$ and $\mathbf{L}_m \in \mathbb{R}^{n\times n}$ be a diagonal matrix with $(\mathbf{L}_m)_{ll} = 1$ if $l \in I_m$ and 0 otherwise. Then, it holds that*

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

$$\frac{1}{k}\operatorname{tr}(\widehat{\mathbf{\Sigma}}_m^+ \mathbf{\Sigma}) \xrightarrow{\text{a.s.}} \frac{\phi_s}{1-\phi_s},$$

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

$$\frac{1}{k}\operatorname{tr}(\widehat{\mathbf{\Sigma}}_l^+ \mathbf{X}^\top \mathbf{L}_l \mathbf{L}_m \widehat{\mathbf{\Sigma}}_m^+ \mathbf{\Sigma}) \xrightarrow{\text{a.s.}} \frac{\phi}{1-\phi}$$

as $n, k, p \to \infty$, $p/n \to \phi \in (0, \infty)$, and $p/k \to \phi_s \in [\phi, \infty) \cap (0, 1)$.

Proof For the first term, from Patil et al. (2022b, Proposition S.3.2) we have that for $m \in [M]$,

$$\frac{1}{k}\operatorname{tr}(\widehat{\mathbf{\Sigma}}_m^+ \mathbf{\Sigma}) \xrightarrow{\text{a.s.}} \begin{cases} \dfrac{\phi_s}{1-\phi_s} & \text{if } \phi_s \in (0,1) \\ \phi_s v_v(0;\phi,\phi_s) \displaystyle\int \frac{r^2}{(1+v(0;\phi_s)r)^2}\,\mathrm{d}H(r) & \text{if } \phi_s \in (1,\infty) \end{cases}. \tag{77}$$

Next we analyze the second term for $\phi_s \in (0, 1)$. It suffices to analyze the case when $(m, l) = (1, 2)$. From Bai and Silverstein (2010), we have

$$r_{\min}(1-\sqrt{\phi_s})^2 \le \liminf \left\|\widehat{\mathbf{\Sigma}}_j\right\|_{\mathrm{op}} \le \limsup \left\|\widehat{\mathbf{\Sigma}}_j\right\|_{\mathrm{op}} \le r_{\max}(1+\sqrt{\phi_s})^2, \quad j = 1, 2.$$

Then $\widehat{\mathbf{\Sigma}}_j$'s are invertible almost surely. From Lemma 40, we have that for $j = 1, 2$,

$$\widehat{\mathbf{\Sigma}}_j^{-1} = \left(\frac{i_0}{k}\widehat{\mathbf{\Sigma}}_0 + \frac{k-i_0}{k}\widehat{\mathbf{\Sigma}}_1^{\mathrm{ind}}\right)^{-1} \simeq \left(\frac{i_0}{k}\widehat{\mathbf{\Sigma}}_0 + (1-\phi_s)\frac{k-i_0}{k}\mathbf{\Sigma}\right)^{-1},$$

where $\widehat{\mathbf{\Sigma}}_0 = \mathbf{X}^\top \mathbf{L}_1 \mathbf{L}_2 \mathbf{X}/i_0$ and $\widehat{\mathbf{\Sigma}}_j^{\mathrm{ind}} = \mathbf{X}^\top \mathbf{L}_j \mathbf{X}/(k-i_0)$ for $j = 1, 2$, defined analogously as in the proof for Theorem 19. Thus, conditional on $\widehat{\mathbf{\Sigma}}_0$ and i_0, we have

$$\begin{aligned} \widehat{\mathbf{\Sigma}}_1^{-1}\widehat{\mathbf{\Sigma}}_0\widehat{\mathbf{\Sigma}}_2^{-1}\mathbf{\Sigma} &\simeq \left(\frac{i_0}{k}\widehat{\mathbf{\Sigma}}_0 + (1-\phi_s)\frac{k-i_0}{k}\mathbf{\Sigma}\right)^{-1}\widehat{\mathbf{\Sigma}}_0\left(\frac{i_0}{k}\widehat{\mathbf{\Sigma}}_0 + (1-\phi_s)\frac{k-i_0}{k}\mathbf{\Sigma}\right)^{-1} \\ &= \frac{i_0^2}{k^2}\left(\widehat{\mathbf{\Sigma}}_0 + (1-\phi_s)\frac{k-i_0}{i_0}\mathbf{\Sigma}\right)^{-1}\widehat{\mathbf{\Sigma}}_0\left(\widehat{\mathbf{\Sigma}}_0 + (1-\phi_s)\frac{k-i_0}{i_0}\mathbf{\Sigma}\right)^{-1}\mathbf{\Sigma} \end{aligned}$$

by applying the conditional product rule from Proposition 35. When $i_0 < k$, let $\widehat{\mathbf{\Sigma}}' = c\mathbf{\Sigma}^{-\frac{1}{2}}\widehat{\mathbf{\Sigma}}_0\mathbf{\Sigma}^{-\frac{1}{2}}$ and $c = (1-\phi_s)(k-i_0)/i_0$, we further have

$$\widehat{\mathbf{\Sigma}}_1^{-1}\widehat{\mathbf{\Sigma}}_0\widehat{\mathbf{\Sigma}}_2^{-1}\mathbf{\Sigma} \simeq \frac{i_0^2}{k^2c^2}\mathbf{\Sigma}^{-\frac{1}{2}}(\widehat{\mathbf{\Sigma}}' + \mathbf{I}_p)^{-1}\widehat{\mathbf{\Sigma}}'(\widehat{\mathbf{\Sigma}}' + \mathbf{I}_p)^{-1}\mathbf{\Sigma}^{-\frac{1}{2}}\mathbf{\Sigma}$$

$$\simeq \frac{i_0^2}{k^2}\widetilde{v}_v(-1;\gamma_0,c^{-1}\boldsymbol{I}_p)(v(-1;\gamma_0,c^{-1}\boldsymbol{I}_p)+c)^{-2}\boldsymbol{I}_p,$$

where $\gamma_0 = p/i_0$, the second equality is from Lemma 38 (2) and the fixed point solutions are defined by

$$\frac{1}{v(-1;\gamma_0,c^{-1}\boldsymbol{I}_p)} = 1+\frac{\gamma_0}{c+v(-1;\gamma_0,c^{-1}\boldsymbol{I}_p)}$$
$$\frac{1}{\widetilde{v}_v(-1;\gamma_0,c^{-1}\boldsymbol{I}_p)} = \frac{1}{v(-1;\gamma_0,c^{-1}\boldsymbol{I}_p)^2}-\frac{\gamma_0}{(c+v(-1;\gamma_0,c^{-1}\boldsymbol{I}_p))^2}.$$

When $i_0 = k$, the above equivalent is also valid, which reduces to the case for $\widehat{\boldsymbol{\Sigma}}_j^{+}\widehat{\boldsymbol{\Sigma}}_j\widehat{\boldsymbol{\Sigma}}_j^{+}$ as in (77). Note that from Lemma 42, $\widetilde{v}_v(-\lambda;\gamma)$ and $v(-\lambda;\gamma)$ are continuous on γ, and from Lemma 48, $i_0/k \xrightarrow{\text{a.s.}} \phi/\phi_s$ where $\phi_s \in (0,\infty)$ is the limiting ratio such that $p/k \to \phi_s$ as $k, p \to \infty$. We have

$$\widehat{\boldsymbol{\Sigma}}_1^{-1}\widehat{\boldsymbol{\Sigma}}_0\widehat{\boldsymbol{\Sigma}}_2^{-1}\boldsymbol{\Sigma} \simeq \frac{\phi_s^2}{\phi_0^2}\widetilde{v}_v(-1;\phi_0,c_0^{-1}\boldsymbol{I}_p)(v(-1;\phi_0,c_0^{-1}\boldsymbol{I}_p)+c_0)^{-2}\boldsymbol{I}_p,$$

where $c_0 = \lim_{p\to\infty} c = (1-\phi_s)(\phi_s-\phi)/\phi$ and the fixed solutions reduce to

$$v(-1;\gamma_0,c_0^{-1}\boldsymbol{I}_p) = 1-\phi_s, \qquad \widetilde{v}_v(-1;\gamma_0,c_0^{-1}\boldsymbol{I}_p) = \frac{(1-\phi_s)^2}{1-\phi}.$$

Then, we have

$$\frac{i_0}{k^2}\operatorname{tr}[\widehat{\boldsymbol{\Sigma}}_1^{+}\widehat{\boldsymbol{\Sigma}}_0\widehat{\boldsymbol{\Sigma}}_2^{+}\boldsymbol{\Sigma}] \xrightarrow{\text{a.s.}} \lim_{p\to\infty}\frac{i_0 p}{k^2}\cdot\frac{1}{p}\operatorname{tr}\left[\frac{\phi_s^2}{\phi^2}\frac{(1-\phi_s)^2}{1-\phi}\left(1-\phi_s+\frac{(1-\phi_s)(\phi_s-\phi)}{\phi}\right)^{-2}\boldsymbol{I}_p\right] = \frac{\phi}{1-\phi}. \tag{78}$$

Combining (77) and (78), the conclusion follows. ■

Lemma 26 (Deterministic approximation of the variance functional when $\phi_s > 1$)**.** *Under Assumptions 1-5, for all $m \in [M]$ and $I_m \in \mathcal{I}_k$, let $\widehat{\boldsymbol{\Sigma}}_m = \boldsymbol{X}^\top \boldsymbol{L}_m \boldsymbol{X}/k$ and $\boldsymbol{L}_m \in \mathbb{R}^{n\times n}$ be a diagonal matrix with $(\boldsymbol{L}_m)_{ll} = 1$ if $l \in I_m$ and 0 otherwise. Then, it holds that*

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

$$\frac{1}{k}\operatorname{tr}(\widehat{\boldsymbol{\Sigma}}_j^{+}\boldsymbol{\Sigma}) \xrightarrow{\text{a.s.}} \frac{1}{2}\widetilde{v}(0;\phi_s,\phi_s),$$

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

$$\frac{1}{k^2}\boldsymbol{\epsilon}^\top \boldsymbol{L}_1\boldsymbol{X}\widehat{\boldsymbol{\Sigma}}_m^{+}\boldsymbol{\Sigma}\widehat{\boldsymbol{\Sigma}}_l^{+}\boldsymbol{X}^\top\boldsymbol{L}_2\boldsymbol{\epsilon} \xrightarrow{\text{a.s.}} \frac{1}{2}\widetilde{v}(0;\phi,\phi_s),$$

as $n, k, p \to \infty$, $p/n \to \phi \in (0,\infty)$, and $p/k \to \phi_s \in [\phi,\infty)\cap(1,\infty)$, where the nonnegative constants $v(0;\phi_s)$ and $\widetilde{v}(0;\phi,\phi_s)$ are as defined in (70) *and* (71).

Proof From (77) we have

$$\frac{1}{k}\operatorname{tr}(\widehat{\boldsymbol{\Sigma}}_m^+\boldsymbol{\Sigma}) \xrightarrow{\text{a.s.}} \widetilde{v}(0;\phi,\phi_s). \tag{79}$$

For the second term, it suffices to consider the case when $(m,l) = (1,2)$. Let $P_0 = \boldsymbol{\epsilon}^\top \boldsymbol{L}_1 \boldsymbol{X}\widehat{\boldsymbol{\Sigma}}_1^+\boldsymbol{\Sigma}\widehat{\boldsymbol{\Sigma}}_2^+\boldsymbol{X}^\top\boldsymbol{L}_2\boldsymbol{\epsilon}/k^2$ and $P_\lambda = \boldsymbol{\epsilon}^\top\boldsymbol{L}_1\boldsymbol{X}\boldsymbol{M}_1\boldsymbol{\Sigma}\ \boldsymbol{M}_2\boldsymbol{X}^\top\boldsymbol{L}_2\boldsymbol{\epsilon}/k^2$ where $\boldsymbol{M}_j = (\widehat{\boldsymbol{\Sigma}}_j + \lambda\boldsymbol{I}_p)^{-1}$. Note that $\lim_{\lambda\to 0^+} P_\lambda = P_0$. Note that $\lim_{\lambda\to 0^+} P_\lambda = P_0$. From Patil et al. (2022a, Lemma S.2.3) and Lemma 21, we have that for any fixed $\lambda > 0$,

$$P_\lambda \xrightarrow{\text{a.s.}} Q_\lambda := \widetilde{v}(-\lambda;\phi,\phi_s),$$

as $n,k,p\to\infty$, $p/n\to\phi\in(0,\infty)$, and $p/k\to\phi_s\in[\phi,\infty)\setminus\{1\}$, where $\widetilde{v}(\lambda,\phi_s,\phi)$ is as defined in (52). Because of the continuity of $\widetilde{v}_v(-\lambda;\phi)$ and $v(-\lambda;\phi)$ in λ from Lemma 44, we have that

$$\lim_{\lambda\to 0^+} Q_\lambda = Q_0 := \widetilde{v}(0;\phi,\phi_s).$$

As $n,p\to\infty$, we have that almost surely

$$|P_\lambda| = \phi|\operatorname{tr}(\boldsymbol{M}_2\widehat{\boldsymbol{\Sigma}}_0\boldsymbol{M}_1\boldsymbol{\Sigma})/p| \le \phi\|\boldsymbol{M}_1\widehat{\boldsymbol{\Sigma}}_0\boldsymbol{M}_2\|_{\mathrm{op}}\|\boldsymbol{\Sigma}\|_{\mathrm{op}} \le \frac{\phi_s^2 r_{\max}}{\phi},$$

where the last inequality is because $\left\|\widehat{\boldsymbol{\Sigma}}_0\right\|_{\mathrm{op}} \le r_{\max}$, and

$$\left\|\boldsymbol{M}_1\widehat{\boldsymbol{\Sigma}}_0\boldsymbol{M}_2\right\|_{\mathrm{op}} \le \frac{k^2}{i_0^2}\cdot\max_i \frac{l_i}{\left(l_i + \frac{k-i_0}{i_0}\lambda\right)^2} \le \frac{k^2}{i_0^2}, \tag{80}$$

where l_i's are the eigenvalues of $\widehat{\boldsymbol{\Sigma}}_0$. Similarly, we have $|P_0|$ is almost surely bounded. Thus, $|P_\lambda|$ is almost surely bounded over $\lambda\in\Lambda[0,\lambda_{\max}]$ for some constant $\lambda_{\max}>0$. Next we consider the derivative

$$\begin{aligned}\frac{\partial P_\lambda}{\partial\lambda} &= \boldsymbol{\epsilon}^\top\boldsymbol{L}_1\boldsymbol{X}\frac{\partial\boldsymbol{M}_1}{\partial\lambda}\boldsymbol{\Sigma}\boldsymbol{M}_2\boldsymbol{X}^\top\boldsymbol{L}_2\boldsymbol{\epsilon}/k^2 + \boldsymbol{\epsilon}^\top\boldsymbol{L}_1\boldsymbol{X}\boldsymbol{M}_1\boldsymbol{\Sigma}\frac{\partial\boldsymbol{M}_2}{\partial\lambda}\boldsymbol{X}^\top\boldsymbol{L}_2\boldsymbol{\epsilon}/k^2\\ &= -\boldsymbol{\epsilon}^\top\boldsymbol{L}_1\boldsymbol{X}\boldsymbol{M}_1^2\boldsymbol{\Sigma}\boldsymbol{M}_2\boldsymbol{X}^\top\boldsymbol{L}_2\boldsymbol{\epsilon}/k^2 - \boldsymbol{\epsilon}^\top\boldsymbol{L}_1\boldsymbol{X}\boldsymbol{M}_1\boldsymbol{\Sigma}\boldsymbol{M}_2^2\boldsymbol{X}^\top\boldsymbol{L}_2\boldsymbol{\epsilon}/k^2\end{aligned}$$

Note that for $\lambda\in\Lambda$, we can bound

$$\left\|\boldsymbol{M}_1^2\widehat{\boldsymbol{\Sigma}}_0\boldsymbol{M}_2\right\|_{\mathrm{op}} \le \frac{k^2}{i_0^2}\cdot\max_i \frac{l_i}{\left(l_i + \frac{k-i_0}{i_0}\lambda\right)^3} \le \frac{k^2}{i_0^2},$$

where l_i's are the eigenvalues of $\widehat{\boldsymbol{\Sigma}}_0$. Similarly, we have that $\left\|\boldsymbol{M}_1\widehat{\boldsymbol{\Sigma}}_0\boldsymbol{M}_2^2\right\|_{\mathrm{op}}$ is almost surely bounded for $\lambda\in\Lambda$. By similar argument as in Patil et al. (2022a, Lemma S.3.3), the following holds almost surely as $n,p\to\infty$,

$$\left|\frac{\partial P_\lambda}{\partial\lambda}\right| = \phi|\operatorname{tr}(\boldsymbol{M}_1^2\widehat{\boldsymbol{\Sigma}}_0\boldsymbol{M}_2\boldsymbol{\Sigma}) + \operatorname{tr}(\boldsymbol{M}_1\widehat{\boldsymbol{\Sigma}}_0\boldsymbol{M}_2^2\boldsymbol{\Sigma})| \le \frac{\phi_s^2 r_{\max}}{\phi}.$$

That is, $|\partial P_\lambda/\partial\lambda|$ is almost surely bounded over $\lambda \in \Lambda[0, \lambda_{\max}]$.

Since $\phi_0 \geq \phi_s > 1$, it follows from Patil et al. (2022b, Lemma S.6.14) that, there exists $M' > 0$ such that the magnitudes of $v(-\lambda;\phi_s)$ and $v_v(\lambda,\phi_s,\phi)/\phi$, and their derivatives with respect to λ are continuous and bounded by M' for all $\lambda \in \Lambda$. Thus, $|Q_\lambda| \leq \phi_0 M' r_{\max}^2$ over $\lambda \in \Lambda$. Similarly, we have $|\partial Q_\lambda/\partial\lambda|_{\lambda=0^+}|$ are uniformly bounded over $\lambda \in \Lambda$. We can now use the Moore-Osgood theorem and the continuity property from Lemma 44 to interchange the limits to obtain

$$\lim_{p\to\infty} P_0 - Q_0 = \lim_{p\to\infty}\lim_{\lambda\to0^+} P_\lambda - Q_\lambda = \lim_{\lambda\to0^+}\lim_{p\to\infty} P_\lambda - Q_\lambda = 0,$$

and the conclusion follows. ■

D.3 Boundary case: diverging subsample aspect ratio

Proposition 27 (Risk approximation when $\phi_s \to \infty$)**.** *Under Assumptions 1-5, it holds for all $M \in \mathbb{N}$*

$$R(\widetilde{f}^{\mathtt{WR}}_{0,M}; \mathcal{D}_n, \{I_\ell\}_{\ell=1}^M) \xrightarrow{\text{a.s.}} \mathscr{R}^{\mathtt{sub}}_{0,M}(\phi,\infty),$$

as $k,n,p\to\infty$, $p/n \to \phi \in (0,\infty)$ and $p/k \to \infty$, where

$$\mathscr{R}^{\mathtt{sub}}_{0,M}(\phi,\infty) := \lim_{\phi_s\to\infty} \mathscr{R}^{\mathtt{sub}}_{0,M}(\phi,\phi_s) = \sigma^2 + \rho^2 \int r\,\mathrm{d}G(r), \tag{81}$$

and $\mathscr{R}^{\mathtt{sub}}_{0,M}(\phi,\phi_s)$ is as defined in Theorem 23.

Proof Note that

$$\begin{aligned}
R(\widetilde{f}^{\mathtt{WR}}_{0,M}; \mathcal{D}_n, \{I_\ell\}_{\ell=1}^M) &= \mathbb{E}_{(\boldsymbol{x}_0,y_0)}[(y_0 - \boldsymbol{x}_0^\top \widetilde{\boldsymbol{\beta}}_{0,M}(\{\mathcal{D}_{I_\ell}\}_{\ell=1}^M))^2] \\
&= \mathbb{E}_{(\boldsymbol{x}_0,y_0)}[(\boldsymbol{\epsilon}_0 + \boldsymbol{x}_0^\top(\boldsymbol{\beta}_0 - \widetilde{\boldsymbol{\beta}}_{0,M}(\{\mathcal{D}_{I_\ell}\}_{\ell=1}^M)))^2] \\
&= \sigma^2 + \mathbb{E}_{(\boldsymbol{x}_0,y_0)}[(\boldsymbol{\beta}_0 - \widetilde{\boldsymbol{\beta}}_{0,M}(\{\mathcal{D}_{I_\ell}\}_{\ell=1}^M))^\top \boldsymbol{x}_0\boldsymbol{x}_0^\top(\boldsymbol{\beta}_0 - \widetilde{\boldsymbol{\beta}}_{0,M}(\{\mathcal{D}_{I_\ell}\}_{\ell=1}^M))] \\
&= \sigma^2 + (\boldsymbol{\beta}_0 - \widetilde{\boldsymbol{\beta}}_{0,M}(\{\mathcal{D}_{I_\ell}\}_{\ell=1}^M))^\top \boldsymbol{\Sigma}(\boldsymbol{\beta}_0 - \widetilde{\boldsymbol{\beta}}_{0,M}(\{\mathcal{D}_{I_\ell}\}_{\ell=1}^M)).
\end{aligned}$$

Then, by the Cauchy-Schwarz inequality, we have

$$\begin{aligned}
R(\widetilde{f}^{\mathtt{WR}}_{0,M}; \mathcal{D}_n, \{I_\ell\}_{\ell=1}^M) - (\boldsymbol{\beta}_0^\top\boldsymbol{\Sigma}\boldsymbol{\beta}_0 + \sigma^2) &= \|\boldsymbol{\Sigma}^{\frac{1}{2}}\widetilde{\boldsymbol{\beta}}_{0,M}(\{\mathcal{D}_{I_\ell}\}_{\ell=1}^M)\|_2^2 - 2\widetilde{\boldsymbol{\beta}}_{0,M}(\{\mathcal{D}_{I_\ell}\}_{\ell=1}^M)^\top\boldsymbol{\Sigma}\boldsymbol{\beta}_0 \\
&\leq \frac{1}{r_{\min}}\|\widetilde{\boldsymbol{\beta}}_{0,M}(\{\mathcal{D}_{I_\ell}\}_{\ell=1}^M)\|_2^2 + 2\|\widetilde{\boldsymbol{\beta}}_{0,M}(\{\mathcal{D}_{I_\ell}\}_{\ell=1}^M)\|_2\|\boldsymbol{\Sigma}\|_2 \\
&\leq \frac{1}{r_{\min}}\|\widetilde{\boldsymbol{\beta}}_{0,M}(\{\mathcal{D}_{I_\ell}\}_{\ell=1}^M)\|_2^2 + 2r_{\max}\rho\|\widetilde{\boldsymbol{\beta}}_{0,M}(\{\mathcal{D}_{I_\ell}\}_{\ell=1}^M)\|_2
\end{aligned}$$

almost surely as $k,n,p\to$ and $p/k\to\infty$. Thus, we have the following holds almost surely:

$$\begin{aligned}
\|\widetilde{\boldsymbol{\beta}}^0_M(\mathcal{D}_n)\|_2 &\leq \frac{1}{M}\sum_{m=1}^M \|(\boldsymbol{X}^\top\boldsymbol{L}_m\boldsymbol{X}/k)^+(\boldsymbol{X}^\top\boldsymbol{L}_m\boldsymbol{y}/k)\|_2 \\
&\leq \frac{1}{M}\sum_{m=1}^M \|(\boldsymbol{X}^\top\boldsymbol{L}_m\boldsymbol{X}/k)^+\boldsymbol{X}^\top\boldsymbol{L}_m/\sqrt{k}\| \cdot \|\boldsymbol{L}_m\boldsymbol{y}/\sqrt{k}\|_2
\end{aligned}$$

$$\leq C\sqrt{\rho^2+\sigma^2}\cdot\frac{1}{M}\sum_{m=1}^{M}\|(\boldsymbol{X}^\top\boldsymbol{L}_m\boldsymbol{X}/k)^+\boldsymbol{X}^\top\boldsymbol{L}_m/\sqrt{k}\|$$

where the last inequality holds eventually almost surely since Assumptions 1-3 imply that the entries of $\boldsymbol{y}$ have bounded 4-th moment, and thus from the strong law of large numbers, $\|\boldsymbol{L}_m\boldsymbol{y}/\sqrt{k}\|_2$ is eventually almost surely bounded above by $C\sqrt{\mathbb{E}[y_1^2]}=C\sqrt{\rho^2+\sigma^2}$ for some constant C. Observe that operator norm of the matrix $(\boldsymbol{X}^\top\boldsymbol{L}_m\boldsymbol{X}/k)^+\boldsymbol{X}\boldsymbol{L}_m/\sqrt{k}$ is upper bounded $1/s_{\min}$, where $s_{\min}$ is the smallest nonzero singular value of $\boldsymbol{X}$. As $k,p\to\infty$ such that $p/k\to\infty$, $s_{\min}\to\infty$ almost surely (e.g., from results in Bloemendal et al. (2016)), and therefore, $\|\widetilde{\boldsymbol{\beta}}_{0,M}(\{\mathcal{D}_{I_\ell}\}_{\ell=1}^M)\|_2\to 0$ almost surely. Thus, we have shown that

$$R(\widetilde{f}^{\texttt{WR}}_{0,M};\mathcal{D}_n,\{I_\ell\}_{\ell=1}^M)\xrightarrow{\text{a.s.}}\sigma^2+\boldsymbol{\beta}_0^\top\boldsymbol{\Sigma}\boldsymbol{\beta}_0,$$

or equivalently,

$$R(\widetilde{f}^{\texttt{WR}}_{0,M};\mathcal{D}_n,\{I_\ell\}_{\ell=1}^M)\xrightarrow{\text{a.s.}}\sigma^2+\rho^2\int r\,\mathrm{d}G(r).$$

From Lemma 43 we have

$$\lim_{\phi_s\to\infty}v(0;\phi_s)=\lim_{\phi_s\to\infty}\widetilde{v}_b(0;\phi_s)=\lim_{\phi_s\to\infty}\widetilde{v}_v(0;\phi_s).$$

Thus,

$$\lim_{\phi_s\to\infty}V_0(\phi_s,\phi_s)=\lim_{\phi_s\to\infty}V_0(\phi,\phi_s)=0,$$

and

$$\lim_{\phi_s\to\infty}B_0(\phi_s,\phi_s)=\lim_{\phi_s\to\infty}B_0(\phi,\phi_s)=\rho^2\int r\,\mathrm{d}G(r).$$

Therefore, we have $\mathscr{R}^{\texttt{WR}}_{0,M}(\phi,\infty):=\lim_{\phi_s\to\infty}\mathscr{R}^{\texttt{sub}}_{0,M}(\phi,\phi_s)=\sigma^2+\rho^2\int r\,\mathrm{d}G(r)$. Thus, $\mathscr{R}^{\texttt{WR}}_{0,M}(\phi,\infty)$ is well defined and $\mathscr{R}^{\texttt{WR}}_{0,M}(\phi,\phi_s)$ is right continuous at $\phi_s=\infty$. ■

Appendix E. Proof of Theorem 8 (splagging without replacement, ridge and ridgeless predictors)

Proof For $M\in\{1,2,\dots,\lfloor\liminf n/k\rfloor\}$, following the proof in Theorem 19, the conditional risk is given by

$$R(\widetilde{f}^{\texttt{WOR}}_{\lambda,M};\{\mathcal{D}_{I_\ell}\}_{\ell=1}^M)=\sigma^2+T_C+T_B+T_V,$$

where T_C, T_B, and T_V are defined as

$$T_C=-\frac{\lambda}{M}\cdot\boldsymbol{\epsilon}^\top\left(\sum_{m=1}^{M}\boldsymbol{M}_m\frac{\boldsymbol{X}^\top\boldsymbol{L}_m}{k}\right)^\top\boldsymbol{\Sigma}\left(\sum_{m=1}^{M}\boldsymbol{M}_m\right)\boldsymbol{\beta}_0,\tag{82}$$

$$T_B=\frac{\lambda^2}{M^2}\cdot\boldsymbol{\beta}_0^\top\left(\sum_{i=1}^{M}\boldsymbol{M}_{I_i}\right)\boldsymbol{\Sigma}\left(\sum_{i=1}^{M}\boldsymbol{M}_{I_i}\right)\boldsymbol{\beta}_0\tag{83}$$

$$
\begin{aligned}
&= \frac{\lambda^2}{M}\sum_{i=1}^M \boldsymbol{\beta}_0^\top \boldsymbol{M}_{I_i}\boldsymbol{\Sigma}\boldsymbol{M}_{I_i}\boldsymbol{\beta}_0 + \frac{\lambda^2(M-1)}{M}\sum_{i,j=1}^M \boldsymbol{\beta}_0^\top \boldsymbol{M}_{I_i}\boldsymbol{\Sigma}\boldsymbol{M}_{I_j}\boldsymbol{\beta}_0,\\
T_V &= \frac{1}{M^2}\cdot \boldsymbol{\epsilon}^\top \left(\sum_{i=1}^M \boldsymbol{M}_{I_i}\frac{\boldsymbol{X}^\top\boldsymbol{L}_i}{k}\right)^\top \boldsymbol{\Sigma}\left(\sum_{i=1}^M \boldsymbol{M}_{I_i}\frac{\boldsymbol{X}^\top\boldsymbol{L}_i}{k}\right)\boldsymbol{\epsilon}\\
&= \frac{1}{M}\sum_{i=1}^M \left(\boldsymbol{M}_{I_i}\frac{\boldsymbol{X}^\top\boldsymbol{L}_i}{k}\right)^\top \boldsymbol{\Sigma}\left(\boldsymbol{M}_{I_i}\frac{\boldsymbol{X}^\top\boldsymbol{L}_i}{k}\right) + \frac{M-1}{M}\sum_{i,j=1}^M \left(\boldsymbol{M}_{I_i}\frac{\boldsymbol{X}^\top\boldsymbol{L}_i}{k}\right)^\top \boldsymbol{\Sigma}\left(\boldsymbol{M}_{I_j}\frac{\boldsymbol{X}^\top\boldsymbol{L}_j}{k}\right),
\end{aligned}
\tag{84}
$$

where $\boldsymbol{M}_{I_\ell} = (\boldsymbol{X}^\top\boldsymbol{L}_\ell\boldsymbol{X}/k + \lambda\boldsymbol{I}_p)^{-1}$ and $\boldsymbol{L}_\ell$ is a diagonal matrix with diagonal entry being 1 if the ℓth sample X_ℓ is in the sub-sampled dataset $\mathcal{D}_{I_\ell}$ and 0 otherwise. Note that for splagging, $I_i \cap I_j = \varnothing$ for all $i \neq j$.

We analyze each term separately for $M \in \{1, 2\}$. From Patil et al. (2022a, Lemma S.2.2), we have that $T_C \xrightarrow{\text{a.s.}} 0$. From Patil et al. (2022a, Lemma S.2.3), we have that

$$
T_V - \frac{1}{M}\sum_{j=1}^M \frac{\sigma^2}{k}\operatorname{tr}(\boldsymbol{M}_{I_j}\widehat{\boldsymbol{\Sigma}}_j\boldsymbol{M}_{I_j}\boldsymbol{\Sigma}) \xrightarrow{\text{a.s.}} 0, \tag{85}
$$

since the datasets have no overlaps and the cross term vanishes because $\boldsymbol{L}_l\boldsymbol{L}_m = \boldsymbol{0}_{n\times n}$ for $l \neq m$. Then, from (57) and (65), we have that for $\ell \in [M]$,

$$
\lambda^2\boldsymbol{\beta}_0^\top \boldsymbol{M}_{I_\ell}\boldsymbol{\Sigma}\boldsymbol{M}_{I_\ell}\boldsymbol{\beta}_0 \xrightarrow{\text{a.s.}} \rho^2\widetilde{v}(-\lambda;\phi_s,\phi_s)\widetilde{c}(-\lambda;\phi_s), \tag{86}
$$

$$
\frac{\sigma^2}{k}\operatorname{tr}(\boldsymbol{M}_{I_\ell}\widehat{\boldsymbol{\Sigma}}_\ell\boldsymbol{M}_{I_\ell}\boldsymbol{\Sigma}) \xrightarrow{\text{a.s.}} \frac{\sigma^2}{2}\widetilde{v}(-\lambda;\phi_s,\phi_s), \tag{87}
$$

as $n, k, p \to \infty$, $p/n \to \phi \in (0,\infty)$, and $p/k \to \phi_s = 2\phi$, where the positive constants $\widetilde{v}(\lambda;\phi_s,\phi)$, and $\widetilde{c}(-\lambda;\phi_s)$ are as defined in (52). For the cross term ($i \neq j$), setting $i_0 = 0$ in (59) yields that

$$
\boldsymbol{M}_{I_i}\boldsymbol{\Sigma}\boldsymbol{M}_{I_j} \simeq (v(-\lambda;\phi_s)\boldsymbol{\Sigma} + \boldsymbol{I}_p)^{-1}\boldsymbol{\Sigma}(v(-\lambda;\phi_s)\boldsymbol{\Sigma} + \boldsymbol{I}_p)^{-1}.
$$

Thus,

$$
\lambda^2\boldsymbol{\beta}_0^\top \boldsymbol{M}_{I_i}\boldsymbol{\Sigma}\boldsymbol{M}_{I_j}\boldsymbol{\beta}_0 \xrightarrow{\text{a.s.}} \rho^2\int \frac{r}{(1+v(-\lambda;\phi_s)r)^2}\,\mathrm{d}G(r) = \rho^2\widetilde{c}(0;\phi_s). \tag{88}
$$

Combining (82)-(88), we have shown that $R(\widetilde{f}^{\texttt{WOR}}_{\lambda,M};\{\mathcal{D}_{I_\ell}\}_{\ell=1}^M) \xrightarrow{\text{a.s.}} \mathscr{R}^{\texttt{spl}}_{\lambda,M}(\phi,\phi_s)$, where

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

and the components are:

$$
\mathscr{B}^{\texttt{spl}}_{\lambda,M}(\phi,\phi_s) = \frac{1}{M}B_\lambda(\phi_s,\phi_s) + \left(1 - \frac{1}{M}\right)C_\lambda(\phi_s), \qquad \mathscr{V}^{\texttt{spl}}_{\lambda,M}(\phi,\phi_s)\frac{1}{M}V_\lambda(\phi_s,\phi_s),
$$

with $B_\lambda(\phi,\phi_s) = \rho^2(1+\widetilde{v}(-\lambda;\phi,\phi_s))\widetilde{c}(-\lambda;\phi_s)$, $C_\lambda(\phi_s) = \rho^2\widetilde{c}(-\lambda;\phi_s)$, and $V_\lambda(\phi,\phi_s) = \sigma^2\widetilde{v}(-\lambda;\phi,\phi_s)$.

From Proposition 22 and Proposition 27, we have that for all $\lambda \in [0, \infty)$ and $M \in \{1, 2\}$,

$$\lim_{\phi_s \to +\infty} R(\widetilde{f}^{\texttt{WOR}}_{\lambda,M}; \{\mathcal{D}_k^{(m)}\}_{m=1}^M) = \sigma^2 + \rho^2 \int r \, \mathrm{d}G(r),$$

$\lim_{\phi_s \to +\infty} B_\lambda(\phi_s, \phi_s) = \rho^2 \int r \, \mathrm{d}G(r)$ and $\lim_{\phi_s \to +\infty} v(-\lambda; \phi_s) = \lim_{\phi_s \to +\infty} V_\lambda(\phi_s, \phi_s) = 0$. Then

$$\lim_{\phi_s \to +\infty} \widetilde{c}(-\lambda; \phi_s) = \lim_{\phi_s \to +\infty} \int r(1 + v(-\lambda; \phi_s) r)^{-2} \, \mathrm{d}G(r) = \int r \, \mathrm{d}G(r).$$

Thus, the approximation holds when $\phi_s = \infty$: $\lim_{\phi_s \to +\infty} R(\widetilde{f}^{\texttt{WOR}}_{\lambda,M}; \{\mathcal{D}_{I_\ell}\}_{\ell=1}^M) = \lim_{\phi_s \to +\infty} \mathscr{R}^{\texttt{spl}}_{\lambda,M}(\phi, \phi_s)$.

Finally, the risk expression for general M and the uniform statement for all $M \leq \lfloor n/k \rfloor$ follow from Theorem 5. ■

Appendix F. Proofs related to bagged risk properties

F.1 Proof of Proposition 7 (bias-variance monotonicities in the number of bags, subagging with replacement)

Proof Recall that from the proof for Theorem 19, we have

$$\mathscr{B}^{\texttt{sub}}_{\lambda,1}(\phi, \phi_s) = \rho^2 (1 + \widetilde{v}(-\lambda, \phi_s, \phi_s)) \widetilde{c}(-\lambda; \phi_s) \qquad \mathscr{V}^{\texttt{sub}}_{\lambda,1}(\phi, \phi_s) = \sigma^2 \widetilde{v}(-\lambda; \phi_s, \phi_s)$$
$$\mathscr{B}^{\texttt{sub}}_{\lambda,\infty}(\phi, \phi_s) = \rho^2 (1 + \widetilde{v}(-\lambda, \phi, \phi_s)) \widetilde{c}(-\lambda; \phi_s) \qquad \mathscr{B}^{\texttt{sub}}_{\lambda,\infty}(\phi, \phi_s) = \sigma^2 \widetilde{v}(-\lambda; \phi, \phi_s)$$

where the nonnegative constants $\widetilde{v}(-\lambda, \phi, \phi_s)$ and $\widetilde{c}(-\lambda; \phi_s)$ are defined in (51). Since H has positive support, $\widetilde{v}(-\lambda; \phi, \phi_s)$ is strictly increasing in ϕ, and thus, $\mathscr{B}^{\texttt{sub}}_{\lambda,\infty}(\phi, \phi_s) = \mathscr{B}^{\texttt{sub}}_{\lambda,1}(\phi, \phi_s)$ when $\phi_s = \phi$, and $\mathscr{B}^{\texttt{sub}}_{\lambda,1}(\phi, \phi_s) > \mathscr{B}^{\texttt{sub}}_{\lambda,\infty}(\phi, \phi_s)$ when $\phi_s > \phi$. Similarly, $\mathscr{V}^{\texttt{sub}}_{\lambda,\infty}(\phi, \phi_s) = \mathscr{V}^{\texttt{sub}}_{\lambda,1}(\phi, \phi_s)$ when $\phi_s = \phi$ and $\mathscr{V}^{\texttt{sub}}_{\lambda,1}(\phi, \phi_s) < \mathscr{V}^{\texttt{sub}}_{\lambda,\infty}(\phi, \phi_s)$ when $\phi_s > \phi$. Recall that the definitions of $\mathscr{B}^{\texttt{sub}}_{\lambda,M}(\phi, \phi_s) = 1/M \cdot B_\lambda(\phi_s, \phi_s) + (1 - 1/M) B_\lambda(\phi, \phi_s)$ and $\mathscr{V}^{\texttt{sub}}_{\lambda,M}(\phi, \phi_s) = 1/M \cdot V_\lambda(\phi_s, \phi_s) + (1 - 1/M) V_\lambda(\phi, \phi_s)$ are a convex combination of $B_\lambda(\phi, \phi_s)$ and $B_\lambda(\phi_s, \phi_s)$, and $V_\lambda(\phi, \phi_s)$ and $V_\lambda(\phi_s, \phi_s)$, respectively. The proof for ridgeless predictor follows by setting $\lambda = 0$ except $B_0(\phi, \phi_s) = B_0(\phi, \phi_s) = 0$ for $\phi_s < 1$. ■

F.2 Proof of Proposition 9 (bias-variance monotonicities in the number of bags, splagging without replacement)

Proof For the variance term, $\mathscr{V}^{\texttt{spl}}_{\lambda,M}(\phi, \phi_s) = M^{-1} V_\lambda(\phi_s, \phi_s)$ as a linear function of M^{-1} is strictly decreasing in M if $\phi_s < \infty$ and is zero if $\phi_s = \infty$ or $\sigma^2 = 0$.

For the bias term, when $\phi_s > 1$, since $\widetilde{c}(-\lambda; \phi_s) > 0$, we have that $B_\lambda(\phi_s, \phi_s) \geq C_\lambda(\phi_s)$ with equality holds if and only if $\widetilde{v}(-\lambda; \phi, \phi_s) = 0$ or $\widetilde{c}(-\lambda; \phi_s) = 0$, if and only if $\phi_s = \infty$. Then we have

$$\begin{aligned}
\mathscr{B}^{\texttt{spl}}_{\lambda,M}(\phi, \phi_s) &= \frac{1}{M} B_\lambda(\phi_s, \phi_s) + \left(1 - \frac{1}{M}\right) C_\lambda(\phi_s) \\
&= \frac{1}{M} (B_\lambda(\phi_s, \phi_s) - C_\lambda(\phi_s)) + C_\lambda(\phi_s)
\end{aligned}$$

$$\geq \frac{1}{M+1}(B_\lambda(\phi_s, \phi_s) - C_\lambda(\phi_s)) + C_\lambda(\phi_s)$$
$$= \frac{1}{M+1} B_\lambda(\phi_s, \phi_s) + \left(1 - \frac{1}{M+1}\right) C_\lambda(\phi_s)$$
$$= \mathscr{B}^{\texttt{spl}}_{\lambda, M+1}(\phi, \phi_s).$$

with equality holds if $\phi_s = \infty$ or $\rho^2 = 0$. When $\phi_s < 1$, $B_\lambda(\phi_s, \phi_s) \geq C_\lambda(\phi_s)$ with equality holds if and only if $\widetilde{c}(-\lambda; \phi_s) = 0$, if and only if $\lambda = 0$. The monotonicity of $\mathscr{V}^{\texttt{spl}}_{\lambda,M}(\phi, \phi_s)$ in M follows analogously.

As $M \leq \phi_s/\phi$, we further have $\mathscr{V}^{\texttt{spl}}_{\lambda,M}(\phi, \phi_s) \geq \mathscr{V}^{\texttt{spl}}_{\lambda,\phi_s/\phi}(\phi, \phi_s)$ and $\mathscr{V}^{\texttt{spl}}_{\lambda,M}(\phi, \phi_s) \geq \mathscr{V}^{\texttt{spl}}_{\lambda,\phi_s/\phi}(\phi, \phi_s)$ for all $M = 1, \ldots, \lfloor \liminf n/k \rfloor$. ■

F.3 Proof of Theorem 10 (risk monotonization of general bagged predictors by cross-validation)

Proof We present the proof for bagging with replacement, and the proof for bagging without replacement follows by restricting the support of $\phi_s \mapsto \mathscr{R}_M(\phi, \phi_s)$ to $[M\phi, \infty]$. From Theorem 5, we have that for any $M \in \mathbb{N}$ and $\{I_\ell\}_{\ell=1}^M$ simple random samples from $\mathcal{I}_k$ or $\mathcal{I}_k^\pi$, it holds that

$$R(\widetilde{f}_M; \mathcal{D}_n, \{I_\ell\}_{\ell=1}^M) \xrightarrow{\text{p}} \mathscr{R}_M(\phi, \phi_s)$$

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

$$\mathscr{R}_M(\phi, \phi_s) := (2\mathscr{R}(\phi, \phi_s) - \mathscr{R}(\phi_s, \phi_s)) + \frac{2}{M}(\mathscr{R}(\phi_s, \phi_s) - \mathscr{R}(\phi, \phi_s)).$$

From Patil et al. (2022b, Lemma 3.8 and Theorem 3.4), we have that

$$\left(R(\widehat{f}^{\text{cv}}_{M,\mathcal{I}_{\widehat{k}}}; \mathcal{D}_n) - \mathscr{R}_M(\phi, \phi_s)\right)_+ \xrightarrow{\text{p}} 0.$$

In Patil et al. (2022b), we have assumed that the risk is bounded away from 0 in order to conclude that the relative error converges to 0. But in Theorem 10, we conclude only the positive part of the absolute error converges to 0, for which we do not require the risk to be bounded away from 0.

Since $\mathscr{R}_M(\phi, \phi_s)$ is increasing in ϕ for any fixed ϕ_s. For $0 < \phi_1 \leq \phi_2 < \infty$,

$$\min_{\phi_s \geq \phi_1} \mathscr{R}_M(\phi_1, \phi_s) \leq \min_{\phi_s \geq \phi_2} \mathscr{R}_M(\phi_1, \phi_s) \leq \min_{\phi_s \geq \phi_2} \mathscr{R}_M(\phi_2, \phi_s)$$

where the first inequality follows because $\{\phi_s : \phi_s \geq \phi_1\} \supseteq \{\phi_s : \phi_s \geq \phi_2\}$, and the second inequality follows because $\mathscr{R}_M(\phi, \phi_s)$ is increasing in ϕ for a fixed ϕ_s. Thus, $\min_{\phi_s \geq \phi} \mathscr{R}_M(\phi, \phi_s)$ is a monotonically increasing function in ϕ. ■

F.4 Proof of Theorem 11 (risk monotonization of ridge bagged predictors by cross-validation)

Proof It suffices to verify the two conditions (i) and (ii) in Theorem 10. From Theorem 6 and Theorem 8, condition (i) holds naturally with $\mathscr{R}_M(\phi,\phi_s)$ being the limiting risk $\mathscr{R}^{\mathtt{sub}}_{\lambda,M}(\phi,\phi_s)$ (or $\mathscr{R}^{\mathtt{spl}}_{\lambda,M}(\phi,\phi_s)$) for fixed $\lambda \geq 0$. For condition (ii), note that when $\lambda > 0$, $\mathscr{R}_M(\phi,\phi_s)$ is continuous over $[\phi,\infty]$. When $\lambda = 0$, $\mathscr{R}_M(\phi,\phi_s)$ is continuous over $[\phi,\infty]\setminus\{1\}$ and can takes value infinity when ϕ_s tends to 1 from both sides. Thus, $\mathscr{R}_M(\phi,\phi_s)$ is lower semi-continuous over $[\phi,\infty]$ and continuous on the set $\mathrm{argmin}_{\psi:\psi\geq\phi}\mathscr{R}_M(\phi,\psi) \subseteq [\phi,\infty]\setminus\{1\}$.

Following the discussion after Theorem 10, the uniform risk closeness condition for $k \in \mathcal{K}_n$ holds. Then by Theorem 10, we have that

$$\left(R(\widehat{f}^{\mathrm{cv}}_M;\mathcal{D}_n,\{I_{\widehat{k},\ell}\}_{\ell=1}^M) - \min_{\phi_s\geq\phi}\mathscr{R}^{\mathtt{sub}}_{\lambda,M}(\phi,\phi_s)\right)_+ \xrightarrow{\mathrm{p}} 0.$$

Recall that for any fixed θ, the function

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

is increasing in ϑ. Then $\mathscr{R}^{\mathtt{sub}}_{\lambda,M}(\phi,\phi_s)$ as a function of $\widetilde{v}(-\lambda;\vartheta,\theta)$ through (19) and (22) is also increasing in ϕ for any fixed ϕ_s. For $0<\phi_1\leq\phi_2<\infty$,

$$\min_{\phi_s\geq\phi_1}\mathscr{R}^{\mathtt{sub}}_{\lambda,M}(\phi_1,\phi_s) \leq \min_{\phi_s\geq\phi_2}\mathscr{R}^{\mathtt{sub}}_{\lambda,M}(\phi_1,\phi_s) \leq \min_{\phi_s\geq\phi_2}\mathscr{R}^{\mathtt{sub}}_{\lambda,M}(\phi_2,\phi_s)$$

where the first inequality follows because $\{\phi_s : \phi_s \geq \phi_1\} \supseteq \{\phi_s : \phi_s \geq \phi_2\}$, and the second inequality follows because $\mathscr{R}^{\mathtt{sub}}_{\lambda,M}(\phi,\phi_s)$ is increasing in ϕ for a fixed ϕ_s. Thus, $\min_{\phi_s\geq\phi}\mathscr{R}^{\mathtt{sub}}_{\lambda,M}(\phi,\phi_s)$ is a monotonically increasing function in ϕ. ■

F.5 Proof of Proposition 12 (optimal subagging versus optimal splagging)

Proof Recall that

$$\mathscr{R}^{\mathtt{sub}}_{\lambda,M}(\phi,\phi_s) = \frac{1}{M}(B_\lambda(\phi_s,\phi_s)+V_\lambda(\phi_s,\phi_s)) + \left(1-\frac{1}{M}\right)(B_\lambda(\phi,\phi_s)+V_\lambda(\phi,\phi_s)), \qquad M\in\mathbb{N}$$

$$\mathscr{R}^{\mathtt{spl}}_{\lambda,M}(\phi,\phi_s) = \frac{1}{M}(B_\lambda(\phi_s,\phi_s)+V_\lambda(\phi_s,\phi_s)) + \left(1-\frac{1}{M}\right)C_\lambda(\phi_s), \qquad M=1,\ldots,\lfloor\frac{n}{k}\rfloor.$$

From Proposition 7, we have that

$$\begin{aligned}\mathscr{R}^{\mathtt{sub}}_{\lambda,M}(\phi,\phi_s) &\geq \mathscr{R}^{\mathtt{sub}}_{\lambda,\infty}(\phi,\phi_s)\\ &= B_\lambda(\phi,\phi_s)+V_\lambda(\phi,\phi_s)\\ &= \rho^2(1+\widetilde{v}(-\lambda;\phi,\phi_s))\widetilde{c}(-\lambda;\phi_s) + \sigma^2\widetilde{v}(-\lambda;\phi,\phi_s)\\ &= \rho^2\widetilde{c}(-\lambda;\phi_s) + \widetilde{v}(-\lambda;\phi,\phi_s))(\rho^2\widetilde{c}(-\lambda;\phi_s)+\sigma^2). \qquad (89)\end{aligned}$$

where $\widetilde{c}(-\lambda;\phi_s) = \int r/(1+v(-\lambda;\phi_s)r)^2 \,\mathrm{d}G(r)$. From Proposition 9, we have that for $M \in \mathbb{N}$,

$$\mathscr{R}^{\texttt{spl}}_{\lambda,M}(\phi,\phi_s) \geq \mathscr{R}^{\texttt{spl}}_{\lambda,\phi_s/\phi}(\phi,\phi_s) = \frac{\phi}{\phi_s}(B_\lambda(\phi_s,\phi_s)+V_\lambda(\phi_s,\phi_s)) + \left(1-\frac{\phi}{\phi_s}\right)C_\lambda(\phi_s). \tag{90}$$

On the other hand,

$$\begin{aligned}\mathscr{R}^{\texttt{spl}}_{\lambda,\phi_s/\phi}(\phi,\phi_s) &= \frac{\phi}{\phi_s}\rho^2(1+\widetilde{v}(-\lambda;\phi_s,\phi_s))\widetilde{c}(-\lambda;\phi_s) + \frac{\phi}{\phi_s}\sigma^2\widetilde{v}(-\lambda;\phi_s,\phi_s)) + \left(1-\frac{\phi}{\phi_s}\right)\rho^2\widetilde{c}(-\lambda;\phi_s)\\ &= \rho^2\widetilde{c}(-\lambda;\phi_s) + \frac{\phi}{\phi_s}\widetilde{v}(-\lambda;\phi_s,\phi_s))(\rho^2\widetilde{c}(-\lambda;\phi_s)+\sigma^2). \end{aligned}\tag{91}$$

Since $v(-\lambda;\phi_s)$ is strictly decreasing in ϕ_s from Lemma 42 and G has nonnegative support from Assumption 5, we have that $\widetilde{c}(-\lambda;\phi_s)$ is nonnegative and increasing in ϕ_s. Also note that

$$\begin{aligned}\frac{\phi}{\phi_s}\widetilde{v}(-\lambda;\phi_s,\phi_s)) &= \frac{\phi\int\frac{r^2}{(1+v(-\lambda;\phi_s)r)^2}\,\mathrm{d}H(r)}{v(-\lambda;\phi_s)^{-2}-\phi_s\int\frac{r^2}{(1+v(-\lambda;\phi_s)r)^2}\,\mathrm{d}H(r)}\\ &\geq \frac{\phi\int\frac{r^2}{(1+v(-\lambda;\phi_s)r)^2}\,\mathrm{d}H(r)}{v(-\lambda;\phi_s)^{-2}-\phi\int\frac{r^2}{(1+v(-\lambda;\phi_s)r)^2}\,\mathrm{d}H(r)}\\ &= \widetilde{v}(-\lambda;\phi,\phi_s). \end{aligned}\tag{92}$$

Suppose that $\phi^* \in \mathrm{argmin}_{\inf_{\phi_s\in[\phi,\infty]}}\mathscr{R}^{\texttt{spl}}_{\lambda,M}(\phi,\phi_s)$, we have

$$\begin{aligned}\inf_{M\in\mathbb{N},\phi_s\in[\phi,\infty]}\mathscr{R}^{\texttt{sub}}_{\lambda,M}(\phi,\phi_s) &= \inf_{\phi_s\in[\phi,\infty]}\mathscr{R}^{\texttt{sub}}_{\lambda,\infty}(\phi,\phi_s)\\ &\leq \mathscr{R}^{\texttt{sub}}_{\lambda,\infty}(\phi,\phi^*)\\ &= \rho^2\widetilde{c}(-\lambda;\phi_s^*) + \widetilde{v}(-\lambda;\phi_s^*,\phi))(\rho^2\widetilde{c}(-\lambda;\phi_s^*)+\sigma^2)\\ &\leq \rho^2\widetilde{c}(-\lambda;\phi_s) + \frac{\phi}{\phi_s}\widetilde{v}(-\lambda;\phi_s,\phi_s))(\rho^2\widetilde{c}(-\lambda;\phi_s)+\sigma^2)\\ &= \mathscr{R}^{\texttt{spl}}_{\lambda,M}(\phi_s,\phi_s^*)\\ &= \inf_{\phi_s\in[\phi,\infty]}\mathscr{R}^{\texttt{spl}}_{\lambda,\phi_s/\phi}(\phi,\phi_s)\\ &\leq \inf_{M\in\mathbb{N},\phi_s\in[\phi,\infty]}\mathscr{R}^{\texttt{spl}}_{\lambda,M}(\phi,\phi_s)\end{aligned}$$

where in the second inequality we use (92) and the last inequality is from (90). ■

F.6 Proof of Proposition 13 (optimal bag size for ridgeless predictors)

Proof The proof of Proposition 13 follows by combining results from Lemma 28 and Lemma 29 for subagged and splagged ridgeless predictors, respectively. ■

Lemma 28 (Optimal risk for subagged ridgeless predictor)**.** *Suppose the conditions in Theorem 6 hold, and $\sigma^2, \rho^2 \geq 0$ are the noise variance and signal strength from Assumptions 2 and 3. Let $\mathtt{SNR} = \rho^2/\sigma^2$. For any $\phi \in (0, \infty)$, the properties of the optimal asymptotic risk $\mathscr{R}^{\mathtt{sub}}_{0,\infty}(\phi, \phi_s^{\mathtt{sub}}(\phi))$ in terms of $\mathtt{SNR}$ and ϕ are characterized as follows:*

(1) $\mathtt{SNR} = 0$ $(\rho^2 = 0, \sigma^2 \neq 0)$*: For all $\phi \geq 0$, the global minimum $\mathscr{R}^{\mathtt{sub}}_{0,\infty}(\phi, \phi_s^{\mathtt{sub}}(\phi)) = \sigma^2$ is obtained with $\phi_s^{\mathtt{sub}}(\phi) = \infty$.*

(2) $\mathtt{SNR} > 0$*: For all $\phi \geq 0$, the global minimum of $\phi_s \mapsto \mathscr{R}^{\mathtt{sub}}_{0,\infty}(\phi, \phi_s)$ is obtained at $\phi_s^{\mathtt{sub}}(\phi) \in (1, \infty)$.*

(3) $\mathtt{SNR} = \infty$ $(\rho^2 \neq 0, \sigma^2 = 0)$*: If $\phi \in (0, 1]$, the global minimum $\mathscr{R}^{\mathtt{sub}}_{0,\infty}(\phi, \phi_s^{\mathtt{sub}}(\phi)) = 0$ is obtained with any $\phi_s^{\mathtt{sub}}(\phi) \in [\phi, 1]$. If $\phi \in (1, \infty)$, then the global minimum $\mathscr{R}^{\mathtt{sub}}_{0,\infty}(\phi, \phi_s^{\mathtt{sub}}(\phi))$ is obtained at $\phi_s^{\mathtt{sub}}(\phi) \in [\phi, \infty)$.*

Proof From Theorem 6, the limiting risk for bagged ridgeless with $M = \infty$ is given by

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

Defined in (70)-(71), $\widetilde{v}(0; \phi, \phi_s) \geq 0$ and $\widetilde{c}(0; \phi_s) \geq 0$ are continuous functions of $v(0; \phi_s)$, which is strictly decreasing over $\phi_s \in (1, \infty)$ and satisfies $\lim_{\phi_s \to \infty} v(0; \phi_s) = 0$ from Lemma 43. Then we have $\widetilde{v}(0; \phi, \phi_s)$ is decreasing in ϕ_s over $(1, \infty)$, $\widetilde{c}(0; \phi_s)$ is increasing in ϕ_s over $(1, \infty)$, and

$$\lim_{\phi_s \to \infty} \widetilde{v}(0; \phi, \phi_s) = 0, \qquad \lim_{\phi_s \to \infty} \widetilde{c}(0; \phi_s) = \int r \, \mathrm{d}G(r).$$

Also, $\widetilde{v}(0; \phi, \phi_s) = \phi/(1 - \phi)$ and $\widetilde{c}(0; \phi_s) = 0$ remain constant for $\phi_s \in (0, 1]$ from (24). Then to determine the global minimum, it suffices to consider the case when $\phi_s \in [1, \infty)$. Next, we consider various cases depending on the value of $\mathtt{SNR}$.

- First, consider the case $\mathtt{SNR} > 0$. We consider further sub-cases depending the value of the pair (ϕ, ϕ_s).

  1. When $\phi \in (0, 1)$ and $\phi_s \in (1, \infty]$,

$$\begin{aligned}
\frac{\partial \mathscr{R}^{\mathtt{sub}}_{0,\infty}(\phi, \phi_s)}{\partial \phi_s} &= \frac{\partial \mathscr{R}^{\mathtt{sub}}_{0,\infty}(\phi, \phi_s)}{\partial v(0; \phi_s)} \frac{\partial v(0; \phi_s)}{\partial \phi_s} \\
&= \rho^2 \frac{\phi \int \frac{v(0;\phi_s) r^2}{(1+v(0;\phi_s)r)^3} \, \mathrm{d}H(r)}{\left(1 - \phi \int \left(\frac{v(0;\phi_s)r}{(1+v(0;\phi_s)r)}\right)^2 \mathrm{d}H(r)\right)^2} \int \frac{r}{(1 + v(0; \phi_s) r)^2} \, \mathrm{d}G(r) \cdot \frac{\partial v(0; \phi_s)}{\partial \phi_s} \\
&\quad - 2\rho^2 \frac{\frac{1}{v(0;\phi_s)^2}}{\frac{1}{v(0;\phi_s)^2} - \phi \int \frac{r^2}{(1+v(0;\phi_s)r)^2} \, \mathrm{d}H(r)} \int \frac{r^2}{(1 + v(0; \phi_s) r)^3} \, \mathrm{d}G(r) \cdot \frac{\partial v(0; \phi_s)}{\partial \phi_s} \\
&\quad + \sigma^2 \frac{\phi \int \frac{v(0;\phi_s) r^2}{(1+v(0;\phi_s)r)^3} \, \mathrm{d}H(r)}{\left(1 - \phi \int \left(\frac{v(0;\phi_s)r}{(1+v(0;\phi_s)r)}\right)^2 \mathrm{d}H(r)\right)^2} \cdot \frac{\partial v(0; \phi_s)}{\partial \phi_s}.
\end{aligned}$$

Note that from Lemma 42, $v(0;\phi_s)$ is differentiable in $\phi_s \in (0,\infty]$ with

$$\frac{\partial v(0;\phi_s)}{\partial \phi_s} = -\frac{\int \frac{r}{1+v(0;\phi_s)r}\,\mathrm{d}H(r)}{\frac{1}{v(0;\phi_s)^2} - \phi_s \int \frac{r^2}{(1+v(0;\phi_s)r)^2}\,\mathrm{d}H(r)}$$

being negative over $\phi_s \in (1,\infty)$ and continuous in $\phi_s \in (1,\infty]$, and

$$\lim_{\phi_s\to 1^+} \frac{\partial v(0;\phi_s)}{\partial \phi_s} = -\infty, \qquad \lim_{\phi_s\to \infty} \frac{\partial v(0;\phi_s)}{\partial \phi_s} = -\lim_{\phi_s\to\infty} \widetilde{v}_v(0;\phi_s) \int \frac{r}{1+v(0;\phi_s)r}\,\mathrm{d}H(r) = 0$$

by Lemma 43 with $\widetilde{v}_v$ defined in (111). We have that $\partial \mathscr{R}^{\mathsf{sub}}_{0,\infty}(\phi,\phi_s)/\partial\phi_s$ is continuous over $\phi_s \in (1,\infty]$. Since $\lim_{\phi_s\to\infty} v(0;\phi_s) = 0$ from Lemma 43, we have that

$$\lim_{\phi_s\to\infty} \frac{\phi \int \frac{v(0;\phi_s)r^2}{(1+v(0;\phi_s)r)^3}\,\mathrm{d}H(r)}{\left(1 - \phi\int \left(\frac{v(0;\phi_s)r}{(1+v(0;\phi_s)r)}\right)^2 \mathrm{d}H(r)\right)^2} = 0 \tag{93}$$

$$\lim_{\phi_s\to\infty} \frac{\frac{1}{v(0;\phi_s)^2}}{\frac{1}{v(0;\phi_s)^2} - \phi\int \frac{r^2}{(1+v(0;\phi_s)r)^2}\,\mathrm{d}H(r)} \int \frac{r^2}{(1+v(0;\phi_s)r)^3}\,\mathrm{d}G(r) = \frac{1}{1-\phi}\int r^2\,\mathrm{d}G(r) > 0. \tag{94}$$

Since $\partial v(0;\phi_s)/\partial\phi_s$ is negative over $(1,\infty)$ and $\lim_{\phi_s\to\infty 0} \partial v(0;\phi_s)/\partial\phi_s = 0$, we have

$$\frac{\partial \mathscr{R}^{\mathsf{sub}}_{0,\infty}(\phi,\phi_s)}{\partial\phi_s}\Big|_{\phi_s=\infty} = -2\rho^2 \int r^2\,\mathrm{d}G(r) \cdot \lim_{\phi_s\to\infty} \frac{\partial v(0;\phi_s)}{\partial\phi_s} = 0. \tag{95}$$

Combining (93)-(95), we have that when ϕ_s is large, $\partial \mathscr{R}^{\mathsf{sub}}_{0,\infty}(\phi,\phi_s)/\partial\phi_s$ approaching zero from above as ϕ_s tends to ∞. On the other hand, since for $k=1,2$,

$$\begin{aligned}&\lim_{\phi_s\to 1^+} \int \frac{r^k}{(1+v(0;\phi_s)r)^{k+1}}\,\mathrm{d}G(r) \cdot \frac{\partial v(0;\phi_s)}{\partial\phi_s} \\ &= \lim_{\phi_s\to 1^+} \int \frac{v(0;\phi_s)r^k}{(1+v(0;\phi_s)r)^{k+1}}\,\mathrm{d}G(r) \cdot \lim_{\phi_s\to 1^+} \frac{1}{v(0;\phi_s)} \frac{\partial v(0;\phi_s)}{\partial\phi_s} = 0,\end{aligned}$$

we have

$$\frac{\partial \mathscr{R}^{\mathsf{sub}}_{0,\infty}(\phi,\phi_s)}{\partial\phi_s}\Big|_{\phi_s=1^+} = \sigma^2 \frac{\phi}{1-\phi} \cdot \lim_{\phi_s\to 1^+} \frac{\partial v(0;\phi_s)}{\partial\phi_s} < 0.$$

Thus, there exists $\phi^* \in (1,\infty)$ such that

$$\mathscr{R}^{\mathsf{sub}}_{0,\infty}(\phi,\phi^*) < \mathscr{R}^{\mathsf{sub}}_{0,\infty}(\phi,1) = \mathscr{R}^{\mathsf{sub}}_{0,\infty}(\phi,\phi).$$

2. When $\phi = 1$, $\mathscr{R}^{\mathsf{sub}}_{0,\infty}(1,1) = \infty$ while $\mathscr{R}^{\mathsf{sub}}_{0,\infty}(\phi,\phi_s) < \infty$ for all $\phi_s \in (1,\infty]$. Since $\mathscr{R}^{\mathsf{sub}}_{0,\infty}(\phi,\phi_s)$ is continuous and finite in $(1,\infty]$, by continuity and (93)-(95) we have $\phi^* \in (1,\infty)$.
3. When $\phi \in (1,\infty)$, the optimal $\phi^* \geq \phi > 1$ must be obtained in $[\phi,\infty)$ because of (93)-(95).

- Next, consider the case when $\texttt{SNR} = 0$, i.e., $\rho^2 = 0$ and $\sigma^2 \neq 0$, since $\mathscr{R}^{\texttt{sub}}_{0,\infty}(\phi, \phi_s) = \sigma^2 + \sigma^2 \widetilde{v}(0; \phi, \phi_s) > 0$ is increasing in $v(0; \phi_s)$ and $v(0; \phi_s) \geq 0$ is decreasing in ϕ_s, we have that $\mathscr{R}^{\texttt{sub}}_{0,\infty}(\phi, \phi_s)$ is decreasing in ϕ_s. Thus, the global minimum $\mathscr{R}^{\texttt{sub}}_{0,\infty}(\infty, \phi_s) = \sigma^2$ is obtained at $\phi_s^* = \infty$.
- Finally, consider the case when $\texttt{SNR} = \infty$, i.e. $\rho^2 \neq 0$ and $\sigma^2 = 0$, $\mathscr{R}^{\texttt{sub}}_{0,\infty}(\phi, \phi_s) = \rho^2(1 + \widetilde{v}(0; \phi, \phi_s))\widetilde{c}(0; \phi_s)$. As the bias term is zero when $\phi_s \in (0, 1]$ and positive when $\phi_s \in (1, \infty]$, we have that $\mathscr{R}^{\texttt{sub}}_{0,\infty}(\phi, \phi_s) \geq \mathscr{R}^{\texttt{sub}}_{0,\infty}(\phi, \phi^*) = 0$ for all $\phi_s^* \in [\phi, 1]$ when $\phi \in (0, 1]$. If $\phi \in (1, \infty)$, since the risk is continuous over $[\phi, \infty]$, the global minimum exists. Since the derivative $\partial \mathscr{R}^{\texttt{sub}}_{0,\infty}(\phi, \phi_s)/\partial \phi_s$ is continuous over $\phi_s \in (1, \infty]$ and (93)-(95), the minimizer satisfies $\phi^* \in [\phi, \infty)$.

■

Lemma 29 (Optimal splagged ridgeless)**.** *Suppose the conditions in Theorem 8 hold, and $\sigma^2, \rho^2 \geq 0$ are the noise variance and signal strength from Assumptions 2 and 3. Let $\texttt{SNR} = \rho^2/\sigma^2$. For any $\phi \in (0, \infty)$, the properties of the optimal asymptotic risk $\mathscr{R}^{\texttt{spl}}_{0,\infty}(\phi, \phi_s^{\texttt{spl}}(\phi))$ in terms of $\texttt{SNR}$ and ϕ are characterized as follows:*

(1) $\texttt{SNR} = 0$ $(\rho^2 = 0, \sigma^2 \neq 0)$*: For all $\phi \geq 0$, the global minimum $\mathscr{R}^{\texttt{spl}}_{0,\infty}(\phi, \phi_s^{\texttt{spl}}(\phi)) = \sigma^2$ is obtained with $\phi_s^{\texttt{spl}}(\phi) = \infty$.*

(2) $\texttt{SNR} > 0$*: For $\phi \geq 1$, there exists global minimum of $\phi_s \mapsto \mathscr{R}^{\texttt{spl}}_{0,\infty}(\phi, \phi_s)$ in $(1, \infty)$. For $\phi \in (0, 1)$, the global minimum is in $\{\phi\} \cup (1, \infty)$.*

(3) $\texttt{SNR} = \infty$ $(\rho^2 \neq 0, \sigma^2 = 0)$*: If $\phi \in (0, 1]$, the global minimum $\mathscr{R}^{\texttt{spl}}_{0,\infty}(\phi, \phi_s^{\texttt{spl}}(\phi)) = 0$ is obtained with any $\phi_s^{\texttt{spl}}(\phi) \in [\phi, 1]$. If $\phi \in (1, \infty)$, then the global minimum $\mathscr{R}^{\texttt{spl}}_{0,\infty}(\phi, \phi_s^{\texttt{spl}}(\phi))$ is obtained at $\phi_s^{\texttt{spl}}(\phi) \in [\phi, \infty)$.*

Proof From Theorem 8, the limiting risk for bagged ridgeless with $M = \phi_s/\phi$ is given by

$$\begin{aligned}\mathscr{R}^{\texttt{spl}}_{0,\phi_s/\phi}(\phi, \phi_s) &= \sigma^2 + \frac{\phi}{\phi_s}\left[\rho^2(1 + \widetilde{v}(0; \phi_s, \phi_s))\widetilde{c}(0; \phi_s) + \sigma^2 \widetilde{v}(0; \phi_s, \phi_s)\right] + \left(1 - \frac{\phi}{\phi_s}\right)\rho^2 \widetilde{c}(0; \phi_s)\\ &= \sigma^2 + \rho^2 \widetilde{c}(0; \phi_s) + \phi \frac{\widetilde{v}(0; \phi_s, \phi_s)}{\phi_s}(\rho^2 \widetilde{c}(0; \phi_s) + \sigma^2).\end{aligned}$$

Defined in (70)-(71), $\widetilde{v}(0; \phi_s, \phi_s) \geq 0$ and $\widetilde{c}(0; \phi_s) \geq 0$ are continuous functions of $v(0; \phi_s)$, which is strictly decreasing over $\phi_s \in (1, \infty)$ and satisfies $\lim_{\phi_s \to \infty} v(0; \phi_s) = 0$ from Lemma 43. Then $\widetilde{c}(0; \phi_s)$ is increasing in ϕ_s over $(1, \infty)$ and $\lim_{\phi_s \to \infty} \widetilde{c}(0; \phi_s) = \int r \, \mathrm{d}G(r)$.

- We first consider the case $\texttt{SNR} > 0$. We consider further sub-cases depending the value of the pair (ϕ, ϕ_s).

  1. When $\phi \in (0, 1)$ and $\phi_s \in (1, \infty]$,
     Define functions h_1 and h_2 as follows:

$$h_1(\phi_s) = \texttt{SNR} \cdot \widetilde{c}(0; \phi_s), \qquad h_2(\phi_s) = \frac{\widetilde{v}(0; \phi_s, \phi_s)}{\phi_s} = \widetilde{v}_v(0; \phi_s) \int \left(\frac{r}{1 + v(0; \phi_s) r}\right)^2 \mathrm{d}H(r), \tag{96}$$

where $\widetilde{v}_v$ is defined in (111). Then $\mathscr{R}^{\texttt{spl}}_{0,\phi_s/\phi}(\phi,\phi_s) = \sigma^2 + \sigma^2(h_1(\phi_s) + \phi h_2(\phi_s)(1 + h_1(\phi_s)))$, with h_1 increasing in ϕ_s and

$$\lim_{\phi_s\to 1^+} h_1(\phi_s) = 0, \quad \lim_{\phi_s\to\infty} h_1(\phi_s) = \texttt{SNR}\int r\,\mathrm{d}G(r), \quad \lim_{\phi_s\to 1^+} h_2(\phi_s) = +\infty, \quad \lim_{\phi_s\to\infty} h_2(\phi_s) = 0.$$

Next, we study the property of h_2. Simple calculation yields that

$$\begin{aligned}
&\frac{\partial h_2(\phi_s)}{\partial \phi_s}\\
&= \widetilde{v}_v(0;\phi_s)^2\left[\frac{2}{v(0;\phi_s)^3}\int \frac{r^2}{(1+v(0;\phi_s)r)^3}\,\mathrm{d}H(r)\cdot\frac{\partial v(0;\phi_s)}{\partial\phi_s} + \left(\int \frac{r^2}{(1+v(0;\phi_s)r)^2}\,\mathrm{d}H(r)\right)^2\right]\\
&= \widetilde{v}_v(0;\phi_s)^2\left[\left(\int \frac{r^2}{(1+v(0;\phi_s)r)^2}\,\mathrm{d}H(r)\right)^2 - \right.\\
&\qquad \left.\frac{2\widetilde{v}_v(0;\phi_s)}{v(0;\phi_s)^3}\int \frac{r^2}{(1+v(0;\phi_s)r)^3}\,\mathrm{d}H(r)\int\frac{r}{1+v(0;\phi_s)r}\,\mathrm{d}H(r)\right].
\end{aligned}$$

From Lemma 43 (4), we have that $\lim_{\phi_s\to\infty}\widetilde{v}_v(0;\phi_s)/v(0;\phi_s)^2\lim_{\phi_s\to\infty}[1+\widetilde{v}_b(0;\phi_s)] = 1$ where $\widetilde{v}_b(0;\phi_s)$ is defined in Lemma 43. Analogously, $\lim_{\phi_s\to 1^+}\widetilde{v}_v(0;\phi_s)/v(0;\phi_s)^2 = +\infty$. Then as in the proof of Proposition 13, one can verify that

$$\begin{aligned}
\frac{\partial \mathscr{R}^{\texttt{spl}}_{0,\phi_s/\phi}(\phi,\phi_s)}{\partial\phi_s} &= -\sigma^2\widetilde{v}_v(0;\phi_s)\left[\texttt{SNR}(1+\phi h_2(\phi_s))\int\frac{r^2}{(1+v(0;\phi_s)r)^3}\,\mathrm{d}G(r)\cdot\int\frac{r}{1+v(0;\phi_s)r}\,\mathrm{d}H(r)\right.\\
&\quad + \phi(1+h_1(\phi_s))\frac{2\widetilde{v}_v(0;\phi_s)^2}{v(0;\phi_s)^3}\int\frac{r^2}{(1+v(0;\phi_s))^3}\,\mathrm{d}H(r)\cdot\int\frac{r}{1+v(0;\phi_s)r}\,\mathrm{d}H(r)\\
&\quad \left. -\widetilde{v}_v(0;\phi_s)\phi(1+h_1(\phi_s))\left(\int\frac{r^2}{(1+v(0;\phi_s)r)^2}\,\mathrm{d}H(r)\right)^2\right]
\end{aligned}$$

satisfies $\lim_{\phi_s\to 1^+}\partial\mathscr{R}^{\texttt{spl}}_{0,\phi_s/\phi}(\phi,\phi_s)/\partial\phi_s = -\infty$ and $\lim_{\phi_s\to\infty}\partial\mathscr{R}^{\texttt{spl}}_{0,\phi_s/\phi}(\phi,\phi_s)/\partial\phi_s = 0$ by utilizing properties in Lemma 43. Furthermore, as

$$\lim_{\phi_s\to\infty}\widetilde{v}_v(0;\phi_s)^{-1}\frac{\partial\mathscr{R}^{\texttt{spl}}_{0,\phi_s/\phi}(\phi,\phi_s)}{\partial\phi_s} = -\rho^2\int r^2\,\mathrm{d}G(r)\cdot\int r\,\mathrm{d}H(r) < 0, \tag{97}$$

we have that when ϕ_s is large, $\partial\mathscr{R}^{\texttt{spl}}_{0,\phi_s/\phi}(\phi,\phi_s)/\partial\phi_s$ approaching zero from above as ϕ_s tends to ∞. Thus, the minimum of $\mathscr{R}^{\texttt{spl}}_{0,\phi_s/\phi}(\phi,\phi_s)$ over $[1,\infty]$ is obtained in the open interval $(1,\infty)$.

2. When $\phi < 1$ and $\phi_s \in [\phi, 1)$, since the term $\widetilde{c}(0;\phi_s)$ is zero, $\mathscr{R}^{\texttt{spl}}_{0,\phi_s/\phi}(\phi,\phi_s) = \sigma^2 + \sigma^2\phi(1-\phi_s)^{-1}$ is increasing in ϕ_s. So the minimum over $[\phi,1]$ is obtained at $\phi_s = \phi$.
3. When $\phi = 1$, $\mathscr{R}^{\texttt{spl}}_{0,\phi_s/\phi}(1,1) = \infty$ while $\mathscr{R}^{\texttt{spl}}_{0,\phi_s/\phi}(\phi,\phi_s) < \infty$ for all $\phi_s\in(1,\infty]$. Since $\mathscr{R}^{\texttt{spl}}_{0,\phi_s/\phi}(\phi,\phi_s)$ is continuous and finite in $(1,\infty]$, by continuity and (97) we have $\phi_s^*\in(1,\infty)$.

4. When $\phi \in (1, \infty)$, the optimal $\phi_s^* \geq \phi > 1$ must be obtained in $[\phi, \infty)$ because of (97).

- Next consider the case when $\mathtt{SNR} = 0$, i.e., $\rho^2 = 0$ and $\sigma^2 \neq 0$. Then $h_1 \equiv 0$ and $\mathscr{R}_{0,\infty}^{\mathtt{spl}}(\phi_s/\phi, \phi_s) = \sigma^2 + \sigma^2 \phi \widetilde{v}(0; \phi_s, \phi_s)/\phi_s$. When $\phi_s \in (0, 1)$, $\widetilde{v}(0; \phi_s, \phi_s)/\phi_s = (1 - \phi_s)^{-1}$ is increasing in ϕ_s; when $\phi_s > 1$, $\widetilde{v}(0; \phi_s, \phi_s)/\phi_s \geq 0 = \lim_{\phi_s \to \infty} \widetilde{v}(0; \phi_s, \phi_s)/\phi_s = 0$. Therefore, the global minimum $\mathscr{R}_{0,\infty}^{\mathtt{sub}}(\infty, \phi_s) = \sigma^2$ is obtained at $\phi_s^* = \infty$.
- Finally, consider the case when $\mathtt{SNR} = \infty$, i.e. $\rho^2 \neq 0$ and $\sigma^2 = 0$, $\mathscr{R}_{0,\infty}^{\mathtt{spl}}(\phi_s/\phi, \phi_s) = \rho^2 \widetilde{c}(0; \phi_s) + \rho^2 \phi \phi_s^{-1} \widetilde{v}(0; \phi, \phi_s) \widetilde{c}(0; \phi_s)$. As the term $\widetilde{c}(0; \phi_s)$ is zero when $\phi_s \in (0, 1]$ and positive when $\phi_s \in (1, \infty]$, we have that $\mathscr{R}_{0,\infty}^{\mathtt{sub}}(\phi, \phi_s) \geq \mathscr{R}_{0,\infty}^{\mathtt{sub}}(\phi, \phi_s^*) = 0$ for all $\phi_s^* \in [\phi, 1]$ when $\phi \in (0, 1]$. If $\phi \in (1, \infty)$, since the risk is continuous over $[\phi, \infty]$, the global minimum exists. Since the derivative $\partial \mathscr{R}_{0,\infty}^{\mathtt{sub}}(\phi, \phi_s)/\partial \phi_s$ is continuous over $\phi_s \in (1, \infty]$ and (97), the minimizer satisfies $\phi_s^* \in [\phi, \infty)$.

■

Appendix G. Proofs in Section 5 (isotropic features)

G.1 Proof of Corollary 14 (bagged risk for ridgeless regression)

Proof Since $\boldsymbol{\Sigma} = \boldsymbol{I}_p$, we have that $\mathrm{d}G = \mathrm{d}H = \delta_1$. Then, $v(0; \phi_s)$, $\widetilde{v}(0; \phi, \phi_s)$ and $\widetilde{c}(0; \phi_s)$ defined in (70) and (71) for $\phi_s > 1$ reduce to

$$v(0; \phi_s) = \frac{1}{\phi_s - 1}, \qquad \widetilde{v}(0; \phi, \phi_s) = \frac{\phi}{\phi_s^2 - \phi}, \qquad \widetilde{c}(0; \phi_s) = \frac{(\phi_s - 1)^2}{\phi_s^2}.$$

Thus, we have

$$B_0(\phi, \phi_s) = \begin{cases} 0, & \phi_s \in (0, 1) \\ \rho^2 \dfrac{\phi_s - 1}{\phi_s}, & \phi_s \in (1, \infty) \end{cases}, \qquad V_0(\phi, \phi_s) = \begin{cases} \sigma^2 \dfrac{\phi_s}{1 - \phi_s}, & \phi_s \in (0, 1) \\ \sigma^2 \dfrac{1}{\phi_s - 1}, & \phi_s \in (1, \infty) \end{cases},$$

and

$$C_0(\phi_s) = \begin{cases} 0, & \phi_s \in (0, 1) \\ \rho^2 \dfrac{(\phi_s - 1)^2}{\phi_s^2}, & \phi_s \in (1, \infty) \end{cases}.$$

■

From Corollary 14, we are able to derive the asymptotic bias and variance for $M = 1$ and $M = \infty$ for ridgeless regression with replacement:

$$\mathscr{B}_{0,1}^{\mathtt{sub}}(\phi, \phi_s) = \begin{cases} 0, & \phi_s \in (0, 1) \\ \rho^2 \dfrac{\phi_s - 1}{\phi_s}, & \phi_s \in (1, \infty) \end{cases} \qquad \mathscr{V}_{0,1}^{\mathtt{sub}}(\phi, \phi_s) = \begin{cases} \sigma^2 \dfrac{\phi_s}{1 - \phi_s}, & \phi_s \in (0, 1) \\ \sigma^2 \dfrac{1}{\phi_s - 1}, & \phi_s \in (1, \infty) \end{cases}$$

$$\mathscr{B}^{\texttt{sub}}_{0,\infty}(\phi,\phi_s) = \begin{cases} 0, & \phi_s \in (0,1) \\ \rho^2 \dfrac{(\phi_s-1)^2}{\phi_s^2-\phi}, & \phi_s \in (1,\infty) \end{cases} \quad \mathscr{V}^{\texttt{sub}}_{0,\infty}(\phi,\phi_s) = \begin{cases} \sigma^2 \dfrac{\phi}{1-\phi}, & \phi_s \in (0,1) \\ \sigma^2 \dfrac{\phi}{\phi_s^2-\phi}, & \phi_s \in (1,\infty) \end{cases}.$$

Then the asymptotic bias and variance for general M would be convex combinations of the above quantities.

On the other hand, the asymptotic bias and variance for splagging without replacement are given by

$$\mathscr{B}^{\texttt{spl}}_{\lambda,M}(\phi,\phi_s) = M^{-1}B_\lambda(\phi_s,\phi_s) + (1-M^{-1})C_\lambda(\phi_s), \qquad \mathscr{V}^{\texttt{spl}}_{\lambda,M}(\phi,\phi_s) = M^{-1}V_\lambda(\phi_s,\phi_s).$$

G.2 Proof of Proposition 15 (optimal subagged ridgeless regression with replacement)

Proof For $\phi \in (0,1)$ and $\phi_s \in (1,\infty]$, we have that

$$\mathscr{R}^{\texttt{sub}}_{0,\infty}(\phi,\phi_s) = \sigma^2 + \rho^2 \frac{(\phi_s-1)^2}{\phi_s^2-\phi} + \sigma^2 \frac{\phi}{\phi_s^2-\phi}.$$

Taking the derivative of the right hand side with respect to ϕ_s

$$\frac{\partial \mathscr{R}^{\texttt{sub}}_{0,\infty}(\phi,\phi_s)}{\partial \phi_s} = 2\sigma^2 \frac{\texttt{SNR}(\phi_s-1)(\phi_s-\phi) - \phi\phi_s}{(\phi_s^2-\phi)^2}$$

and setting it to zero yields that

$$\phi_s = A \pm \sqrt{A^2-\phi}. \tag{98}$$

where $A = (\phi + 1 + \phi/\texttt{SNR})/2$. Since $A - \sqrt{A^2-\phi} < \sqrt{\phi} \leq 1$, we have $\phi_s^* = A + \sqrt{A^2-\phi}$ is a minimizer and

$$\begin{aligned} \mathscr{R}^{\texttt{sub}}_{0,\infty}(\phi,\phi^*) &= \sigma^2 + \sigma^2 \frac{\phi + A - \sqrt{A^2-\phi} + \texttt{SNR}(1-\phi)/\phi(A-\phi-\sqrt{A^2-\phi})}{2\sqrt{A^2-\phi}} \\ &= \frac{\sigma^2}{2}\left[1 + \frac{\phi-1}{\phi}\texttt{SNR} + \frac{2\texttt{SNR}}{\phi}\sqrt{A^2-\phi}\right] \\ &= \frac{\sigma^2}{2}\left[1 + \frac{\phi-1}{\phi}\texttt{SNR} + \sqrt{\left(1 - \frac{\phi-1}{\phi}\texttt{SNR}\right)^2 + 4\texttt{SNR}}\right], \end{aligned} \tag{99}$$

which gives the simplified formula. Note that

$$\begin{aligned} \mathscr{R}^{\texttt{sub}}_{0,\infty}(\phi,\phi^*) &= \sigma^2 + \sigma^2 \left(\frac{\phi}{2\sqrt{A^2-\phi}} + \frac{A-\sqrt{A^2-\phi}}{2\sqrt{A^2-\phi}} + \frac{1-\phi}{\phi}\texttt{SNR}\frac{A-\phi-\sqrt{A^2-\phi}}{2\sqrt{A^2-\phi}}\right) \\ &= \sigma^2 + \sigma^2 h(\texttt{SNR}) - \sigma^2 \delta(\texttt{SNR}) \end{aligned} \tag{100}$$

where for all $r \geq 0$, the functions h and δ are defined as $h(r) = h_1(r) + h_2(r) + h_3(r)$ and $\delta(r) = (1-\phi)rh_1(r)/\phi$, with $A(r) = (\phi + 1 + \phi/r)/2$ and

$$h_1(r) = \frac{\phi}{2\sqrt{A(r)^2-\phi}},$$

$$h_2(r) = \frac{A(r) - \sqrt{A(r)^2 - \phi}}{2\sqrt{A(r)^2 - \phi}} = \frac{1}{2\sqrt{1 - \phi/A(r)^2}} - \frac{1}{2},$$
$$h_3(r) = \frac{1-\phi}{\phi} r h_2(r).$$

Since h_1, h_2, and h_3 are nonngative over $(0, \infty)$, h and δ are also nonnegative. Also noted that

$$\delta(0) = \frac{1-\phi}{\phi} \lim_{r \to 0^+} r h_1(r) = 0, \qquad \delta(\infty) = \frac{1-\phi}{\phi} \lim_{r \to +\infty} r h_1(r) = +\infty,$$

we obtain the upper bound for $\mathscr{R}^{\mathtt{sub}}_{0,\infty}(\phi, \phi^*)$ as follows:

$$\mathscr{R}^{\mathtt{sub}}_{0,\infty}(\phi, \phi^*) \le \sigma^2 + \sigma^2 h(\mathtt{SNR}), \tag{101}$$

with equality obtained if and only if $\mathtt{SNR} = 0$.

Next we analyze the function $h(r)$. Note that $A(r) > 0$ is decreasing in r, we have that the functions h_1 and h_2 are nonnegative and monotone increasing in $\mathtt{SNR}$. Hence h_3 as the product of nonnegative and monotone increasing functions, is also nonnegative and monotone increasing in $\mathtt{SNR}$. Thus, h is monotone increasing in $\mathtt{SNR}$ and

$$\begin{aligned}
h(\mathtt{SNR}) &\le \lim_{r \to \infty} h(r) \\
&= \lim_{r \to \infty} h_1(r) + \lim_{r \to \infty} h_2(r) + \lim_{r \to \infty} r h_2(r) \\
&= \frac{\phi}{1-\phi} + \frac{\phi}{1-\phi} + \frac{1}{\phi} \lim_{r \to \infty} \frac{A(r) - \sqrt{A(r)^2 - \phi}}{\frac{1}{r}} \\
&= \frac{\phi}{1-\phi} + \frac{\phi}{1-\phi} + \frac{1}{\phi} \lim_{r \to \infty} \frac{-\frac{\phi}{2r^2} + \frac{\frac{A(r)\phi}{r^2}}{2\sqrt{A(r)^2 - \phi}}}{-\frac{1}{r^2}} \\
&= \frac{\phi}{1-\phi},
\end{aligned}$$

where the third equality is due to the L'Hospital's rule. Note that the risk for $\phi_s \in [\phi, 1)$ is given by $\sigma^2 + \sigma^2 \phi/(1-\phi)$, we have that ϕ_s^* obtained the global minimum of $\mathscr{R}^{\mathtt{sub}}_{0,\infty}(\phi, \phi_s)$ over $\phi_s \in [\phi, \infty]$.

For $\phi \in [1, \infty)$ and $\phi_s \in [\phi, \infty)$, from (98) and $A - \sqrt{A^2 - \phi} \le \sqrt{\phi} \le \phi$, we have again $\phi_s^* = A + \sqrt{A^2 - \phi}$ is a minimizer.

When $\mathtt{SNR} = 0$, since the bias term is zero and variance term is increasing over $\phi_s < 1$ and increasing over $\phi_s > 1$, we have that when $\phi_s > 1$ (whenever $\phi \le \phi_s$),

$$\mathscr{R}^{\mathtt{sub}}_{0,\infty}(\phi, \phi_s) = \sigma^2 + V_0(\phi, \phi_s) \ge \sigma^2 + V_0(\phi, \infty) = \sigma^2.$$

When $\phi < 1$, we have $\mathscr{R}^{\mathtt{sub}}_{0,\infty}(\phi, \phi_s) \ge \mathscr{R}^{\mathtt{sub}}_{0,\infty}(\phi, \phi) = \sigma^2/(1-\phi) > \sigma^2$. Therefore, $\mathscr{R}^{\mathtt{sub}}_{0,\infty}(\phi, \phi_s) \ge \mathscr{R}^{\mathtt{sub}}_{0,\infty}(\phi, \infty) = \sigma^2$ for all $\phi \in (1, \infty]$.

When $\mathtt{SNR} = \infty$, the variance term $V_0(\phi, \phi_s) = 0$ for all $\phi_s \in [\phi, \infty]$. If $\phi \in (0, 1]$, then $B_0(\phi, \phi_s) = 0$ for all $\phi_s \in [\phi, 1]$. If $\phi \in (1, \infty)$, then $B_0(\phi, \phi_s)$ is increasing over $\phi_s \in [\phi, \infty]$. Hence, the conclusions follow. ■

G.3 Proof of Theorem 16 (comparison between subagged and optimal ridge regression)

Proof As $n, p \to \infty$ and $p/n \to \phi$, the optimal regularization parameter is given by $\lambda^* = \phi\sigma^2/\rho^2$ under the isotopic model (Dobriban and Wager, 2018). The limiting risk of the optimal ridge regression is given by

$$\mathscr{R}^{\mathtt{WR}}_{\lambda^*,1}(\phi,\phi) = \frac{\sigma^2}{2}\left[1 + \frac{\phi-1}{\phi}\mathtt{SNR} + \sqrt{\left(1 - \frac{\phi-1}{\phi}\mathtt{SNR}\right)^2 + 4\mathtt{SNR}}\right]$$

which is the same the formula given in Proposition 15. The conclusion thus follows. ■

Fixed-point equation details for ridge regression

For isotopic features $\mathbf{\Sigma} = \boldsymbol{I}_p$, $\mathrm{d}G = \mathrm{d}H = \delta_1$. When $n, p \to$ and $p/n \to \phi \in (0, \infty)$, (109)-(111) reduce to

$$\begin{aligned}
v(-\lambda;\phi)^{-1} &= \lambda + \phi(1 + v(-\lambda;\phi))^{-1} \\
\widetilde{v}_b(-\lambda;\phi) &= \frac{\phi(1+v(-\lambda;\phi))^{-2}}{v(-\lambda;\phi)^{-2} - \phi(1+v(-\lambda;\phi))^{-2}} \\
\widetilde{v}_v(-\lambda;\phi)^{-1} &= v(-\lambda;\phi)^{-2} - \phi(1+v(-\lambda;\phi))^{-2}.
\end{aligned}$$

Solving the first equation for $v(-\lambda;\phi) \geq 0$ gives

$$v(-\lambda;\phi) = \frac{1}{2\lambda}(-(\phi+\lambda-1) + \sqrt{(\phi+\lambda-1)^2 + 4\lambda}). \tag{102}$$

Then the asymptotic bias and variance defined in Theorem 19 can be evaluated accordingly.

Appendix H. Auxiliary asymptotic equivalency results

H.1 Preliminaries

We use the notion of asymptotic equivalence of sequences of random matrices in various proofs. In this section, we provide a basic review of the related definitions and corresponding calculus rules.

Definition 30 (Asymptotic equivalence: deterministic version)**.** *Consider sequences* $\{\boldsymbol{A}_p\}_{p\geq 1}$ *and* $\{\boldsymbol{B}_p\}_{p\geq 1}$ *of (random or deterministic) matrices of growing dimensions. We say that* $\boldsymbol{A}_p$ *and* $\boldsymbol{B}_p$ *are equivalent and write* $\boldsymbol{A}_p \simeq_D \boldsymbol{B}_p$ *if* $\lim_{p\to\infty} |\operatorname{tr}[\boldsymbol{C}_p(\boldsymbol{A}_p - \boldsymbol{B}_p)]| = 0$ *almost surely for any sequence of matrices* $\boldsymbol{C}_p$ *with bounded trace norm such that* $\limsup_{p\to\infty} \|\boldsymbol{C}_p\|_{\mathrm{tr}} < \infty$*.*

We emphasize that recent work (Dobriban and Sheng, 2021; Patil et al., 2022b) used the deterministic version of the asymptotic equivalence, implicitly assuming that $\boldsymbol{C}_p$ in the definition is deterministic. However, in this paper we need to investigate the asymptotic equivalence relationship conditional on some other sequences. In that direction, we first extend Definition 30 to allow for random $\boldsymbol{C}_p$, as in Definition 31.

Definition 31 (Asymptotic equivalence: random version)**.** *Consider sequences* $\{\boldsymbol{A}_p\}_{p\geq 1}$ *and* $\{\boldsymbol{B}_p\}_{p\geq 1}$ *of (random or deterministic) matrices of growing dimensions. We say that* $\boldsymbol{A}_p$ *and* $\boldsymbol{B}_p$ *are equivalent and write* $\boldsymbol{A}_p \simeq_R \boldsymbol{B}_p$ *if* $\lim_{p\to\infty} |\operatorname{tr}[\boldsymbol{C}_p(\boldsymbol{A}_p - \boldsymbol{B}_p)]| = 0$ *almost surely for any sequence of random matrices* $\boldsymbol{C}_p$ *independent of* $\boldsymbol{A}_p$ *and* $\boldsymbol{B}_p$*, with bounded trace norm such that* $\limsup_{p\to\infty} \|\boldsymbol{C}_p\|_{\operatorname{tr}} < \infty$ *almost surely.*

Even though Definition 30 seems to be more restrictive than Definition 31, they are indeed equivalent as shown in Proposition 32. The latter definition allows for more general definition for "conditional" asymptotic equivalents.

Proposition 32 (Equivalence of $\simeq_D$ and $\simeq_R$)**.** *The asymptotic equivalent relations* $\simeq_D$ *in Definition 30 and* $\simeq_R$ *in Definition 31 are equivalent.*

Proof Let $\{\boldsymbol{A}_p\}$ and $\{\boldsymbol{B}\}_p$ be two sequences of random matrices. Suppose that $\boldsymbol{A}_p \simeq_D \boldsymbol{B}_p$. We next show that $\boldsymbol{A}_p \simeq_R \boldsymbol{B}_p$ holds. For any sequence of random matrices $\boldsymbol{C}_p$ that is independent of $\boldsymbol{A}_p$ and $\boldsymbol{B}_p$ for all $p \in \mathbb{N}$, and has bounded trace norm such that $\limsup \|\boldsymbol{C}_p\|_{\operatorname{tr}} < \infty$ as $p \to \infty$ almost surely. Let A denote the event that $\lim_{p\to\infty} |\operatorname{tr}[\boldsymbol{C}_p(\boldsymbol{A}_p - \boldsymbol{B}_p)]| = 0$. Then

$$\mathbb{P}(A) = \mathbb{E}[\mathbb{1}_A] \overset{(a)}{=} \mathbb{E}[\mathbb{E}[\mathbb{1}_A \mid \{\boldsymbol{C}_p\}_{p\geq 1}]] \overset{(b)}{=} \mathbb{E}[1] = 1.$$

Above, equality (a) follows from the law of total expectation. Inequality (b) holds almost surely because $\boldsymbol{A}_p \simeq_D \boldsymbol{B}_p$ and $\boldsymbol{C}_p$ is independent of $\boldsymbol{A}_p$ and $\boldsymbol{B}_p$. This can be seen as follows. Note that $\mathbb{1}_A(\{\boldsymbol{C}_p\}, (\{\boldsymbol{A}_p\}, \{\boldsymbol{B}_p\}))$ is a function of random variables $\{\boldsymbol{C}_p\}$ and $(\{\boldsymbol{A}_p\}, \{\boldsymbol{B}_p\})$. Let

$$\mathbb{E}[\mathbb{1}_A(\{\boldsymbol{c}_p\}, (\{\boldsymbol{A}_p\}, \{\boldsymbol{B}_p\}))] = h(\{\boldsymbol{c}_p\}),$$

where the expectation is taken over the randomness in $(\{\boldsymbol{A}_p\}, \{\boldsymbol{B}_p\})$. Since $\{\boldsymbol{C}_p\}$ and $(\{\boldsymbol{A}_p\}, \{\boldsymbol{B}_p\})$ are independent and $\mathbb{E}[| \mathbb{1}_A |] \leq 1 < \infty$, we have that (see, e.g., Shiryaev (2016, Chapter 2, Section 7, Equation (16)), or Durrett (2010, Example 5.1.5))

$$\mathbb{E}[\mathbb{1}_A \mid \{\boldsymbol{C}_p\}] = h(\{\boldsymbol{C}_p\}),$$

and from Definition 30, we have $h(\{\boldsymbol{C}_p\}) = 1$ almost surely. Thus, we can conclude that $\boldsymbol{A}_p \simeq_R \boldsymbol{B}_p$.

On the other hand, by definition, $\boldsymbol{A}_p \simeq_R \boldsymbol{B}_p$ directly implies $\boldsymbol{A}_p \simeq_D \boldsymbol{B}_p$, which completes the proof. ■

The properties for the two types of deterministic equivalents are summarized in Lemma 33. Though most of the calculus rules are the direct consequences from Dobriban and Wager (2018); Dobriban and Sheng (2021), the product rule involving random matrices $\boldsymbol{C}_p$ does not immediately follow from previous work.

Lemma 33 (Calculus of deterministic equivalents)**.** *Let* $\boldsymbol{A}_p$*,* $\boldsymbol{B}_p$*,* $\boldsymbol{C}_p$ *and* $\boldsymbol{D}_p$ *be sequences of random matrices. The calculus of deterministic equivalents (*$\simeq_D$ *and* $\simeq_R$*) satisfies the following properties:*

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

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

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

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

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

Proof The conclusions for $\simeq_D$ directly follow from Dobriban and Wager (2018); Dobriban and Sheng (2021). Then, the proof of property (1), (2), (4), and (5) for $\simeq_R$ follows from Proposition 32. It remains to show that the product rule holds for $\simeq_R$. Since $\boldsymbol{B}_p \simeq_R \boldsymbol{C}_p$, we have $\boldsymbol{B}_p \simeq_D \boldsymbol{C}_p$. Then for any sequence of random matrices $\{\boldsymbol{D}_p\}_{p\geq 1}$ that have bounded trace norm and are independent of $\boldsymbol{B}_p$ and $\boldsymbol{C}_p$, we have

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

Because $|\,\mathrm{tr}[\boldsymbol{D}_p(\boldsymbol{A}_p\boldsymbol{B}_p - \boldsymbol{A}_p\boldsymbol{C}_p)]| \leq \|\boldsymbol{A}_p\|_{\mathrm{op}} |\,\mathrm{tr}[\boldsymbol{D}_p(\boldsymbol{B}_p - \boldsymbol{C}_p)]|$ and $\limsup_{p\to\infty} \|\boldsymbol{A}_p\|_{\mathrm{op}} < \infty$, we have that $\lim_{p\to\infty} |\,\mathrm{tr}[\boldsymbol{D}_p(\boldsymbol{B}_p - \boldsymbol{C}_p)]| = 0$ implies $\lim_{p\to\infty} |\,\mathrm{tr}[\boldsymbol{D}_p(\boldsymbol{A}_p\boldsymbol{B}_p - \boldsymbol{A}_p\boldsymbol{C}_p)]| = 0$ conditioning on $\{\boldsymbol{A}_p\}_{p\geq 1}$. Thus,

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

and by law of total expectation

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

which holds for any sequence of random matrices $\{\boldsymbol{D}_p\boldsymbol{A}_p\}_{p\geq 1}$ that have bounded trace norm and are independent of $\boldsymbol{B}_p$ and $\boldsymbol{C}_p$. By definition, we have $\boldsymbol{A}_p\boldsymbol{B}_p \simeq \boldsymbol{A}_p\boldsymbol{C}_p$. ■

Since the asymptotic equivalent relation $\simeq_D$ is equivalent to $\simeq_R$, we will just ignore the subscript and use the notation "$\simeq$" for simplicity. The subscript will be specified when needed.

H.2 Conditioning and calculus

In this section, we extend the notion of asymptotic equivalence of two sequences of random matrices from Definitions 30 and 31 to incorporate conditioning on another sequence of random matrices.

Definition 34 (Conditional asymptotic equivalence)**.** *Consider sequences $\{\boldsymbol{A}_p\}_{p\geq 1}$, $\{\boldsymbol{B}_p\}_{p\geq 1}$ and $\{\boldsymbol{D}_p\}_{p\geq 1}$ of (random or deterministic) matrices of growing dimensions. We say that $\boldsymbol{A}_p$*

and $\boldsymbol{B}_p$ *are equivalent given* $\boldsymbol{D}_p$ *and write* $\boldsymbol{A}_p \simeq \boldsymbol{B}_p \mid \boldsymbol{D}_p$ *if* $\lim_{p\to\infty} |\operatorname{tr}[\boldsymbol{C}_p(\boldsymbol{A}_p - \boldsymbol{B}_p)]| = 0$ *almost surely conditional on* $\{\boldsymbol{D}_p\}_{p\geq 1}$*, i.e.,*

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

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

Below we formalize additional calculus rules that hold for conditional asymptotic equivalence Definition 34.

Proposition 35 (Calculus of conditional asymptotic equivalents)**.** *Let* $\boldsymbol{A}_p$*,* $\boldsymbol{B}_p$*,* $\boldsymbol{C}_p$*, and* $\boldsymbol{E}_p$ *be sequences of random matrices.*

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

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

Proof Proofs for the two parts appear below.

Part (1). For any sequence of deterministic matrices $\boldsymbol{C}_p$ with bounded trace norm, we have

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

because $\boldsymbol{A}_p \simeq \boldsymbol{B}_p \mid \boldsymbol{E}_p$. By the law of total expectation, we have

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

Thus, $\boldsymbol{A}_p \simeq_D \boldsymbol{B}_p$. By Proposition 32, we further have $\boldsymbol{A}_p \simeq_R \boldsymbol{B}_p$.

Part (2). For any sequence of random matrices $\boldsymbol{D}_p$, let E_1 and E_2 respectively denote the following events: $\lim_{p\to\infty} |\operatorname{tr}[\boldsymbol{D}_p(\boldsymbol{B}_p - \boldsymbol{C}_p)]| = 0$ and $\lim_{p\to\infty} |\operatorname{tr}[\boldsymbol{D}_p(\boldsymbol{A}_p\boldsymbol{B}_p - \boldsymbol{A}_p\boldsymbol{C}_p)]| = 0$. Because $\boldsymbol{B}_p \simeq \boldsymbol{C}_p \mid \boldsymbol{E}_p$, by definition we have

$$\mathbb{P}\left(E_1 \mid \{\boldsymbol{E}_p\}_{p\geq 1}\right) = 1$$

Because $|\operatorname{tr}[\boldsymbol{D}_p(\boldsymbol{A}_p\boldsymbol{B}_p - \boldsymbol{A}_p\boldsymbol{C}_p)]| \leq \|\boldsymbol{A}_p\|_{\mathrm{op}} |\operatorname{tr}[\boldsymbol{D}_p(\boldsymbol{B}_p - \boldsymbol{C}_p)]|$ and $\limsup_{p\to\infty} \|\boldsymbol{A}_p\|_{\mathrm{op}} < \infty$, we have E_1 implies E_2 conditioning on $\{\boldsymbol{E}_p\}_{p\geq 1}$. Thus we have

$$\mathbb{P}\left(E_2 \mid \{\boldsymbol{E}_p\}_{p\geq 1}\right) = 1$$

holds for any $\{\boldsymbol{D}_p\}_{p\geq 1}$. This implies that $\boldsymbol{A}_p\boldsymbol{B}_p \simeq \boldsymbol{A}_p\boldsymbol{C}_p \mid \boldsymbol{E}_p$. ∎

Other rules in Lemma 33 also hold for conditional asymptotic equivalents. A direct implication of this is that the deterministic equivalents for resolvents we will derive in Section H.3 based on these rules can be naturally generalized to allow for conditional asymptotic equivalents given a common sequence of random matrices that are independent of the source sequence.

H.3 Standard ridge resolvents and extensions

In this section, we collect various asymptotic equivalents that are used in the proofs of Lemmas 20 and 21, and Lemmas 24 to 26, which serve to prove Theorem 6. These equivalents are also subsequently used in the proof of Theorem 8.

H.3.1 Standard ridge resolvents

The following lemma provides a deterministic equivalent for the standard ridge resolvent and implies Corollary 37. It is adapted from Theorem 1 of Rubio and Mestre (2011). See also Theorem 3 of Dobriban and Sheng (2021).

Lemma 36 (Deterministic equivalent for standard ridge resolvent)**.** *Suppose $\boldsymbol{x}_i \in \mathbb{R}^p$, $1 \le i \le n$, are i.i.d. random vectors such that each $\boldsymbol{x}_i = \boldsymbol{z}_i \boldsymbol{\Sigma}^{1/2}$, where $\boldsymbol{z}_i$ is a random vector consisting of i.i.d. entries z_{ij}, $1 \le j \le p$, satisfying $\mathbb{E}[z_{ij}] = 0$, $\mathbb{E}[z_{ij}^2] = 1$, and $\mathbb{E}[|z_{ij}|^{8+\alpha}] \le M_\alpha$ for some constants $\alpha > 0$ and $M_\alpha < \infty$, and $\boldsymbol{\Sigma} \in \mathbb{R}^{p \times p}$ is a positive semidefinite matrix satisfying $0 \preceq \boldsymbol{\Sigma} \preceq r_{\max} I_p$ for some constant $r_{\max} < \infty$ (independent of p). Let $\boldsymbol{X} \in \mathbb{R}^{n \times p}$ the concatenated matrix with $\boldsymbol{x}_i^\top$, $1 \le i \le n$, as rows, and let $\widehat{\boldsymbol{\Sigma}} \in \mathbb{R}^{p \times p}$ denote the random matrix $\boldsymbol{X}^\top \boldsymbol{X}/n$. Let $\gamma := p/n$. Then, for $z \in \mathbb{C}^+$, as $n, p \to \infty$ such that $0 < \liminf \gamma \le \limsup \gamma < \infty$, we have*

$$(\widehat{\boldsymbol{\Sigma}} - z\boldsymbol{I}_p)^{-1} \simeq (c(e(z;\gamma))\boldsymbol{\Sigma} - z\boldsymbol{I}_p)^{-1}, \tag{103}$$

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

$$c(e(z;\gamma)) = \frac{1}{1 + \gamma e(z;\gamma)}, \tag{104}$$

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

$$e(z;\gamma) = \operatorname{tr}[\boldsymbol{\Sigma}(c(e(z;\gamma))\boldsymbol{\Sigma} - zI_p)^{-1}]/p. \tag{105}$$

Note that both the scalars $c(e(z;\gamma))$ and $e(z;\gamma)$ also implicitly depend on $\boldsymbol{\Sigma}$. For notation brevity, we do not always explicitly indicate this dependence. However, we will be explicit in such dependence for certain extensions to follow. See the remark after Lemma 38 for more details. Additionally, observe that one can eliminate $e(z;\gamma)$ in the statement of Lemma 36 by combining (104) and (105) so that for $z \in \mathbb{C}^+$, one has

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

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

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

The following corollary is a simple consequence of Lemma 36, which supplies a deterministic equivalent for the (regularization) scaled ridge resolvent. We will work with a real regularization parameter λ from here on.

Corollary 37 (Deterministic equivalent for scaled ridge resolvent)**.** *Assume the setting of Lemma 36. For $\lambda > 0$, we have*

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

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

$$\frac{1}{v(-\lambda;\gamma)} = \lambda + \gamma \int \frac{r}{1 + v(-\lambda;\gamma) r} \, \mathrm{d}H_n(r). \tag{106}$$

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

As a side note, the parameter $v(-\lambda;\gamma)$ in Corollary 37 is also the companion Stieltjes transform of the spectral distribution of the sample covariance matrix $\widehat{\mathbf{\Sigma}}$, which is also the Stieltjes transform of the spectral distribution of the gram matrix $\boldsymbol{X}\boldsymbol{X}^\top/n$.

The following lemma uses Corollary 37 along with calculus of deterministic equivalents (from Lemma 33), and provides deterministic equivalents for resolvents needed to obtain limiting bias and variance of standard ridge regression. It is adapted from Lemma S.6.10 of Patil et al. (2022b). These equivalents are standard and well-established in the prior literature. For example, the variance resolvent in Lemma 38 can be obtained from results in Dobriban and Sheng (2021), while the bias resolvent in Lemma 38 can be obtained from results in Hastie et al. (2022).

Lemma 38 (Deterministic equivalents for bias and variance resolvents of ridge regression)**.** *Suppose $\boldsymbol{x}_i \in \mathbb{R}^p$, $1 \leq i \leq n$, are i.i.d. random vectors with each $\boldsymbol{x}_i = \boldsymbol{z}_i \mathbf{\Sigma}^{1/2}$, where $\boldsymbol{z}_i \in \mathbb{R}^p$ is a random vector that contains i.i.d. random variables z_{ij}, $1 \leq j \leq p$, each with $\mathbb{E}[z_{ij}] = 0$, $\mathbb{E}[z_{ij}^2] = 1$, and $\mathbb{E}[|z_{ij}|^{8+\alpha}] \leq M_\alpha$ for some constants $\alpha > 0$ and $M_\alpha < \infty$, and $\mathbf{\Sigma} \in \mathbb{R}^{p \times p}$ is a positive semidefinite matrix with $r_{\min} \boldsymbol{I}_p \preceq \mathbf{\Sigma} \preceq r_{\max} \boldsymbol{I}_p$ for some constants $r_{\min} > 0$ and $r_{\max} < \infty$ (independent of p). Let $\boldsymbol{X} \in \mathbb{R}^{n \times p}$ be the concatenated random matrix with $\boldsymbol{x}_i$, $1 \leq i \leq n$, as its rows, and define $\widehat{\mathbf{\Sigma}} := \boldsymbol{X}^\top \boldsymbol{X}/n \in \mathbb{R}^{p \times p}$. Let $\gamma := p/n$. Then, for $\lambda > 0$, as $n, p \to \infty$ with $0 < \liminf \gamma \leq \limsup \gamma < \infty$, the following statements hold:*

(1) Bias of ridge regression:

$$\lambda^2 (\widehat{\mathbf{\Sigma}} + \lambda \boldsymbol{I}_p)^{-1} \mathbf{\Sigma} (\widehat{\mathbf{\Sigma}} + \lambda \boldsymbol{I}_p)^{-1} \simeq (v(-\lambda;\gamma)\mathbf{\Sigma} + \boldsymbol{I}_p)^{-1} (\widetilde{v}_b(-\lambda;\gamma)\mathbf{\Sigma} + \mathbf{\Sigma})(v(-\lambda;\gamma)\mathbf{\Sigma} + \boldsymbol{I}_p)^{-1}. \tag{107}$$

(2) Variance of ridge regression:

$$(\widehat{\mathbf{\Sigma}} + \lambda \boldsymbol{I}_p)^{-2} \widehat{\mathbf{\Sigma}} \mathbf{\Sigma} \simeq \widetilde{v}_v(-\lambda;\gamma)(v(-\lambda;\gamma)\mathbf{\Sigma} + \boldsymbol{I}_p)^{-2} \mathbf{\Sigma}\mathbf{\Sigma}. \tag{108}$$

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

$$\frac{1}{v(-\lambda;\gamma,\mathbf{\Sigma})} = \lambda + \int \frac{\gamma r}{1 + v(-\lambda;\gamma,\mathbf{\Sigma}) r} \, \mathrm{d}H_n(r;\mathbf{\Sigma}), \tag{109}$$

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

$$\widetilde{v}_b(-\lambda;\gamma,\mathbf{\Sigma}) = \frac{\int \gamma r^2 (1 + v(-\lambda;\gamma,\mathbf{\Sigma}) r)^{-2} \, \mathrm{d}H_n(r;\mathbf{\Sigma})}{v(-\lambda;\gamma,\mathbf{\Sigma})^{-2} - \int \gamma r^2 (1 + v(-\lambda;\gamma,\mathbf{\Sigma}) r)^{-2} \, \mathrm{d}H_n(r;\mathbf{\Sigma})}, \tag{110}$$

$$\widetilde{v}_v(-\lambda; \gamma, \mathbf{\Sigma})^{-1} = v(-\lambda; \gamma, \mathbf{\Sigma})^{-2} - \int \gamma r^2 (1 + v(-\lambda; \gamma, \mathbf{\Sigma}) r)^{-2} \, \mathrm{d}H_n(r; \mathbf{\Sigma}), \tag{111}$$

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

A couple of remarks on Lemma 38 follow.

- The dependency of various scalar parameters appearing in Lemma 38 on the matrix $\mathbf{\Sigma}$ is explicitly highlighted the statement. This is because when we extend the current results later in Lemma 39, these parameters depend on the distributions of eigenvalues of matrices other than $\mathbf{\Sigma}$. In places where it is clear from context, we will write $H_n(r)$, $v(-\lambda; \gamma)$, $\widetilde{v}_b(-\lambda; \gamma)$, and $\widetilde{v}_v(-\lambda; \gamma)$ to denote $H_n(r; \mathbf{\Sigma})$, $v(-\lambda; \gamma, \mathbf{\Sigma})$, $\widetilde{v}_b(-\lambda; \gamma, \mathbf{\Sigma})$, and $\widetilde{v}_v(-\lambda; \gamma, \mathbf{\Sigma})$, respectively, for notational simplicity.
- Lemmas 36 and 38 assume existence of moments of order $8 + \alpha$ for some $\alpha > 0$ on the entries of $\boldsymbol{z}_i$, $1 \le i \le k_m$, mentioned in assumption 2. As done in the proof of Theorem 6 of Hastie et al. (2022) (in Appendix A.4 therein), this can be relaxed to only requiring existence of moments of order $4 + \alpha$ by a truncation argument. We omit the details and refer the readers to Hastie et al. (2022).

H.3.2 Extended ridge resolvents

The lemma below extends the deterministic equivalents of the ridge resolvents in Lemma 38 to provide deterministic equivalents for Tikhonov resolvents, where the regularization matrix $\lambda \boldsymbol{I}_p$ is replaced with $\lambda(\boldsymbol{I}_p + \boldsymbol{C})$ and $\boldsymbol{C} \in \mathbb{R}^{p \times p}$ is an arbitrary positive semidefinite random matrix. While the derivation of extended resolvents in Lemma 39 naturally follows from Lemma 38, we have specifically isolated these extensions. This abstraction facilitates their repeated application in our conditioning arguments, especially in the proofs of Theorems 6 and 8.

Lemma 39 (Tikhonov resolvents)**.** *Suppose the conditions in Lemma 38 holds. Let $\boldsymbol{C} \in \mathbb{R}^{p \times p}$ be any symmetric and positive semidefinite random matrix with uniformly bounded operator norm in p that is independent of $\boldsymbol{X}$ for all $n, p \in \mathbb{N}$, and let $\boldsymbol{N} = (\widehat{\mathbf{\Sigma}} + \lambda \boldsymbol{I}_p)^{-1}$. Then the following statements hold:*

(1) Tikhonov resolvent:

$$\lambda(\boldsymbol{N}^{-1} + \lambda \boldsymbol{C})^{-1} \simeq \widetilde{\mathbf{\Sigma}}_{\boldsymbol{C}}^{-1}. \tag{112}$$

(2) Bias of Tikhonov regression:

$$\lambda^2 (\boldsymbol{N}^{-1} + \lambda \boldsymbol{C})^{-1} \mathbf{\Sigma} (\boldsymbol{N}^{-1} + \lambda \boldsymbol{C})^{-1} \simeq \widetilde{\mathbf{\Sigma}}_{\boldsymbol{C}}^{-1} (\widetilde{v}_b(-\lambda; \gamma, \mathbf{\Sigma}_{\boldsymbol{C}}) \mathbf{\Sigma} + \mathbf{\Sigma}) \widetilde{\mathbf{\Sigma}}_{\boldsymbol{C}}^{-1}. \tag{113}$$

(3) Variance of Tikhonov regression:

$$(\boldsymbol{N}^{-1} + \lambda \boldsymbol{C})^{-1} \widehat{\mathbf{\Sigma}} (\boldsymbol{N}^{-1} + \lambda \boldsymbol{C})^{-1} \mathbf{\Sigma} \simeq \widetilde{v}_v(-\lambda; \gamma, \mathbf{\Sigma}_{\boldsymbol{C}}) \widetilde{\mathbf{\Sigma}}_{\boldsymbol{C}}^{-1} \mathbf{\Sigma} \widetilde{\mathbf{\Sigma}}_{\boldsymbol{C}}^{-1} \mathbf{\Sigma}, \tag{114}$$

where $\boldsymbol{\Sigma}_{\boldsymbol{C}} = (\boldsymbol{I}_p + \boldsymbol{C})^{-\frac{1}{2}}\boldsymbol{\Sigma}(\boldsymbol{I}_p + \boldsymbol{C})^{-\frac{1}{2}}$, $\widetilde{\boldsymbol{\Sigma}}_{\boldsymbol{C}} = v(-\lambda;\gamma,\boldsymbol{\Sigma}_{\boldsymbol{C}})\boldsymbol{\Sigma} + \boldsymbol{I}_p + \boldsymbol{C}$. *Here,* $v(-\lambda;\gamma,\boldsymbol{\Sigma}_{\boldsymbol{C}})$, $\widetilde{v}_b(-\lambda;\gamma,\boldsymbol{\Sigma}_{\boldsymbol{C}})$, *and* $\widetilde{v}_v(-\lambda;\gamma,\boldsymbol{\Sigma}_{\boldsymbol{C}})$ *defined by* (109)-(111) *simplify to*

$$\frac{1}{v(-\lambda;\gamma,\boldsymbol{\Sigma}_{\boldsymbol{C}})} = \lambda + \gamma \operatorname{tr}[(v(-\lambda;\gamma,\boldsymbol{\Sigma}_{\boldsymbol{C}})\boldsymbol{\Sigma} + \boldsymbol{I}_p + \boldsymbol{C})^{-1}\boldsymbol{\Sigma}]/p, \tag{115}$$

$$\frac{1}{\widetilde{v}_v(-\lambda;\gamma,\boldsymbol{\Sigma}_{\boldsymbol{C}})} = \frac{1}{v(-\lambda;\gamma,\boldsymbol{\Sigma}_{\boldsymbol{C}})^2} - \gamma \operatorname{tr}[(v(-\lambda;\gamma,\boldsymbol{\Sigma}_{\boldsymbol{C}})\boldsymbol{\Sigma} + \boldsymbol{I}_p + \boldsymbol{C})^{-2}\boldsymbol{\Sigma}^2]/p, \tag{116}$$

$$\widetilde{v}_b(-\lambda;\gamma,\boldsymbol{\Sigma}_{\boldsymbol{C}}) = \gamma \operatorname{tr}[(v(-\lambda;\gamma,\boldsymbol{\Sigma}_{\boldsymbol{C}})\boldsymbol{\Sigma} + \boldsymbol{I}_p + \boldsymbol{C})^{-2}\boldsymbol{\Sigma}^2]/p \cdot \widetilde{v}_v(-\lambda;\gamma,\boldsymbol{\Sigma}_{\boldsymbol{C}}). \tag{117}$$

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

Proof Proofs for the different parts are separated below.

Part (1). Note that

$$\lambda(\boldsymbol{N}^{-1} + \lambda\boldsymbol{C})^{-1} = \lambda(\widehat{\boldsymbol{\Sigma}} + \lambda(\boldsymbol{I}_p + \boldsymbol{C}))^{-1} = (\boldsymbol{I}_p + \boldsymbol{C})^{-\frac{1}{2}}\lambda(\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{C}} + \lambda\boldsymbol{I}_p)^{-1}(\boldsymbol{I}_p + \boldsymbol{C})^{-\frac{1}{2}}, \tag{118}$$

where $\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{C}} = \boldsymbol{\Sigma}_{\boldsymbol{C}}^{\frac{1}{2}}(\boldsymbol{Z}^\top\boldsymbol{Z}/n)\boldsymbol{\Sigma}_{\boldsymbol{C}}^{\frac{1}{2}}$, and $\boldsymbol{\Sigma}_{\boldsymbol{C}} = (\boldsymbol{I}_p + \boldsymbol{C})^{-\frac{1}{2}}\boldsymbol{\Sigma}(\boldsymbol{I}_p + \boldsymbol{C})^{-\frac{1}{2}}$. Using Lemma 36, we have

$$\lambda(\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{C}} + \lambda\boldsymbol{I}_p)^{-1} \simeq (v(-\lambda;\gamma,\boldsymbol{\Sigma}_{\boldsymbol{C}})\boldsymbol{\Sigma}_{\boldsymbol{C}} + \boldsymbol{I}_p)^{-1}, \tag{119}$$

where $v(-\lambda;\gamma,\boldsymbol{\Sigma}_{\boldsymbol{C}})$ is the unique solution to the fixed point equation (106) such that

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

Note that $\left\|(\boldsymbol{I}_p + \boldsymbol{C})^{-1}\right\|_{\mathrm{op}} \le 1$. We can apply the product rule from Lemma 33 (3) and get

$$\lambda(\boldsymbol{N}^{-1} + \lambda\boldsymbol{C})^{-1} \simeq (\boldsymbol{I}_p + \boldsymbol{C})^{-\frac{1}{2}}(v(-\lambda;\gamma,\boldsymbol{\Sigma}_{\boldsymbol{C}})\boldsymbol{\Sigma}_{\boldsymbol{C}} + I_p)^{-1}(\boldsymbol{I}_p + \boldsymbol{C})^{-\frac{1}{2}} = (v(-\lambda;\gamma,\boldsymbol{\Sigma}_{\boldsymbol{C}})\boldsymbol{\Sigma} + \boldsymbol{I}_p + \boldsymbol{C})^{-1},$$

by combining (118)-(119).

Part (2). From Lemma 38 (1), we have

$$\begin{aligned}
&\lambda^2(\boldsymbol{N}^{-1} + \lambda\boldsymbol{C})^{-1}\boldsymbol{\Sigma}(\boldsymbol{N}^{-1} + \lambda\boldsymbol{C})^{-1} \\
&= \lambda^2(\boldsymbol{I}_p + \boldsymbol{C})^{-\frac{1}{2}} \cdot [(\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{C}} + \lambda\boldsymbol{I}_p)^{-1}(\boldsymbol{I}_p + \boldsymbol{C})^{-\frac{1}{2}} \cdot \boldsymbol{\Sigma} \cdot (\boldsymbol{I}_p + \boldsymbol{C})^{-\frac{1}{2}}(\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{C}} + \lambda\boldsymbol{I}_p)^{-1}] \cdot (\boldsymbol{I}_p + \boldsymbol{C})^{-\frac{1}{2}} \\
&\simeq (\boldsymbol{I}_p + \boldsymbol{C})^{-\frac{1}{2}} \cdot [(v(-\lambda;\gamma,\boldsymbol{\Sigma}_{\boldsymbol{C}})\boldsymbol{\Sigma}_{\boldsymbol{C}} + \boldsymbol{I}_p)^{-1} \\
&\qquad \cdot (\widetilde{v}_b(-\lambda;\gamma,\boldsymbol{\Sigma}_{\boldsymbol{C}})\boldsymbol{\Sigma}_{\boldsymbol{C}} + (\boldsymbol{I}_p + \boldsymbol{C})^{-\frac{1}{2}}\boldsymbol{\Sigma}(\boldsymbol{I}_p + \boldsymbol{C})^{-\frac{1}{2}}) \cdot (v(-\lambda;\gamma,\boldsymbol{\Sigma}_{\boldsymbol{C}})\boldsymbol{\Sigma}_{\boldsymbol{C}} + \boldsymbol{I}_p)^{-1}] \cdot (\boldsymbol{I}_p + \boldsymbol{C})^{-\frac{1}{2}} \\
&= (v(-\lambda;\gamma,\boldsymbol{\Sigma}_{\boldsymbol{C}})\boldsymbol{\Sigma} + \boldsymbol{I}_p + \boldsymbol{C})^{-1}(\widetilde{v}_b(-\lambda;\gamma,\boldsymbol{\Sigma}_{\boldsymbol{C}})\boldsymbol{\Sigma} + \boldsymbol{\Sigma})(v(-\lambda;\gamma,\boldsymbol{\Sigma}_{\boldsymbol{C}})\boldsymbol{\Sigma} + \boldsymbol{I}_p + \boldsymbol{C})^{-1}.
\end{aligned}$$

Part (3). Similar to Part (2), from Lemma 38 (2), we have

$$\begin{aligned}
&(\boldsymbol{N}^{-1} + \lambda\boldsymbol{C})^{-1}\widehat{\boldsymbol{\Sigma}}(\boldsymbol{N}^{-1} + \lambda\boldsymbol{C})^{-1}\boldsymbol{\Sigma} \\
&= (\boldsymbol{I}_p + \boldsymbol{C})^{-\frac{1}{2}} \cdot (\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{C}} + \lambda\boldsymbol{I}_p)^{-1}\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{C}}(\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{C}} + \lambda\boldsymbol{I}_p)^{-1} \cdot (\boldsymbol{I}_p + \boldsymbol{C})^{-\frac{1}{2}}\boldsymbol{\Sigma}
\end{aligned}$$

$$\simeq (\boldsymbol{I}_p + \boldsymbol{C})^{-\frac{1}{2}} \cdot \widetilde{v}_v(-\lambda; \gamma, \boldsymbol{\Sigma_C})(v(-\lambda; \gamma, \boldsymbol{\Sigma_C})\boldsymbol{\Sigma_C} + \boldsymbol{I}_p)^{-1}\boldsymbol{\Sigma_C}(v(-\lambda; \gamma, \boldsymbol{\Sigma_C})\boldsymbol{\Sigma_C} + \boldsymbol{I}_p)^{-1} \cdot (\boldsymbol{I}_p + \boldsymbol{C})^{-\frac{1}{2}}\boldsymbol{\Sigma}$$
$$= \widetilde{v}_v(-\lambda; \gamma, \boldsymbol{\Sigma_C})(v(-\lambda; \gamma, \boldsymbol{\Sigma_C})\boldsymbol{\Sigma} + \boldsymbol{I}_p + \boldsymbol{C})^{-1}\boldsymbol{\Sigma}(v(-\lambda; \gamma, \boldsymbol{\Sigma_C})\boldsymbol{\Sigma} + \boldsymbol{I}_p + \boldsymbol{C})^{-1}\boldsymbol{\Sigma}.$$

Note that the distribution of $\boldsymbol{\Sigma_C}$'s eigenvalue has positive support. By the continuity of $v(-\lambda; \cdot, \boldsymbol{\Sigma_C})$, $\widetilde{v}_b(-\lambda; \cdot, \boldsymbol{\Sigma_C})$, and $\widetilde{v}_v(-\lambda; \cdot, \boldsymbol{\Sigma_C})$ from Lemma 42 (2), (4) and (3), γ can by replaced by its limit ϕ as $n, p \to \infty$. ■

The following lemma concerns the deterministic equivalents of the precision matrix as the weighted average of two sample covariance matrices of subsamples, when the full sample covariance matrix is invertible almost surely. It is useful when we aim to condition on one of the subsampled covariance matrix, which is used in the proof of Lemma 25.

Lemma 40 (Deterministic equivalent of subsamples in the underparameterized regime)**.** *Suppose the conditions in Lemma 38 holds. Let $\widehat{\boldsymbol{\Sigma}}_0$ be the sample covariance matrix computed using i observations of $\boldsymbol{X}$, and $\widehat{\boldsymbol{\Sigma}}_1$ be the sample covariance matrix computed using the remaining $n - i$ samples. Let $\pi_0 = i/n$ and $\pi_1 = (n-i)/n$. Suppose that $p/n \to \phi \in (0, 1)$ as $n, p \to \infty$. Then, we have*

$$(\pi_0\widehat{\boldsymbol{\Sigma}}_0 + \pi_1\widehat{\boldsymbol{\Sigma}}_1)^{-1} \simeq (\pi_0\widehat{\boldsymbol{\Sigma}}_0 + (1-\phi)\pi_1\boldsymbol{\Sigma})^{-1}.$$

Proof We first note that when $\phi \in (0, 1)$, the eigenvalues of $\widehat{\boldsymbol{\Sigma}} = \pi_0\widehat{\boldsymbol{\Sigma}}_0 + \pi_1\widehat{\boldsymbol{\Sigma}}_1$ are bounded away from zero almost surely (Bai and Silverstein, 2010) and hence the inverse is well defined almost surely as $n, p \to \infty$.

The idea for the proof is to consider the perturbed resolvent $(\pi_0\widehat{\boldsymbol{\Sigma}}_0 + \mu\boldsymbol{I}_p + \pi_1\widehat{\boldsymbol{\Sigma}}_1)^{-1}$ for $\mu > 0$. Note that since the matrix $(\pi_0\widehat{\boldsymbol{\Sigma}}_0 + \pi_1\widehat{\boldsymbol{\Sigma}}_1)$ is almost surely invertible. Then,

$$\lim_{\mu \to 0^+} (\pi_0\widehat{\boldsymbol{\Sigma}}_0 + \mu\boldsymbol{I}_p + \pi_1\widehat{\boldsymbol{\Sigma}}_1)^{-1} = (\pi_0\widehat{\boldsymbol{\Sigma}}_0 + \pi_1\widehat{\boldsymbol{\Sigma}}_1)^{-1}.$$

Conditioned on $(\pi_0\widehat{\boldsymbol{\Sigma}}_0 + \mu\boldsymbol{I}_p)$, we have

$$\begin{aligned}
(\pi_0\widehat{\boldsymbol{\Sigma}}_0 + \mu\boldsymbol{I}_p + \pi_1\widehat{\boldsymbol{\Sigma}}_1)^{-1} &= a(\boldsymbol{A} + \widehat{\boldsymbol{\Sigma}}_1)^{-1} \\
&= a\boldsymbol{A}^{-\frac{1}{2}}(\boldsymbol{I}_p + \boldsymbol{A}^{-\frac{1}{2}}\widehat{\boldsymbol{\Sigma}}_1\boldsymbol{A}^{-\frac{1}{2}})\boldsymbol{A}^{-\frac{1}{2}} \\
&= a\boldsymbol{A}^{-\frac{1}{2}}(\boldsymbol{I}_p + \widehat{\boldsymbol{\Sigma}}_{1,\boldsymbol{A}})^{-1}\boldsymbol{A}^{-\frac{1}{2}} \\
&\simeq a\boldsymbol{A}^{-\frac{1}{2}}(\boldsymbol{I}_p + c\boldsymbol{\Sigma}_{\boldsymbol{A}})\boldsymbol{A}^{-\frac{1}{2}} \\
&= a(\boldsymbol{A} + c\boldsymbol{\Sigma})^{-1} \\
&= (\pi_0\widehat{\boldsymbol{\Sigma}}_0 + \mu\boldsymbol{I}_p + c\pi_1\boldsymbol{\Sigma})^{-1},
\end{aligned}$$

where the intermediate constants are $a = \pi_1^{-1}$, $\boldsymbol{A} = a\pi_0\widehat{\boldsymbol{\Sigma}}_0 + a\mu\boldsymbol{I}_p$, $\widehat{\boldsymbol{\Sigma}}_{1,\boldsymbol{A}} = \boldsymbol{A}^{-\frac{1}{2}}\widehat{\boldsymbol{\Sigma}}\boldsymbol{A}^{-\frac{1}{2}}$, $\boldsymbol{\Sigma}_{\boldsymbol{A}} = \boldsymbol{A}^{-\frac{1}{2}}\boldsymbol{\Sigma}\boldsymbol{A}^{-\frac{1}{2}}$, and c satisfy the fixed-point equation

$$\begin{aligned}
\frac{1}{c} &= 1 + \frac{p}{n-i}\operatorname{tr}[\boldsymbol{\Sigma}_{\boldsymbol{A}}(c\boldsymbol{\Sigma}_{\boldsymbol{A}} + \boldsymbol{I}_p)^{-1}]/p \\
&= 1 + \frac{p}{k}\frac{k}{n-i}\operatorname{tr}[\boldsymbol{A}^{-1/2}\boldsymbol{\Sigma}\boldsymbol{A}^{-1/2}(c\boldsymbol{A}^{-1/2}\boldsymbol{\Sigma}\boldsymbol{A}^{-1/2} + \boldsymbol{I}_p)^{-1}]/p
\end{aligned}$$

$$
\begin{aligned}
&= 1 + \phi a \operatorname{tr}[\mathbf{\Sigma}(c\mathbf{\Sigma} + \boldsymbol{A})^{-1}]/p \\
&= 1 + \phi \operatorname{tr}[\mathbf{\Sigma}(c\pi_1\mathbf{\Sigma} + \pi_0\widehat{\mathbf{\Sigma}}_0 + \mu\boldsymbol{I}_p)^{-1}]/p \\
&= 1 + \phi \operatorname{tr}[\mathbf{\Sigma}(\pi_0\widehat{\mathbf{\Sigma}}_0 + \mu\boldsymbol{I}_p + \pi_1\widehat{\mathbf{\Sigma}}_1)^{-1}]/p,
\end{aligned}
$$

where in final equality, we used the trace property of the asymptotic equivalence

$$(\pi_0\widehat{\mathbf{\Sigma}}_0 + \mu\boldsymbol{I}_p + c\pi_1\mathbf{\Sigma})^{-1} \simeq (\pi_0\widehat{\mathbf{\Sigma}}_0 + \mu\boldsymbol{I}_p + \pi_1\widehat{\mathbf{\Sigma}}_1)^{-1}.$$

Now note that

$$(\pi_0\widehat{\mathbf{\Sigma}}_0 + \mu\boldsymbol{I}_p + \pi_1\widehat{\mathbf{\Sigma}}_1)^{-1} = (\widehat{\mathbf{\Sigma}} + \mu\boldsymbol{I}_p)^{-1} \simeq (c'\mathbf{\Sigma} + \mu\boldsymbol{I}_p)^{-1}$$

where c' solves the fixed-point equation

$$\frac{1}{c'} = 1 + \phi \operatorname{tr}[\mathbf{\Sigma}(c'\mathbf{\Sigma} + \mu\boldsymbol{I}_p)^{-1}]/p.$$

Thus, the fixed-point in c can be written as

$$\frac{1}{c} = 1 + \phi \operatorname{tr}[\mathbf{\Sigma}(c'\mathbf{\Sigma} + \mu\boldsymbol{I}_p)^{-1}]/p.$$

We note that $c = c'$ satisfy the fixed-point equation for c (from the fixed-point equation for c'). Since c is a unique solution, this must be the solution. Letting $\mu \to 0^+$, we observe that $c' = 1 - \phi$ is the solution for the fixed-point equation in c'. Thus, we also have $c = 1 - \phi$. ■

H.4 Analytic properties of associated fixed-point equations

In this section, we compile results regarding analytical properties of the fixed-point solution $v(-\lambda;\phi)$ as defined in (106).

The subsequent lemma affirms the existence and uniqueness of the solution $v(-\lambda;\phi)$. It establishes a connection between the properties of the derivatives described in Lemma 41 and the properties of $\widetilde{v}_v(-\lambda;\phi)$ as defined in (111). Note the latter equals $-f'(x)$, where the function f is as defined in (120)

Lemma 41 (Properties of the solution of the fixed-point equation)**.** *Let $\lambda, \phi, a > 0$ and $b < \infty$ be real numbers. Let P be a probability measure supported on $[a, b]$. Define the function as follows:*

$$f(x) = \frac{1}{x} - \phi \int \frac{r}{1 + rx} \mathrm{d}P(r) - \lambda. \tag{120}$$

Then the following properties hold:

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

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

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

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

Proof We consider different parts separately below.

Part (1). Observe that

$$f(x) = \frac{1}{x} - \phi \int \frac{r}{xr + 1} \, \mathrm{d}P(r) = g_1(x) h_1(x),$$

where

$$g_1(x) = \frac{1}{x}, \qquad h_1(x) = 1 - \phi \int \frac{xr}{xr + 1} \, \mathrm{d}P(r).$$

Note that g_1 is positive and strictly decreasing over $(0, \infty)$ with $\lim_{x \to 0^+} g_1(x) = \infty$ and $\lim_{x \to \infty} g_1(x) = 0$, while h_1 is strictly decreasing over $(0, \infty)$ with $h_1(0) = 1$ and $\lim_{x \to \infty} h_1(x) = 1 - \phi < 0$. Thus, there is a unique $0 < x_0 < \infty$ such that $h_1(x_0) = 0$, and consequently $f(x_0) = 0$. Because h_1 is positive over $(0, x_0)$, and negative over (x_0, ∞), f is positive strictly decreasing over $(0, x_0)$ and negative over (x_0, ∞), with $\lim_{x \to 0^+} f(x) = \infty$ and $\lim_{x \to \infty} f(x) = 0$.

Part (2). Note that $f(x) = g_1(x) h_1(x) - \lambda$. Since from (1) $\lim_{x \to 0} g_1(x) h_1(x) = \infty$ and $\lim_{x \to 0} g_1(x) h_1(x) = 0$, we have that $\lim_{x \to 0^+} f(x) = +\infty$ and $\lim_{x \to \infty} f(x) = -\lambda < 0$.

For $\phi > 1$, since $g_1(x) h_1(x)$ is positive and strictly decreasing over $(0, x_0)$ and negative over (x_0, ∞), and $\lim_{x \to 0^+} g_1(x) h_1(x) = \infty$, we have that there exists $x_0^\lambda \in (0, x_0)$ such that $f(x_0^\lambda) = 0$. The properties of f over $(0, x_0^\lambda)$ and (x_0^λ, ∞) follow analogously as in (1).

For $\phi \in (0, 1]$, since $g_1 h_1$ is continuous, positive and strictly decreasing over $(0, \infty)$, by intermediate value theorem, there exists $x_0^\lambda \in (0, \infty)$ such that $f(x_0^\lambda) = 0$, f is positive and strictly decreasing for $x < x_0^\lambda$ and negative for $x > x_0^\lambda$, with $\lim_{x \to 0^+} f(x) = \infty$ and $\lim_{x \to \infty} f(x) = -\lambda$.

Part (3). Since f is monotone and continuous, it is differentiable. The derivative f' at x is given by

$$f'(x) = -\frac{1}{x^2} + \phi \int \frac{r^2}{(xr + 1)^2} \, \mathrm{d}P(r) = -g_2(x) h_2(x),$$

where

$$g_2(x) = \frac{1}{x^2}, \qquad h_2(x) = \left(1 - \phi \int \left(\frac{xr}{xr + 1} \right)^2 \mathrm{d}P(r). \right)$$

Note that the function g_2 is positive and strictly decreasing over $(0, \infty)$ with $\lim_{x \to 0^+} g_2(x) = \infty$ and $\lim_{x \to \infty} g_2(x) = 0$. On the other hand, the function h_2 is strictly decreasing over $(0, \infty)$ with $h_2(0) = 1$ and $h_2(x_0) > 0$. This follows because for $x \in (0, x_0]$,

$$\phi \int \left(\frac{xr}{xr + 1} \right)^2 \mathrm{d}P(r) \le \frac{x_0 b}{x_0 b + 1} \phi \int \left(\frac{xr}{xr + 1} \right) \mathrm{d}P(r) < \int \frac{\phi xr}{xr + 1} \, \mathrm{d}P(r) \le \int \frac{\phi x_0 r}{x_0 r + 1} \, \mathrm{d}P(r) = 1,$$

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

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

Thus, $-f'$, a product of two positive strictly decreasing functions, is strictly decreasing, and in turn, f' is strictly increasing. Moreover, $\lim_{x \to 0^+} f'(x) = -\infty$ and $f'(x_0) < 0$.

Part (4). The conclusion follows by noting that $h_2(x_0^\lambda) > h_2(x_0) > 0$ from (3). ■

Lemma 42 provides the continuity and limiting behavior of the function $\phi \mapsto v(-\lambda; \phi)$ for ridge regression ($\lambda > 0$). Lemma 43 does the same for ridgeless regression ($\lambda = 0$).

Lemma 42 (Continuity properties in the aspect ratio for ridge regression)**.** *Let $\lambda, a > 0$ and $b < \infty$ be real numbers. Let P be a probability measure supported on $[a, b]$. Consider the function $v(-\lambda; \cdot) : \phi \mapsto v(-\lambda; \phi)$, over $(0, \infty)$, where $v(-\lambda; \phi) > 0$ is the unique solution to the following fixed-point equation:*

$$\frac{1}{v(-\lambda; \phi)} = \lambda + \phi \int \frac{r}{1 + r v(-\lambda; \phi)} \mathrm{d}P(r) \tag{121}$$

Then the following properties hold:

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

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

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

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

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

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

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

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

Proof Proofs for the different parts appear below.

Part (1). Since P has positive support, we have

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

which implies that $0 < (\lambda + \phi b)^{-1} < v(-\lambda; \phi) < \lambda^{-1}$.

Part (2). Rearranging (121) yields

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

From Patil et al. (2022b, Lemma S.6.13), the function

$$h_1 : t \mapsto 1 - \int \frac{1}{1 + rt} \mathrm{d}P(r)$$

is strictly increasing and continuous over $(0, \infty)$, $\lim_{t\to 0} h_1(t) = 0$, and $\lim_{t\to\infty} h_1(t) = 1$. It is also positive from (1). Since $h_2 : t \mapsto 1/(1 - \lambda t)$ is positive, strictly increasing and continuous over $t \in (0, \lambda^{-1})$, we have that the function

$$T : t \mapsto \frac{1}{1 - \lambda t} \left(1 - \int \frac{1}{1 + rt} \mathrm{d}P(r)\right)$$

is strictly increasing and continuous over $(0, \lambda^{-1})$. By the continuous inverse theorem, we have T^{-1} is strictly increasing and continuous. Note that $v(-\lambda;\phi) = T^{-1}(\phi^{-1})$. Since $\phi \mapsto \phi^{-1}$ is continuous and strictly decreasing in ϕ, we have $\phi \mapsto v(-\lambda;\phi)$ is continuous and strictly decreasing over $\phi \in (0, \infty)$. Moreover, $\lim_{\phi\to 0^+} T^{-1}(\phi^{-1}) = \lambda^{-1}$ and $\lim_{\phi\to\infty} T^{-1}(\phi^{-1}) = 0$.

Part (3). From (2), $\phi \mapsto v(-\lambda;\phi)^{-2}$ is continuous in ϕ and

$$T_2 : \phi \mapsto \phi \int \frac{r^2}{(1 + r v(-\lambda;\phi))^2} \, \mathrm{d}P(r)$$

is also continuous in ϕ. Thus, the function $\widetilde{v}_v(-\lambda;\cdot)^{-1}$ is continuous. Note that

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

where the inequality holds because $rv(-\lambda;\phi)/(1+rv(-\lambda;\phi))$ is strictly positive and $P(r)$ has positive support. Then we have that $\phi \mapsto \widetilde{v}_v(-\lambda;\phi)^{-1} > 0$ and $\widetilde{v}_v(-\lambda;\cdot)$ is continuous over $(0, \infty)$. Since $\lim_{\phi\to 0^+} v(-\lambda;\phi) = \lambda^{-1}$, it follows that $\lim_{\phi\to 0^+} \widetilde{v}_v(-\lambda;\phi) = \lambda^{-2}$. Similarly, from $\lim_{\phi\to\infty} v(-\lambda;\phi) = 0$, $\lim_{\phi\to\infty} \phi v(-\lambda;\phi) = 1$ and the fact that

$$\lim_{\phi\to\infty} \int \frac{r^2}{(1 + r v(-\lambda;\phi))^2} \, \mathrm{d}P(r) \in [a^2, b^2],$$

it follows that

$$\lim_{\phi\to\infty} \widetilde{v}_v(-\lambda;\phi) = \lim_{\phi\to\infty} v(-\lambda;\phi)^2 \cdot \left(1 - v(-\lambda;\phi) \cdot \phi v(-\lambda;\phi) \cdot \int r^2 (1 + r v(-\lambda;\phi))^{-2} \, \mathrm{d}P(r)\right)^{-1} = 0.$$

Part (4). The continuity of $\widetilde{v}_b(-\lambda;\cdot)$ follows from the continuity of $v(-\lambda;\cdot)$ and $\widetilde{v}_v(-\lambda;\cdot)$. Note that

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

From the proof in (3), we have

$$\lim_{\phi\to 0^+}\frac{1}{1+\widetilde{v}_b(-\lambda;\phi)} = 1 - \lim_{\phi\to 0^+} v(-\lambda;\phi)\cdot\phi v(-\lambda;\phi)\cdot\int \frac{r^2}{(1+rv(-\lambda;\phi))^2}\,\mathrm{d}P(r) = 1$$
$$\lim_{\phi\to \infty}\frac{1}{1+\widetilde{v}_b(-\lambda;\phi)} = 1 - \lim_{\phi\to \infty} v(-\lambda;\phi)\cdot\phi v(-\lambda;\phi)\cdot\int \frac{r^2}{(1+rv(-\lambda;\phi))^2}\,\mathrm{d}P(r) = 1$$

and thus, $\lim_{\phi\to 0^+}\widetilde{v}_b(-\lambda;\phi) = \lim_{\phi\to\infty}\widetilde{v}_b(-\lambda;\phi) = 0$. ■

Lemma 43 (Continuity properties in the aspect ratio for ridgeless regression)**.** *Let $a > 0$ and $b < \infty$ be real numbers. Let P be a probability measure supported on $[a, b]$. Consider the function $v(0;\cdot) : \phi \mapsto v(0;\phi)$, over $(1,\infty)$, where $v(0;\phi) > 0$ is the unique solution to the followinn fixed-point equation:*

$$\frac{1}{\phi} = \int \frac{v(0;\phi)r}{1+v(0;\phi)r}\,\mathrm{d}P(r). \tag{122}$$

Then the following properties hold:

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

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

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

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

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

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

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

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

Lemma 44, adapted from Patil et al. (2022b), confirms the continuity and differentiability of the function $\lambda \mapsto v(-\lambda;\phi)$ on the closed interval $[0,\lambda_{\max}]$ for a certain constant $\lambda_{\max}$, provided $\phi \in (1,\infty)$. This ensures that $v(0;\phi) = \lim_{\lambda\to 0^+} v(-\lambda;\phi)$ is well-defined for $\phi > 1$, and also implies that related functions are bounded.

Lemma 44 (Differentiability properties in the regularization parameter for $\phi \in (1, \infty)$)**.** *Let $0 < a \le b < \infty$ be real numbers. Let P be a probability measure supported on $[a, b]$. Let $\phi \in (1, \infty)$ be a real number. Let $\Lambda = [0, \lambda_{\max}]$ for some constant $\lambda_{\max} \in (0, \infty)$. For $\lambda \in \Lambda$, let $v(-\lambda; \phi) > 0$ denote the solution to the fixed-point equation*

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

Then, the function $\lambda \mapsto v(-\lambda; \phi)$ is twice differentiable over Λ. Furthermore, over Λ, $v(-\lambda; \phi)$, $\partial/\partial\lambda[v(-\lambda; \phi)]$, and $\partial^2/\partial\lambda^2[v(-\lambda; \phi)]$ are bounded.

Lemma 45 (Substitutability of the fixed-point solution)**.** *Let $v : \mathbb{R}^{p\times p} \to \mathbb{R}$ and $f(v(\boldsymbol{C}), \boldsymbol{C}) : \mathbb{R}^{p\times p} \to \mathbb{R}^{p\times p}$ be a matrix function for matrix $\boldsymbol{C} \in \mathbb{R}^{p\times p}$ and $p \in \mathbb{N}$, that is continuous in the first augment with respect to operator norm. If $v(\boldsymbol{C}) \overset{a.s.}{=} v(\boldsymbol{D})$ such that $\boldsymbol{C}$ is independent of $\boldsymbol{D}$, then $f(v(\boldsymbol{C}), \boldsymbol{C}) \simeq f(v(\boldsymbol{D}), \boldsymbol{C}) \mid \boldsymbol{C}$.*

Proof For any matrix $\boldsymbol{T}$ whose trace norm is bounded by M, conditioning on $\{\boldsymbol{C}\}_{p\ge1}$, we have

$$\begin{aligned} |\operatorname{tr}[(f(v(\boldsymbol{C}), \boldsymbol{C}) - f(v(\boldsymbol{D}), \boldsymbol{C}))\boldsymbol{T}]| &\le \|f(v(\boldsymbol{C}), \boldsymbol{C}) - f(v(\boldsymbol{D}), \boldsymbol{C})\|_{\mathrm{op}} \operatorname{tr}(\boldsymbol{T}) \\ &\le M \|f(v(\boldsymbol{C}), \boldsymbol{C}) - f(v(\boldsymbol{D}), \boldsymbol{C})\|_{\mathrm{op}}. \end{aligned}$$

Since $v(\boldsymbol{C}) \xrightarrow{\text{a.s.}} v(\boldsymbol{D})$ and f is continuous in the first argument with respect to operator norm, we have $\lim_{p\to\infty} \|f(v(\boldsymbol{C}), \boldsymbol{C}) - f(v(\boldsymbol{D}), \boldsymbol{C})\|_{\mathrm{op}} = 0$. Thus,

$$\lim_{p\to\infty} |\operatorname{tr}[(f(v(\boldsymbol{C}), \boldsymbol{C}) - f(v(\boldsymbol{D}), \boldsymbol{C}))\boldsymbol{T}]| = 0,$$

conditioning on $\{\boldsymbol{C}\}_{p\ge1}$. ■

The lemma below specializes the solution to the fixed-point equations under the isotopic model.

Lemma 46 (Properties of the fixed-point solution with isotopic features)**.** *Let P be a probability measure supported on $\{a\}$ for $a > 0$. For $\lambda > 0$ and $\phi > 0$, the fixed-point equation*

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

has a closed-form solution given by:

$$v(-\lambda; \phi) = \frac{-(\lambda/a + \phi - 1) + \sqrt{(\lambda/a + \phi - 1)^2 + 4\lambda/a}}{2\lambda}.$$

Define $\widetilde{v}_b(-\lambda; \phi)$ and $\widetilde{v}_v(-\lambda; \phi)$ via the follow equations:

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

As $\lambda \to 0^+$, we have the following different cases:

$$
\begin{aligned}
&(1) \quad \phi \in (0,1): && v(0;\phi) = \infty, && \widetilde{v}_b(0;\phi) = \frac{\phi}{1-\phi}, && \widetilde{v}_v(0;\phi) = \infty,\\
&(2) \quad \phi = 1: && v(0;\phi) = \infty, && \widetilde{v}_b(0;\phi) = \infty, && \widetilde{v}_v(0;\phi) = \infty,\\
&(3) \quad \phi \in (1,\infty): && v(0;\phi) = \frac{1}{a(\phi-1)}, && \widetilde{v}_b(0;\phi) = \frac{1}{\phi-1}, && \widetilde{v}_v(0;\phi) = \frac{\phi}{a^2(\phi-1)^3},\\
&(4) \quad \phi = \infty: && v(0;\phi) = 0, && \widetilde{v}_b(0;\phi) = 0, && \widetilde{v}_v(0;\phi) = 0,
\end{aligned}
$$

Proof For $\phi \in (0,1)$, we have $v(0;\phi) = \lim_{\lambda \to 0^+} v(-\lambda;\phi) = \infty$. For $\phi > 1$,

$$
v(0;\phi) = \lim_{\lambda \to 0^+} v(-\lambda;\phi) = \frac{1}{2a} \lim_{\lambda \to 0^+} \left(-1 + \frac{\lambda/a + \phi + 1}{\sqrt{(\lambda/a + \phi - 1)^2 + 4\lambda/a}} \right) = \frac{1}{a(\phi-1)},
$$

by applying L'Hospital's rule for indeterminate forms. When $\phi = 1$, we have

$$
v(0;1) = \lim_{\lambda \to 0^+} v(-\lambda;1) = \lim_{\lambda \to 0^+} \frac{1}{2a} \left(-1 + \sqrt{1 + \frac{a}{\lambda}} \right) = \infty.
$$

Since $\widetilde{v}_b(0;\phi)$ and $\widetilde{v}_v(0;\phi)$ are continuous functions of $v(0;\phi)$, we have

$$
\widetilde{v}_v(0;\phi) = \begin{cases} \infty, & \phi \in (0,1] \\ \dfrac{\phi}{a^2(\phi-1)^3}, & \phi \in (1,\infty) \end{cases}
$$

and $\widetilde{v}_b(0;\phi) = 1/(\phi-1)$ for $\phi \in (1,\infty)$. For $\phi \in (0,1]$, we apply the L'Hospital' rule to obtain $\widetilde{v}_b(0;\phi) = \phi/(1-\phi)$. ■

Appendix I. Helper concentration results

I.1 Size of the intersection of randomly sampled datasets

In this section, we collect various helper results concerned with concentrations and convergences that are used in the proofs of Lemma 4, Lemmas 20, 21 and 25.

Below we recall the definition of a hypergeometric random variable, along with its mean and variance. See, e.g., Greene and Wellner (2017) for more related details.

Definition 47 (Hypergeometric random variable)**.** *A random variable X follows the hypergeometric distribution $X \sim \mathrm{Hypergeometric}(n, K, N)$ if the probability mass function of X is as follows:*

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

The expectation and variance of X are given by:

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

The following lemma provides tail bounds for the number of shared observations in two simple random samples adapted from (Hoeffding, 1963; Serfling, 1974). See also Greene and Wellner (2017).

Lemma 48 (Concentration bounds for the number of shared observations)**.** *For $n \in \mathbb{N}$, define $\mathcal{I}_k := \{\{i_1, i_2, \dots, i_k\} : 1 \le i_1 < i_2 < \dots < i_k \le n\}$. Let $I_1, I_2 \overset{\texttt{SRSWR}}{\sim} \mathcal{I}_k$, define the random variable $i_0^{\texttt{SRSWR}} := |I_1 \cap I_2|$ to be the number of shared samples, and define $i_0^{\texttt{SRSWOR}}$ accordingly. Then the following statements hold:*

(1) $i_0^{\texttt{SRSWR}}$ follows a binomial distribution, $i_0^{\texttt{SRSWR}} \sim \mathrm{Binomial}(k, k/n)$ with mean $\mathbb{E}[i_0^{\texttt{SRSWR}}] = k^2/n$. It holds that for all $t > 0$,

$$\mathbb{P}\left(i_0^{\texttt{SRSWR}} - \mathbb{E}[i_0^{\texttt{SRSWR}}] \ge kt\right) \le \exp\left(-2kt^2\right).$$

(2) $i_0^{\texttt{SRSWOR}}$ follows a hypergeometric distribution, $i_0^{\texttt{SRSWOR}} \sim \mathrm{Hypergeometric}(k, k, n)$ with mean $\mathbb{E}[i_0^{\texttt{SRSWOR}}] = k^2/n$. It holds that for all $t > 0$,

$$\mathbb{P}\left(i_0^{\texttt{SRSWOR}} - \mathbb{E}[i_0^{\texttt{SRSWOR}}] \ge kt\right) \le \exp\left(-\frac{2nkt^2}{n-k+1}\right). \tag{123}$$

The following lemma characterizes the limiting proportions of shared observations in two simple random samples under proportional asymptotics when both the subsample and full data sizes tend to infinity.

Lemma 49 (Asymptotic proportions of the shared observations)**.** *Consider the setting in Lemma 48. Let $\{k_m\}_{m=1}^\infty$ and $\{n_m\}_{m=1}^\infty$ be two sequences of positive integers such that n_m is strictly increasing in m, $n_m^\nu \le k_m \le n_m$ for some constant $\nu \in (0,1)$, and $k_m/n_m \to \omega_s \in [0,1]$. Then, $i_0^{\texttt{SRSWR}}/k_m \xrightarrow{\text{a.s.}} \omega_s$, and $i_0^{\texttt{SRSWOR}}/k_m \xrightarrow{\text{a.s.}} \omega_s$.*

Proof Proofs for the two parts are split below.

Part (1). For all $\delta > 0$,

$$\sum_{m=1}^\infty \mathbb{P}\left(\frac{1}{k_m}|i_0^{\texttt{SRSWR}} - \mathbb{E}[i_0^{\texttt{SRSWR}}]| > \delta\right) \le 2\sum_{m=1}^\infty \exp\left(-2k_m\delta^2\right).$$

Because $k_m, n_m \to \infty$ and $k_m = \Omega(n_m^\nu)$, there exists $m_0 \in \mathbb{N}$, such that for all $m > m_0$, $\exp(-2k_m\delta^2) \le n_m^{-(1+\nu)}$. Thus,

$$\sum_{m=1}^\infty \mathbb{P}\left(\frac{1}{k_m}|i_0^{\texttt{SRSWR}} - \mathbb{E}[i_0^{\texttt{SRSWR}}]| > \delta\right) \le 2\sum_{m=1}^{m_0} \exp\left(-2k_m\delta^2\right) + 2\sum_{m=m_0}^\infty \frac{1}{n_m^{1+\nu}} < \infty.$$

By the Borel–Cantelli lemma, we have

$$\frac{i_0^{\texttt{SRSWR}}}{k_m} - \frac{\mathbb{E}[i_0^{\texttt{SRSWR}}]}{k_m} \xrightarrow{\text{a.s.}} 0.$$

As $\lim_{m\to\infty} \mathbb{E}[i_0^{\texttt{SRSWR}}]/k_m = \lim_{m\to\infty} k_m/n_m = \omega_s$, we further have $i_0^{\texttt{SRSWR}}/k_m \xrightarrow{\text{a.s.}} \omega_s$.

Part (2). Note that

$$\mathbb{P}\left(i_0^{\texttt{SRSWOR}} - \mathbb{E}[i_0^{\texttt{SRSWOR}}] \geq kt\right) \leq \exp\left(-\frac{2nkt^2}{n-k+1}\right) \leq \exp\left(-2kt^2\right).$$

The conclusion then follows analogously, as in Part 1. ■

I.2 Convergence of Ces̀aro-type mean and max for triangular array

In this section, we collect a helper lemma on deducing almost sure convergence of a Ces̀aro-type mean from almost sure convergence of the original sequence. It is used in the proof of Proposition 3 and Lemma 4.

Lemma 50 (Convergence of conditional expectation). *For $n \in \mathbb{N}$, suppose $\{R_{n,\ell}\}_{\ell=1}^{N_n}$ is a set of N_n random variables defined over the probability space $(\Omega, \mathcal{F}, \mathbb{P})$, with $1 < N_n < \infty$ almost surely. If there exists a constant c such that $R_{n,p_n} \xrightarrow{\text{a.s.}} c$ for all deterministic sequences $\{p_n \in [N_n]\}_{n=1}^{\infty}$, then the following statements hold:*

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

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

Proof [Proof of Lemma 50] Proofs for the two parts are split below.

Part (1). We concatenate the sets $\{R_{n,\ell}\}_{\ell=1}^{N_n}$ for all $n \in \mathbb{N}$ to form a new sequence

$$W = (W_1, W_2, \cdots) = (R_{1,1}, \cdots, R_{1,N_1}, R_{2,1}, \cdots, R_{2,N_2}, \cdots).$$

That is, $W_t = R_{n,\ell}$ for $t = \sum_{j=1}^{n} N_j + \ell$. See Figure 13 for an illustration. Because $N_n \to \infty$ if and only if $n \to \infty$ if and only if $t \to \infty$, it holds that $W_t \xrightarrow{\text{a.s.}} c$ as $t \to \infty$. Then, by Shiryaev (2016, Chapter 2, Section 10, Theorem 1), we have that for all $\epsilon > 0$,

$$\lim_{s\to\infty} \mathbb{P}\left(\bigcup_{t=s}^{\infty}\{\omega \in \Omega : |W_t(\omega) - c| > \epsilon\}\right) = 0.$$

Now, for $s \in \mathbb{N}$, let m be the smallest natural number such that $\sum_{j=1}^{m} N_j \geq s$. Since

$$\begin{aligned}\bigcup_{t=s}^{\infty}\{\omega \in \Omega : |W_t(\omega) - c| > \epsilon\} &\supseteq \bigcup_{n=m}^{\infty}\bigcup_{\ell=1}^{N_n}\{\omega \in \Omega : |R_{n,\ell}(\omega) - c| > \epsilon\} \\ &= \bigcup_{n=m}^{\infty}\left\{\omega \in \Omega : \max_{\ell \in [N_n]} |R_{n,\ell}(\omega) - c| > \epsilon\right\}.\end{aligned}$$

We further have

$$0 \leq \lim_{m\to\infty} \mathbb{P}\left(\bigcup_{n=m}^{\infty}\left\{\omega \in \Omega : \max_{\ell \in [N_n]} |R_{n,\ell}(\omega) - c| > \epsilon\right\}\right) \leq \lim_{s\to\infty} \mathbb{P}\left(\bigcup_{t=s}^{\infty}\{\omega \in \Omega : |W_t(\omega) - c| > \epsilon\}\right) = 0,$$

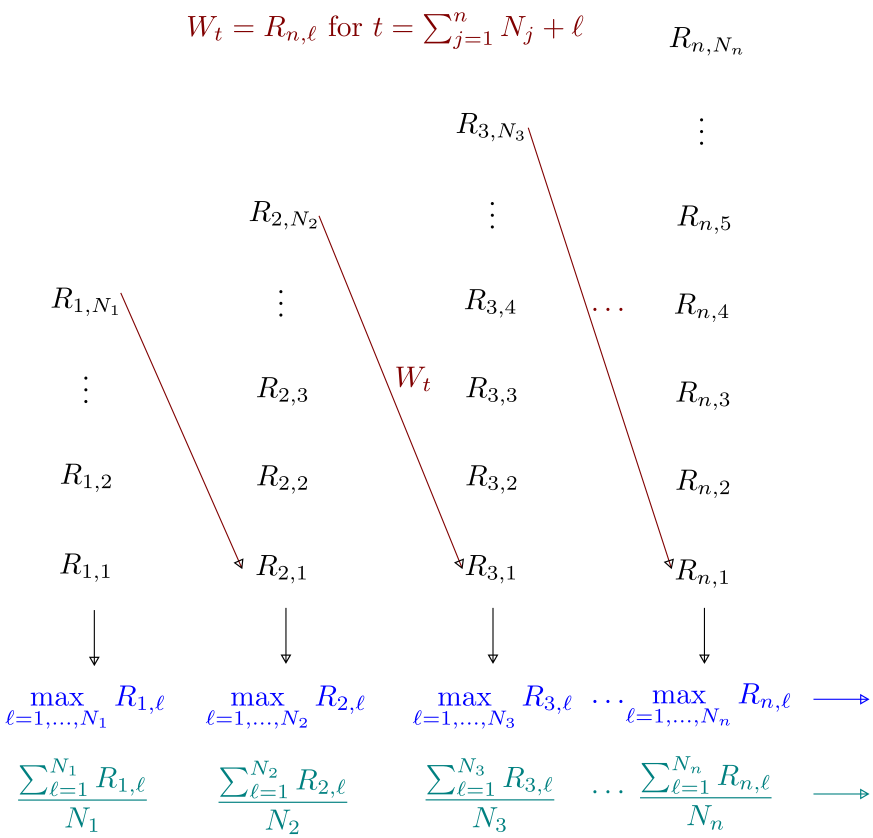

Figure 13: Illustration of the concatenated sequence $\{W_t\}$ (in maroon) constructed from the triangle array $\{R_{n,\ell}\}_{\ell=1}^{N_n}, n \in \mathbb{N}$ (in black), used in the proof of Lemma 50, along with the max sequence (in blue) and the average sequence (in teal).

or in other words,

$$\lim_{m\to\infty} \mathbb{P}\left(\bigcup_{n=m}^{\infty} \left\{ \omega \in \Omega : \max_{\ell\in[N_n]} |R_{n,\ell}(\omega) - c| > \epsilon \right\} \right) = 0.$$

Thus, we have that $\max_{\ell\in[N_n]} |R_{n,\ell}(\omega) - c| \xrightarrow{\text{a.s.}} 0$ by Shiryaev (2016, Chapter 2, Section 10, Theorem 1).

Part (2). We will use the first part. Note that by triangle inequality,

$$\left| N_n^{-1} \sum_{\ell=1}^{N_n} R_{n,\ell} - c \right| \le N_n^{-1} \sum_{\ell=1}^{N_n} |R_{n,\ell} - c| \le \max_{\ell\in[N_n]} |R_{n,\ell}(\omega) - c|\,.$$

Invoking the first part, we have that $N_n^{-1} \sum_{\ell=1}^{N_n} R_{n,\ell} \xrightarrow{\text{a.s.}} c$. ∎

Appendix J. Additional numerical illustrations

J.1 Additional illustrations for Theorem 6

J.1.1 Prediction risk curves for subagged ridgeless and ridge predictors with varying M

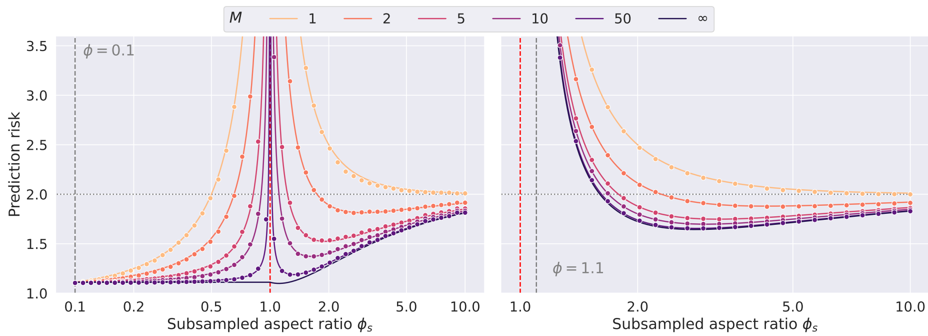

Figure 14: Asymptotic prediction risk curves in (19) for ridgeless predictors ($\lambda = 0$), under model (M-ISO-LI) when $\rho^2 = 1$ and $\sigma^2 = 1$ for varying bag size $k = \lfloor p/\phi_s \rfloor$ and number of bags M. The null risk is marked as a dotted line. For each value of M, the points denote finite-sample risks averaged over 100 dataset repetitions, with $n = 1000$ and $p = \lfloor n\phi \rfloor$. The left and the right panels correspond to the cases when $p < n$ ($\phi = 0.1$) and $p > n$ ($\phi = 1.1$), respectively.

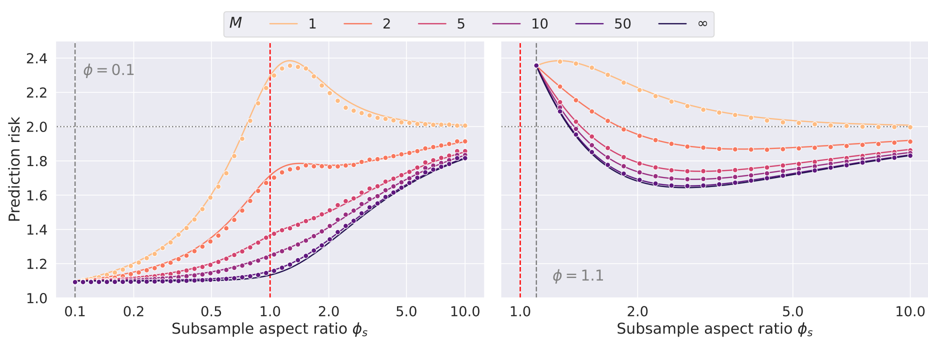

Figure 15: Asymptotic prediction risk curves in (19) for subagged ridge predictors ($\lambda = 0.1$), under model (M-ISO-LI) when $\rho^2 = 1$ and $\sigma^2 = 1$ for varying bag size $k = \lfloor p/\phi_s \rfloor$ and number of bags M. The null risk is marked as a dotted line. For each value of M, the points denote finite-sample risks averaged over 100 dataset repetitions, with $n = 1000$ and $p = \lfloor n\phi \rfloor$. The left and the right panels correspond to the cases when $p < n$ ($\phi = 0.1$) and $p > n$ ($\phi = 1.1$), respectively.

J.1.2 Bias-variance curves for subagged ridgeless and ridge predictors with varying M

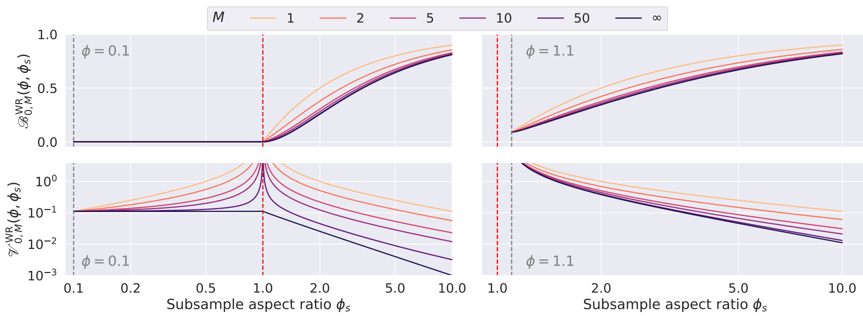

Figure 16: Asymptotic bias and variance curves in (22) for subagged ridgeless predictors ($\lambda = 0$), under model (M-ISO-LI) when $\rho^2 = 1$ and $\sigma^2 = 0.25$ for varying bag size $k = \lfloor p/\phi_s \rfloor$ and number of bags M. The left and the right panels correspond to the cases when $p < n$ ($\phi = 0.1$) and $p > n$ ($\phi = 1.1$), respectively. The values of $\mathscr{V}_{0,M}^{\mathtt{sub}}(\phi, \phi_s)$ are shown on a log-10 scale.

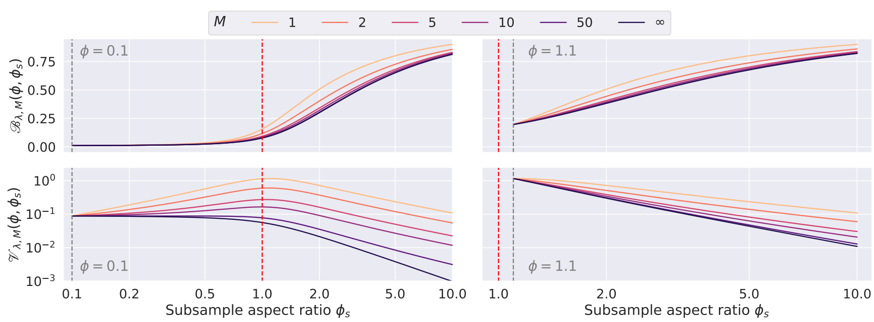

Figure 17: Asymptotic bias and variance curves in (22) for subagged ridge predictors ($\lambda = 0.1$), under model (M-ISO-LI) when $\rho^2 = 1$ and $\sigma^2 = 1$ for varying bag size $k = \lfloor p/\phi_s \rfloor$ and number of bags M. The left and the right panels correspond to the cases when $p < n$ ($\phi = 0.1$) and $p > n$ ($\phi = 1.1$), respectively. The values of $\mathscr{V}_{0,M}^{\mathtt{sub}}(\phi, \phi_s)$ are shown in log-10 scale.

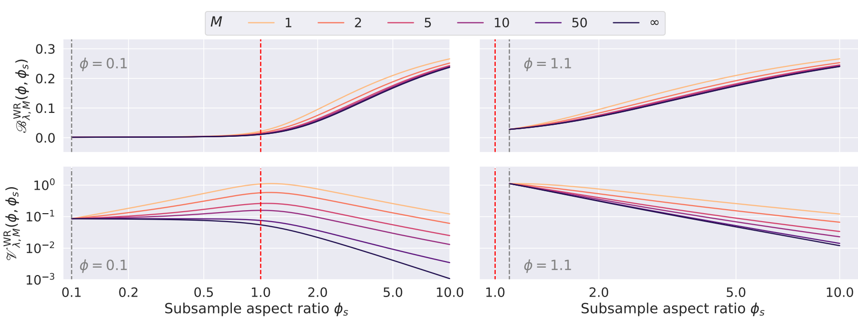

Figure 18: Asymptotic bias and variance curves in (22) for subagged ridge predictors ($\lambda = 0.1$), under model (M-AR1-LI) when $\rho^2 = 1$ and $\sigma^2 = 1$ for varying bag size $k = \lfloor p/\phi_s \rfloor$ and number of bags M. The left and the right panels correspond to the cases when $p < n$ ($\phi = 0.1$) and $p > n$ ($\phi = 1.1$), respectively. The values of $\mathscr{V}^{\mathtt{sub}}_{0,M}(\phi, \phi_s)$ are shown in log-10 scale.

J.1.3 Bias-variance curves for subagged ridge predictors with varying λ ($M = 1$)

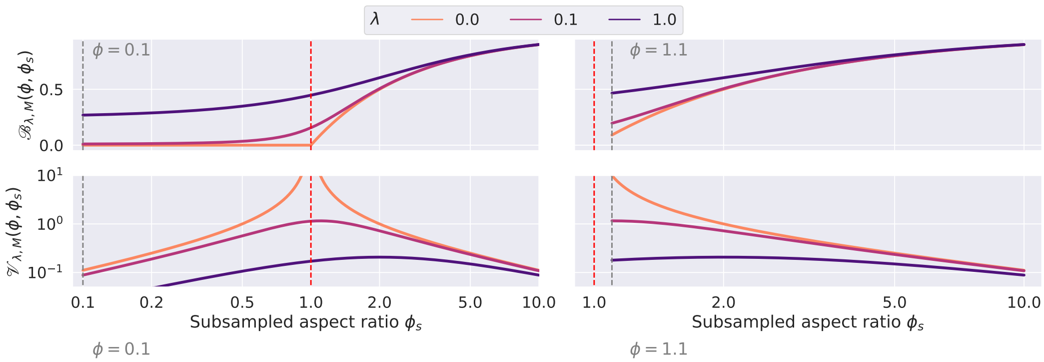

Figure 19: Asymptotic bias and variance curves in (22) for subagged ridge and ridgeless predictors with number of bags $M = 1$, under model (M-ISO-LI) when $\rho^2 = 1$ and $\sigma^2 = 1$ for varying regularization parameter λ. The left and the right panels correspond to the cases when $p < n$ ($\phi = 0.1$) and $p > n$ ($\phi = 1.1$), respectively. The values of $\mathscr{V}^{\mathsf{sub}}_{0,M}(\phi, \phi_s)$ are shown in log-10 scale.

J.1.4 Bias-variance curves for subagged ridge predictors with varying λ ($M = \infty$)

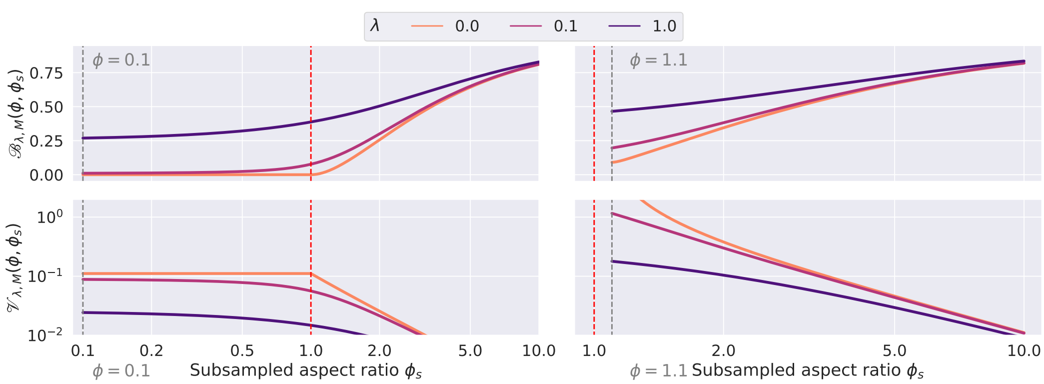

Figure 20: Asymptotic bias and variance curves in (22) for subagged ridge and ridgeless predictors with number of bags $M = \infty$, under model (M-ISO-LI) when $\rho^2 = 1$ and $\sigma^2 = 1$ for varying regularization parameter λ. The left and the right panels correspond to the cases when $p < n$ ($\phi = 0.1$) and $p > n$ ($\phi = 1.1$), respectively. The values of $\mathscr{V}^{\mathsf{sub}}_{0,M}(\phi, \phi_s)$ are shown in log-10 scale.

J.2 Additional illustrations for Theorem 8

J.2.1 Prediction risk curves for splagged ridgeless and ridge predictors with varying M

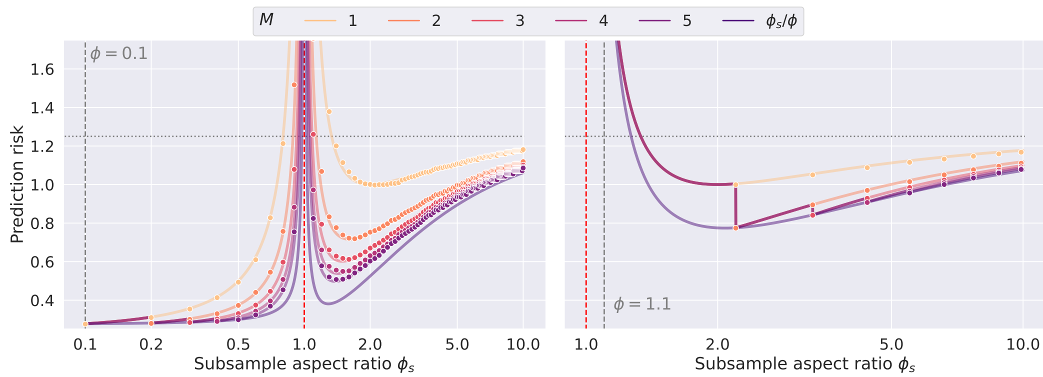

Figure 21: Asymptotic prediction risk curves in (28) for splagged ridgeless predictors ($\lambda = 0$), under model (M-ISO-LI) when $\rho^2 = 1$ and $\sigma^2 = 0.25$ for varying bag size $k = \lfloor p/\phi_s \rfloor$ and number of bags M without replacement. The left and the right panels correspond to the cases when $p < n$ ($\phi = 0.1$) and $p > n$ ($\phi = 1.1$), respectively. The null risk is marked as a dotted line. For each value of M, the points denote finite-sample risks averaged over 100 dataset repetitions, with $n = 1000$ and $p = \lfloor n\phi \rfloor$.

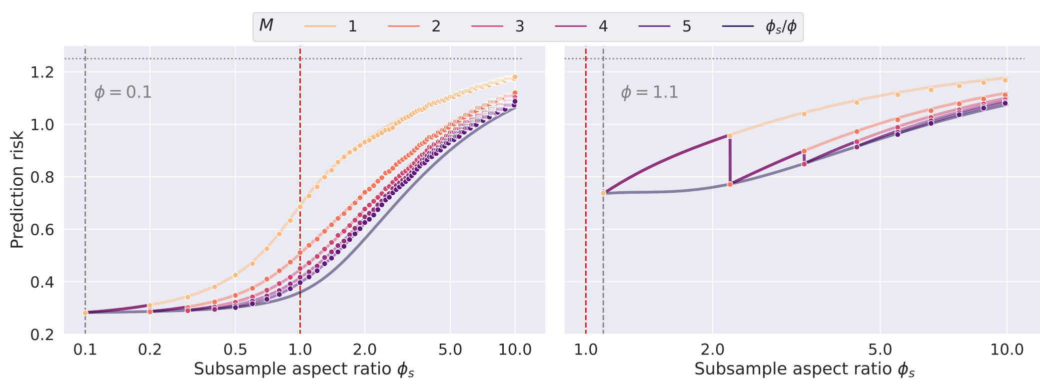

Figure 22: Asymptotic prediction risk curves in (28) for splagged ridge predictors ($\lambda = 0.1$), under model (M-ISO-LI) when $\rho^2 = 1$ and $\sigma^2 = 0.25$ for varying bag size $k = \lfloor p/\phi_s \rfloor$ and number of bags M without replacement. The left and the right panels correspond to the cases when $p < n$ ($\phi = 0.1$) and $p > n$ ($\phi = 1.1$), respectively. The null risk is marked as a dotted line. For each value of M, the points denote finite-sample risks averaged over 100 dataset repetitions, with $n = 1000$ and $p = \lfloor n\phi \rfloor$.

J.2.2 Bias-variance curves for ridgeless and ridge predictors with varying M

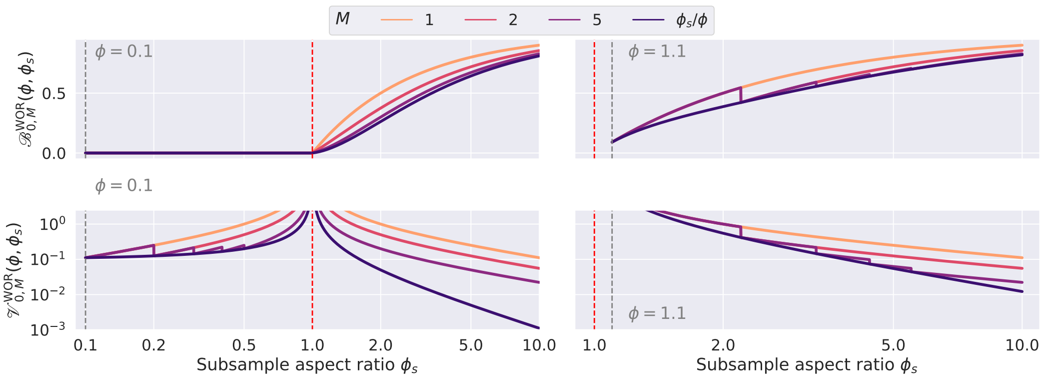

Figure 23: Asymptotic bias and variance curves in (22) for splagged ridgeless predictors ($\lambda = 0$), under model (M-ISO-LI) when $\rho^2 = 1$ and $\sigma^2 = 1$ for varying bag size $k = \lfloor p/\phi_s \rfloor$ and number of bags M without replacement. The left and the right panels correspond to the cases when $p < n$ ($\phi = 0.1$) and $p > n$ ($\phi = 1.1$), respectively. The values of $\mathscr{V}^{\mathtt{spl}}_{0,M}(\phi, \phi_s)$ are shown in log-10 scale.

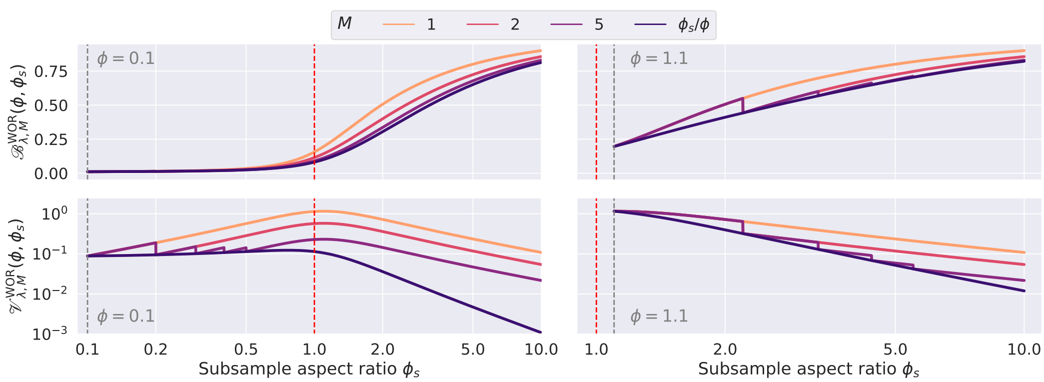

Figure 24: Asymptotic bias and variance curves in (22) for splagged ridge predictors ($\lambda = 0.1$), under model (M-ISO-LI) when $\rho^2 = 1$ and $\sigma^2 = 1$ for varying bag size $k = \lfloor p/\phi_s \rfloor$ and number of bags M without replacement. The left and the right panels correspond to the cases when $p < n$ ($\phi = 0.1$) and $p > n$ ($\phi = 1.1$), respectively. The values of $\mathscr{V}^{\mathtt{spl}}_{\lambda,M}(\phi, \phi_s)$ are shown in log-10 scale.

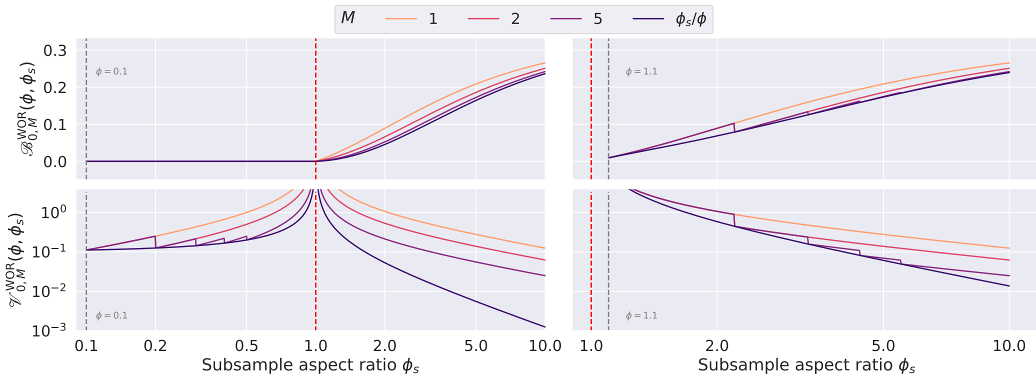

Figure 25: Asymptotic bias and variance curves in (22) for splagged ridgeless predictors ($\lambda = 0$), under model (M-AR1-LI) when $\rho_{\mathrm{ar1}} = 0.25$ and $\sigma^2 = 1$ for varying bag size $k = \lfloor p/\phi_s \rfloor$ and number of bags M without replacement. The left and the right panels correspond to the cases when $p < n$ ($\phi = 0.1$) and $p > n$ ($\phi = 1.1$), respectively. The values of $\mathscr{V}_{0,M}^{\mathsf{spl}}(\phi, \phi_s)$ are shown in log-10 scale.

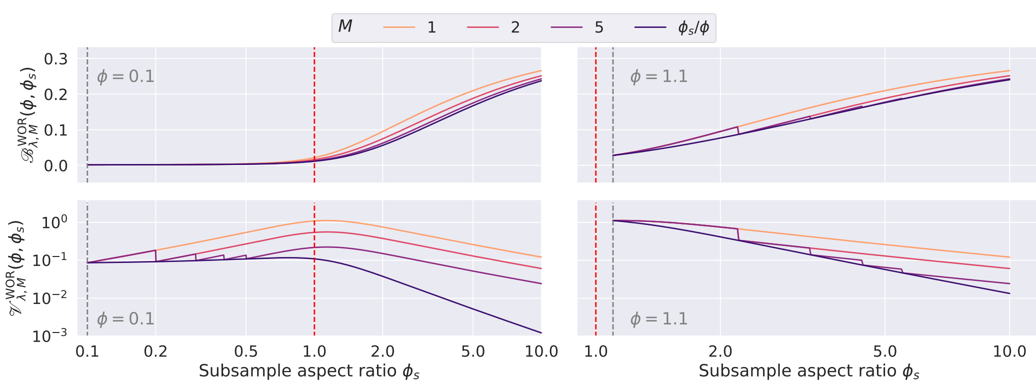

Figure 26: Asymptotic bias and variance curves in (22) for splagged ridge predictors ($\lambda = 0.1$), under model (M-AR1-LI) when $\rho_{\mathrm{ar1}} = 0.25$ and $\sigma^2 = 1$ for varying bag size $k = \lfloor p/\phi_s \rfloor$ and number of bags M without replacement. The left and the right panels correspond to the cases when $p < n$ ($\phi = 0.1$) and $p > n$ ($\phi = 1.1$), respectively. The values of $\mathscr{V}_{0,M}^{\mathsf{spl}}(\phi, \phi_s)$ are shown in log-10 scale.

J.3 Additional illustrations for Theorem 10

J.3.1 Risk monotonization for subagged ridgeless and ridge predictors

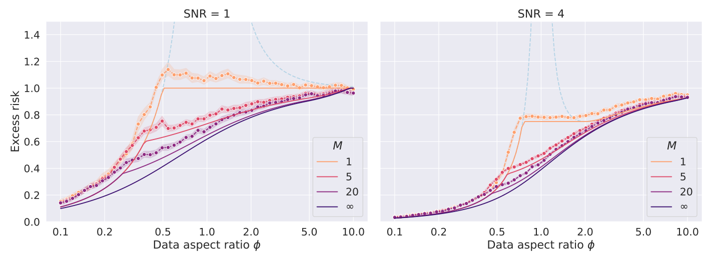

Figure 27: Asymptotic excess risk curves for cross-validated subagged ridgeless predictors ($\lambda = 0$), under model (M-ISO-LI) when $\rho^2 = 1$ for varying SNR, subsample sizes $k = \lfloor p/\phi_s \rfloor$, and numbers of bags M with replacement. The left and the right panels correspond to the cases when $\texttt{SNR} = 1$ and 4, respectively. The null risk is marked as a dotted line, and the risk for the unbagged ridgeless predictor is denoted by the dashed line. For each value of M, the points denote finite-sample risks, and the shaded regions denote the values within one standard deviation, with $n = 1000$, $n_{\rm te} = 63$, and $p = \lfloor n\phi \rfloor$.

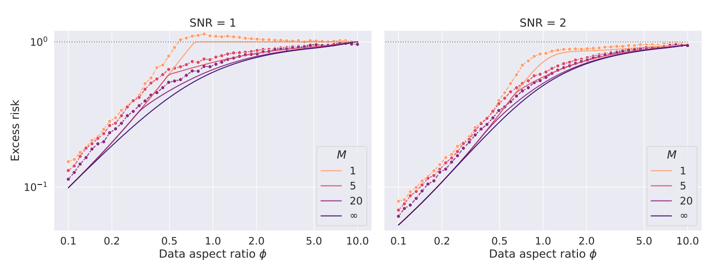

Figure 28: Asymptotic prediction risk curves for cross-validated subagged ridge predictors ($\lambda = 0.1$), under model (M-ISO-LI) when $\rho^2 = 1$ for varying SNR, subsample sizes $k = \lfloor p/\phi_s \rfloor$ and numbers of bags M with replacement. The left and the right panels correspond to the cases when $\texttt{SNR} = 1$ and 2, respectively. The null risk is marked as a dotted line. For each value of M, the points denote finite-sample risks averaged over 100 dataset repetitions, with $n = 1000$, $n_{\rm te} = 63$, and $p = \lfloor n\phi \rfloor$.

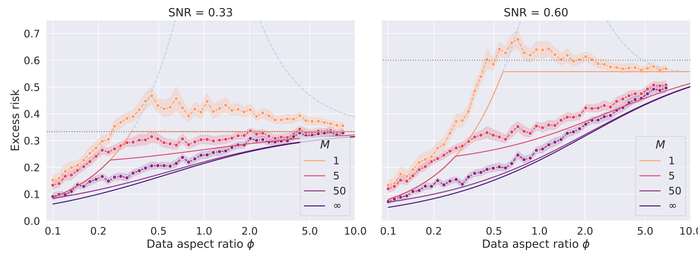

Figure 29: Asymptotic excess risk curves for cross-validated subagged ridge predictors ($\lambda = 0.1$), under model (M-AR1-LI) when $\sigma^2 = 1$ for varying SNR, subsample sizes $k = \lfloor p/\phi_s \rfloor$ and numbers of bags M. The left and the right panels correspond to the cases when $\texttt{SNR} = 0.33$ ($\rho_{\text{ar1}} = 0.25$) and 0.6 ($\rho_{\text{ar1}} = 0.5$), respectively. The excess null risk is marked as a dotted line, and the risk for the unbagged ridgeless predictor is denoted by the dashed line. For each value of M, the points denote finite-sample risks averaged over 100 dataset repetitions, and the shaded regions denote the values within one standard deviation, with $n = 1000$, $n_{\text{te}} = 63$, and $p = \lfloor n\phi \rfloor$.

J.3.2 Risk monotonization for splagged ridgeless and ridge predictors

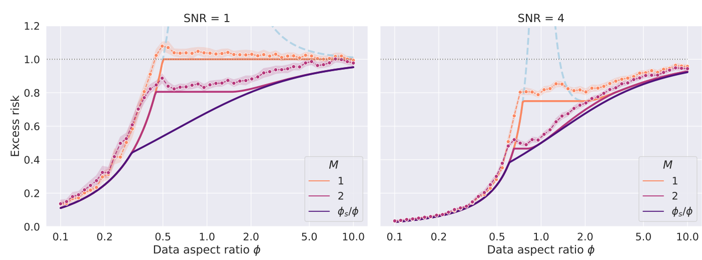

Figure 30: Asymptotic excess risk curves for cross-validated splagged ridgeless predictors ($\lambda = 0$), under model (M-ISO-LI) when $\rho^2 = 1$ for varying SNR, subsample sizes $k = \lfloor p/\phi_s \rfloor$, and numbers of bags M without replacement. The left and the right panels correspond to the cases when $\texttt{SNR} = 1$ and 4, respectively. The null risk is marked as a dotted line, and the risk for the unbagged ridgeless predictor is denoted by the dashed line. For each value of M, the points denote finite-sample risks averaged over 100 dataset repetitions, and the shaded regions denote the values within one standard deviation, with $n = 1000$, $n_{\text{te}} = 63$, and $p = \lfloor n\phi \rfloor$.

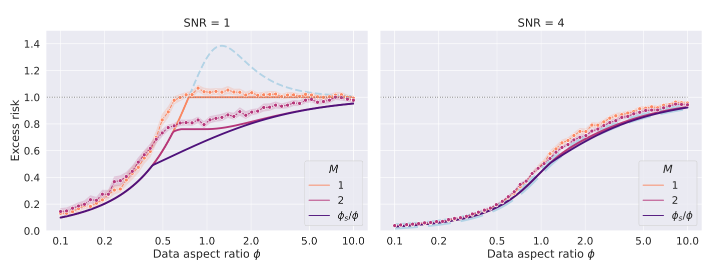

Figure 31: Asymptotic prediction risk curves for cross-validated splagged ridge predictors ($\lambda = 0.1$), under model (M-ISO-LI) when $\rho^2 = 1$ for varying SNR, subsample sizes $k = \lfloor p/\phi_s \rfloor$, and numbers of bags M without replacement. The left and the right panels correspond to the cases when $\texttt{SNR} = 1$ and 4, respectively. The null risk is marked as a dotted line. For each value of M, the points denote finite-sample risks averaged over 100 dataset repetitions, with $n = 1000$, $n_{\text{te}} = 63$, and $p = \lfloor n\phi \rfloor$.

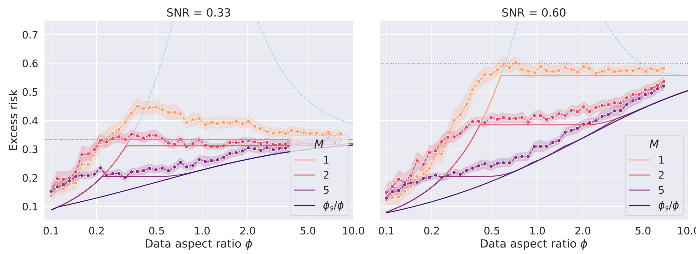

Figure 32: Asymptotic excess risk curves for cross-validated splagged ridge predictors ($\lambda = 0.1$), under model (M-AR1-LI) when $\sigma^2 = 1$ for varying SNR, subsample sizes $k = \lfloor p/\phi_s \rfloor$ and numbers of bags M. The left and the right panels correspond to the cases when SNR $= 0.33$ ($\rho_{\text{ar1}} = 0.25$) and 0.6 ($\rho_{\text{ar1}} = 0.5$), respectively. The excess null risk is marked as a dotted line, and risk for the unbagged ridgeless predictor is denoted by the dashed line. For each value of M, the points denote finite-sample risks averaged over 100 dataset repetitions and the shaded regions denote the values within one standard deviation, with $n = 1000$, $n_{\text{te}} = 63$, and $p = \lfloor n\phi \rfloor$.

J.4 Additional illustrations in Section 5

J.4.1 Subagging with replacement and splagging without replacement

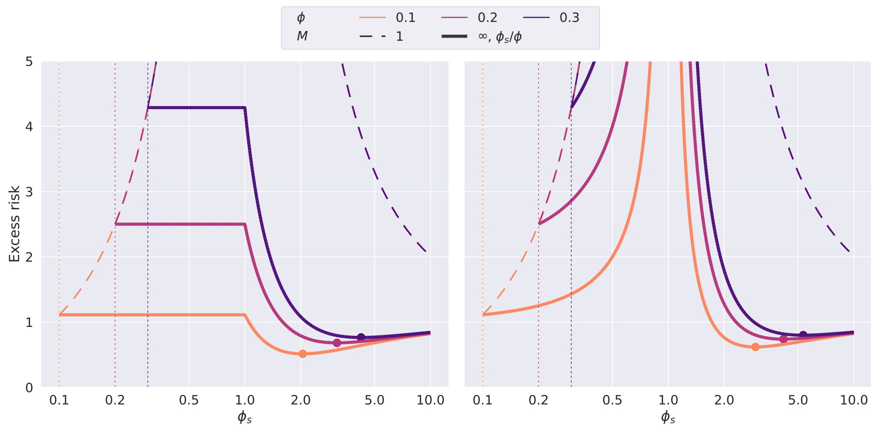

Figure 33: Asymptotic excess risk (the difference between the prediction risk and the noise level σ^2) curves of bagged ridgeless predictors ($\lambda = 0$) for subagging (left panel) and splagging (right panel), under model (M-ISO-LI) when $\rho^2 = 1$ and $\mathtt{SNR} = 0.1$, for varying ϕ ($p < n$), bag size $k = \lfloor p/\phi_s \rfloor$ and number of bags M. The solid lines represent the optimal risks with respect to M for either with replacement ($M = \infty$) or without replacement ($M = \phi_s/\phi$); the dashed lines represent the risks for $M = 1$; the dotted lines indicates the aspect ratio ϕ of the full dataset; the solid dots represent the optimal risk with respect to both M and ϕ_s.

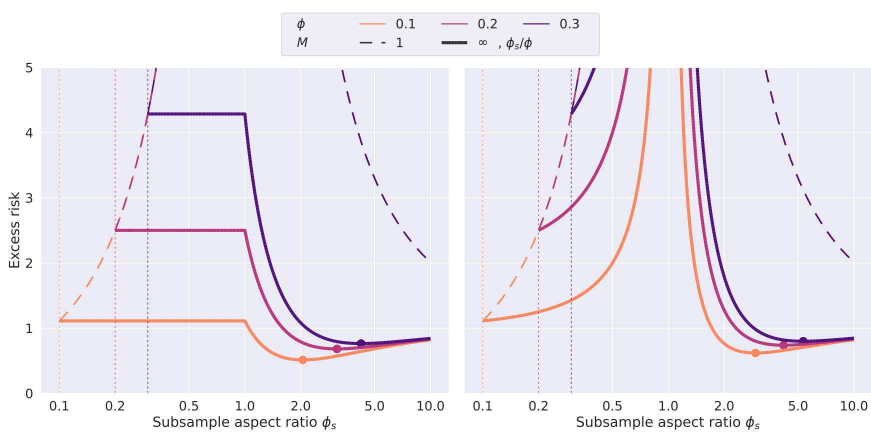

Figure 34: Asymptotic excess risk (the difference between the prediction risk and the noise level σ^2) curves of bagged ridgeless predictors ($\lambda = 0$) for subagging (left panel) and splagging (right panel), under model (M-ISO-LI) when $\rho^2 = 1$ and $\mathtt{SNR} = 0.5$, for varying ϕ ($p \geq n$), bag size $k = \lfloor p/\phi_s \rfloor$ and number of bags M. The solid lines represent the optimal risks with respect to M for either with replacement ($M = \infty$) or without replacement ($M = \phi_s/\phi$); the dashed lines represent the risks for $M = 1$; the dotted lines indicates the aspect ratio ϕ of the full dataset; the solid dots represent the optimal risk with respect to both M and ϕ_s.

References

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

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.

Bogachev, V. I. (2007). *Measure Theory*. Springer. First edition.

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

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

Durrett, R. (2010). *Probability: Theory and Examples*. Cambridge University Press. Fourth edition.

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

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

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.

Hoeffding, W. (1963). Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, pages 13–30.

Lee, A. J. (1990). *U-statistics: Theory and Practice*. Routledge. First edition.

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

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

Resnick, S. (2019). *A Probability Path*. Springer.

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

Sen, P. K. (1970). The Hájek-Rényi inequality for sampling from a finite population. *Sankhyā: The Indian Journal of Statistics, Series A*, pages 181–188.

Serfling, R. J. (1974). Probability inequalities for the sum in sampling without replacement. *The Annals of Statistics*, pages 39–48.

Shiryaev, A. N. (2016). *Probability-1*. Springer. Third edition.