Extrapolated cross-validation for randomized ensembles

Jin-Hong Du[†‡] Pratik Patil[§] Kathryn Roeder[†] Arun Kumar Kuchibhotla[†]

Abstract

Ensemble methods such as bagging and random forests are ubiquitous in various fields, from finance to genomics. Despite their prevalence, the question of the efficient tuning of ensemble parameters has received relatively little attention. This paper introduces a cross-validation method, ECV (Extrapolated Cross-Validation), for tuning the ensemble and subsample sizes in randomized ensembles. Our method builds on two primary ingredients: initial estimators for small ensemble sizes using out-of-bag errors and a novel risk extrapolation technique that leverages the structure of prediction risk decomposition. By establishing uniform consistency of our risk extrapolation technique over ensemble and subsample sizes, we show that ECV yields δ-optimal (with respect to the oracle-tuned risk) ensembles for squared prediction risk. Our theory accommodates general predictors, only requires mild moment assumptions, and allows for high-dimensional regimes where the feature dimension grows with the sample size. As a practical case study, we employ ECV to predict surface protein abundances from gene expressions in single-cell multiomics using random forests under a computational constraint on the maximum ensemble size. Compared to sample-split and K-fold cross-validation, ECV achieves higher accuracy by avoiding sample splitting. Meanwhile, its computational cost is considerably lower owing to the use of the risk extrapolation technique.

Keywords: Ensemble learning; Bagging; Random forest; Risk extrapolation; Tuning and model selection; Distributed learning.

[†]Department of Statistics and Data Science, Carnegie Mellon University, Pittsburgh, PA 15213, USA.
[‡]Machine Learning Department, Carnegie Mellon University, Pittsburgh, PA 15213, USA.
[§]Department of Statistics, University of California, Berkeley, CA 94720, USA.

1 Introduction

Bagging and its variants are popular randomized ensemble methods in statistics and machine learning. These methods combine multiple models, each fitted on different bootstrapped or subsampled datasets, to improve prediction accuracy and stability (Breiman, 1996; Pugh, 2002), which is well-suited for large-scale distributed computation. The success of these methods lies in the careful tuning of key parameters: the *ensemble size* M and the *subsample size* k (Hastie et al., 2009). As M grows, the predictive accuracy improves while prediction variance decreases and stabilizes, a concept known as algorithmic convergence (Lopes, 2019; Lopes et al., 2020). However, in the era of big data (Politis, 2023), achieving a precise approximation in the infinity ensemble is challenging due to computational costs. This necessitates the selection of a suitable value of M to strike a balance between data-dependent considerations and budget constraints, without the requirement for it to scale proportionally with the sample size. Further, the number of subsampled/bootstrapped observations, k, used for each predictor plays a crucial role in ensemble learning (Martínez-Muñoz and Suárez, 2010). In low-dimensional scenarios, a smaller k can yield consistent results (Politis and Romano, 1994; Bickel et al., 1997); however, in high-dimensional scenarios, the prediction risk may not have a straightforward monotonic relationship with subsample size, exhibiting instead multiple descent behaviors (Hastie et al., 2022; Chen et al., 2020; Patil et al., 2022). By restricting the subsample size, there has been some theoretical work on random forests showing that consistency and/or asymptotic normality can be achieved under certain regularity conditions (Scornet et al., 2015; Mentch and Hooker, 2016; Peng et al., 2022; Wager and Athey, 2018). Selecting the right subsample size is thus also of paramount importance for optimal predictive performance.

Several strategies have been proposed to tune either the ensemble size M or the subsample size k. For instance, to choose an ensemble size M, a variance stabilization strategy is proposed by Lopes (2019) and Lopes et al. (2020). This approach relies on the convergence rate of variance or quantile estimators, using these metrics to gauge the point at which the ensemble's performance stabilizes as M approaches infinity. Such an approach effectively helps reduce computational expenses. On the other hand, determining the optimal subsample size k is a more difficult task and generally tuned by standard cross-validation (CV) methods. As one of the most basic of the CV methods, sample-split CV estimates the predictive risk of every predictor associated with a configuration of parameters using independent hold-out observations (Patil et al., 2023). Another commonly used CV method, K-fold CV, repeatedly fits each candidate predictor on K different subsets of the data and uses their average to estimate the prediction risk.

While these aforementioned methods offer ways to tune M and k, they have several drawbacks. In terms of turning over the ensemble size M, the specialized method proposed by Lopes (2019) and Lopes et al. (2020) involves monitoring the variability of the test errors as a function of M. However, as this approach focuses solely on *the scale of variance of the risk* rather than the *prediction risk* itself, it does not provide any suboptimality guarantee compared to the optimal risk of an infinite ensemble (LeJeune et al., 2020). Furthermore, the method does not provide any estimators for the prediction risks for any finite ensemble size M. In terms of tuning the subsample size k, the traditional CV methods such as sample-split CV and K-fold CV can be significantly impacted by finite-sample effects due

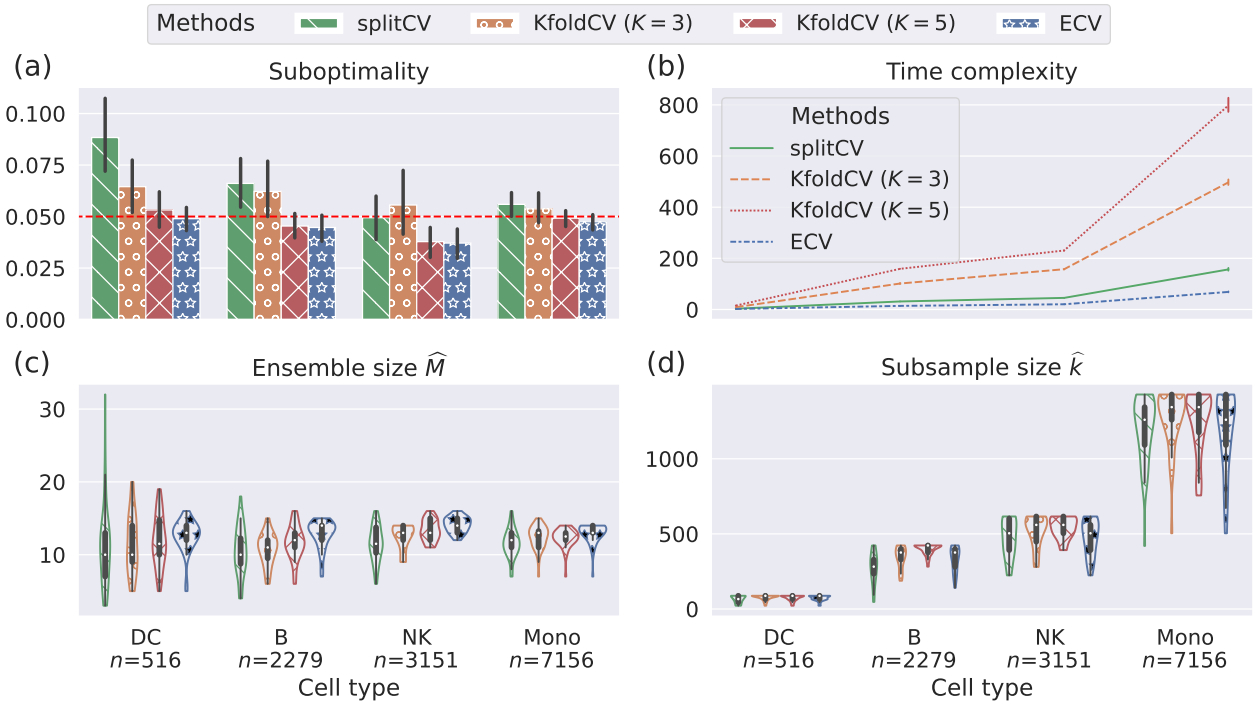

Figure 1: Overview of different cross-validation methods for predicting 50 surface proteins based on 5,000 genes in the single-cell sequencing multiomic datasets from Hao et al. (2021) using random forests. The proposed ECV method estimates the prediction risks based on 20 trees. Each panel shows cell types DC, B, NK, and Mono. (a) The out-of-sample suboptimality compared to the optimal random forest with 50 trees. The CV predictors are tuned to be δ-optimal in terms of normalized mean squared errors. The error bars show the standard deviations over 50 proteins, and the red dashed line indicates the optimality threshold $\delta = 0.05$. (b) The time consumption in seconds. (c) The cross-validated ensemble size $\widehat{M}$. (d) The cross-validated subsample size $\widehat{k}$.

to sample splitting, particularly in high-dimensional scenarios (Wang et al., 2018; Rad and Maleki, 2020; Patil et al., 2023). Furthermore, these generic CV methods must evaluate every possible ensemble and subsample size within an arbitrarily chosen search space, which often requires exploring larger ensemble sizes, thus demanding more computational resources. Yet, even with these, certifying any optimality outside this predefined search is not generally possible. These drawbacks highlight the need for more efficient tuning methods for ensemble learning.

This paper seeks to address both the theoretical and practical challenges associated with ensemble parameter tuning. To this end, we develop a CV method that can efficiently and consistently tune both the ensemble and subsample sizes. We focus primarily on randomized ensembles such as bagging and subagging (Breiman, 1996; Bühlmann and Yu, 2002) and rely on out-of-bag observations to estimate the conditional prediction risk. Our proposed method, termed ECV (Extrapolated Cross-Validation), enjoys several advantages over the previously mentioned approaches. We highlight two of them below: (1) *Statistical consistency*: ECV is versatile and model-agnostic and is applicable to general ensemble predictors. It provides uniform consistency in estimating the actual prediction risk of ensembles over all ensemble and subsample sizes under mild conditions. It also notably outperforms standard CV meth-

ods in finite samples, especially in high-dimensional settings. (2) *Computational efficiency*: ECV operates by estimating the risk of ensembles with small ensemble sizes ($M = 1, 2$) using out-of-bag observations and then extrapolates the risk estimates to arbitrary ensemble sizes. Unlike sample-split and K-fold CV, our method does not require fitting an ensemble for every ensemble size or explicitly estimating prediction risks for every ensemble size, and can serve as a valuable tool for assessing the fitness of ensemble predictors. Though our focus in the current paper is on ensemble and subsample size tuning, ECV is flexible for tuning other hyperparameters efficiently by risk extrapolation. As a result, ECV significantly reduces the computational burden, making it an efficient method for ensemble parameter tuning.

Before delving into the details of our method, we take the opportunity to demonstrate these key points through a real-world example. We apply ECV on four single-cell datasets and aim to select a δ-optimal random forest in Section 6, so that its prediction risk is no more than $\delta = 0.05$ away from the best random forest consisting of 50 trees. In Figure 1, we compare the performance of the sample-split CV (Patil et al., 2023), K-fold CV with $K = 3, 5$, and our method ECV applied to four datasets, each corresponding to a different cell type obtained from (Hao et al., 2021). Further details regarding this application can be found in Section 6. Both of the commonly used CV methods require estimating the risks with all possible choices of ensemble size M and subsample size k. To ensure a fair comparison, we use the same search space of (M, k) for all methods. Because sample splitting introduces additional randomness and the reduced sample size has significant finite sample effects, we observe from Figure 1(a) that sample-split CV does not control the out-of-sample error within the specified tolerance of $\delta = 0.05$ away from the best possible error. On the other hand, even though K-fold CV gives the valid error control as ECV, it costs extra computational time, which significantly increases as the sample size increases; see Figure 1(b). Overall, the distributions of tuned ensemble parameters $(\widehat{M}, \widehat{k})$ are similar between K-fold CV and ECV; see Figure 1(c)-(d). These results demonstrate the practical effectiveness of ECV in addressing the above drawbacks for ensemble parameter tuning.

We next provide an overview of the main results and delineate the structure for the remainder of the paper. In Section 2, we outline the setup of the tuning problem in the context of randomized ensemble learning. We also provide a comprehensive review of prior work on cross-validation and tuning and contrast our method to earlier work. In Section 3, we lay the foundation for our proposed method by constructing the theoretical ingredients essential to our main algorithm, in particular, the extrapolation risk estimation strategy. Through a non-asymptotic analysis, we further demonstrate the convergence rate and uniform consistency of the proposed risk estimator for general predictors and data distributions under a mild moment condition (see Theorem 3.3).

In Section 4, we introduce our main algorithm and discuss its theoretical properties and various practical considerations. We prove that the resulting tuned ensemble obtains the best possible ensemble risk over all ensemble and subsample sizes up to a specified tolerance of δ (see Theorem 4.1). In Section 5, we examine ECV's generality and effectiveness with various types of predictors. In Section 6, comparisons with sample-split and K-fold CV in the protein prediction problem highlight the statistical and computational benefits of ECV on low- and high-dimensional datasets, under a computational budget for the maximum ensemble size. In Section 7, we conclude the paper with a brief discussion that acknowledges some limitations

of the method and provides avenues for future work. The code to replicate all our experiments can be obtained from https://jaydu1.github.io/overparameterized-ensembling/ecv and the Python package implementing the ECV method can be found on the GitHub repository https://github.com/jaydu1/ensemble-cross-validation.

2 Randomized ensembles

We consider a supervised regression setup. Suppose $\mathcal{D}_n = \{(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_n, y_n)\}$ represents a dataset with independent and identically distributed random vectors from $\mathbb{R}^p \times \mathbb{R}$. We will not assume any specific data model, only that the second moment of the response is finite, i.e., $\mathbb{E}(y_1^2) < \infty$. 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.

Bagging (as in bootstrap-aggregating) traditionally refers to computing predictors multiple times based on bootstrapped data (Breiman, 1996), which can involve repeated observations. There is another version of bagging called subagging (as in subsample-aggregating) in Bühlmann and Yu (2002, Section 3.2) where we sample observations without replacement. Our method and analysis apply to both of these sampling strategies. Formally, these can be understood as simple random samples from a finite set, commonly used in survey sampling. Fix any $k \in [n]$, we define the indices $\{I_\ell\}_{\ell=1}^M$ to be M independent samples with replacement from $\mathcal{I}_k$ (denoted by $\{I_\ell\}_{\ell=1}^M \overset{\text{SRS}}{\sim} \mathcal{I}_k$). Here $\mathcal{I}_k$ is defined for bagging and subagging as

$$\mathcal{I}_k = \begin{cases} \{\{i_1, i_2, \ldots, i_k\} : 1 \le i_1 \le i_2 \le \cdots \le i_k \le n\}, & \text{(bagging)} \\ \{\{i_1, i_2, \ldots, i_k\} : 1 \le i_1 < i_2 < \cdots < i_k \le n\}. & \text{(subagging)} \end{cases} \tag{2.1}$$

For bagging, $\mathcal{I}_k$ represents the set of all possible independent draws from $[n]$ with replacement, and there are n^k many of them. For subagging, $\mathcal{I}_k$ represents the set of all k subset choices from $[n]$, and there are $n!/(k!(n-k)!)$ many of them. Throughout the paper, we mainly focus on subagging but the results apply equally well to bagging. For this reason, we do not distinguish different choices of $\mathcal{I}_k$. For any $I \in \mathcal{I}_k$, let $\mathcal{D}_I$ and the corresponding subsampled predictor be defined as $\mathcal{D}_I = \{(\boldsymbol{x}_j, y_j) : j \in I\}$ and $\widehat{f}(\boldsymbol{x}; \mathcal{D}_I) = \widehat{f}(\boldsymbol{x}; \{(\boldsymbol{x}_j, y_j) : j \in I\})$. Then the randomized ensemble using either bootstrap or subsampling is defined as follows:

$$\widetilde{f}_{M,k}(\boldsymbol{x}; \{\mathcal{D}_{I_\ell}\}_{\ell=1}^M) = \frac{1}{M} \sum_{\ell=1}^M \widehat{f}(\boldsymbol{x}; \mathcal{D}_{I_\ell}) \quad \text{with} \quad I_1, \ldots, I_M \overset{\text{SRS}}{\sim} \mathcal{I}_k. \tag{2.2}$$

In the context where we want to highlight the size of bootstrap/subsample, k, we write $I_{k,1}, \ldots, I_{k,M}$ instead of $I_1, \ldots, I_M$.

We are interested in the performance of our predictors (computed on $\mathcal{D}_n$) on future data from the same distribution P. We consider the behavior of the predictors conditional on $\mathcal{D}_n$ and $\{I_\ell\}_{\ell=1}^M$. More specifically, for a predictor $\widehat{f}$ fitted on $\mathcal{D}_n$ and its bagged predictor $\widetilde{f}_{M,k}$ fitted on $\{\mathcal{D}_{I_\ell}\}_{\ell=1}^M$, with $\{I_\ell\}_{\ell=1}^M \overset{\text{SRS}}{\sim} \mathcal{I}_k$, the data and subsample conditioned risks are defined as:

$$R(\widehat{f}; \mathcal{D}_n) = \int (y - \widehat{f}(\boldsymbol{x}; \mathcal{D}_n))^2 \, \mathrm{d}P(\boldsymbol{x}, y),$$

$$R(\widetilde{f}_{M,k};\, \mathcal{D}_n, \{I_\ell\}_{\ell=1}^M) = \int \left(y - \widetilde{f}_{M,k}(\boldsymbol{x}; \{\mathcal{D}_{I_\ell}\}_{\ell=1}^M)\right)^2 \mathrm{d}P(\boldsymbol{x}, y). \tag{2.3}$$

The two critical quantities for ensemble learning are the ensemble size M and the subsample size k. Different values of M trade-off model stability and computational burden. As M increases, the bagged predictors get more stabilized while requiring more time to fit them. On the other hand, the subsample size trades off the bias and variance of the bagged predictors. A smaller subsample size has a considerable bias but may reduce the variance. For example, in the context of subagging minimum norm least square predictors with $M = \infty$, a properly chosen subsample size k strictly less than n can have a higher variance reduction compared to the inflation of bias. This raises the question: how to efficiently choose both the ensemble size (M) and the subsample size (k) to minimize prediction risk (2.3) for general predictors. We address this question in the next section. Before that, we review some related work on cross-validation and situate our work in the context of other related work.

There is extensive literature on cross-validation (CV) approaches; see Appendix S1 for a detailed survey. In the context of bagging and subagging, Liu et al. (2019) study parameters selection for bagging in sparse regression based on the derived error bound. For subsample size tuning, a sample-split CV method is proposed in (Patil et al., 2022, 2023). To estimate the prediction risk without sample splitting, the other line of research uses the out-of-bag (OOB) observations (Breiman, 2001). For example, the algorithmic variance of ensemble regression functions at a fixed test point is studied in (Oshiro et al., 2012; Wager et al., 2014); in Lopes (2019); Lopes et al. (2020), the authors extrapolate the algorithmic variance and quantile of random forests for classification and regression problems, respectively. Their extrapolated estimators based on the heuristic scaling improve computation empirically, but theoretically, the statistical property is still unclear. Politis (2023) discuss scalable subbagging estimator when the subsample size k and the ensemble size M scale with the sample size n.

The current paper differs from the previously mentioned works in two significant aspects. First, the consistency of sample-split CV is shown in Patil et al. (2023) for a fixed ensemble size M for subagging, which suffers from the finite-sample effects because of sample splitting. Additionally, their results rely on stringent assumptions that require the asymptotic risk to satisfy certain analytic properties and do not provide any convergence rates. In contrast, our work establishes uniform consistency over both the ensemble size M and the subsample size k for bagging as well as subagging. More importantly, we also characterize the proposed estimators' convergence rate and require much milder assumptions on the risk, in the form of certain moment conditions. Second, Lopes (2019) and Lopes et al. (2020) rely on the convergence rate of variance to extrapolate the fluctuations of the estimates and require the ensemble size M to approach infinity. In contrast, ECV directly estimates the extrapolated risk (not just the scale of variances or quantiles) for an arbitrary range of ensemble sizes in a consistent manner. Additionally, while their papers only focus on tuning the ensemble size M, we also tune the subsample size k, which is crucial to minimizing the predictive risks, especially in high-dimensional scenarios, as alluded to in the introduction.

3 Method motivation

In this section, we derive preliminary results that serve as the foundation for our extrapolated cross-validation method in Section 4. Our approach relies on utilizing out-of-bag (OOB) observations to tune both the ensemble size M and the subsample size k. Let us now describe the main components behind our methodology.

3.1 Decomposition and risk estimation

To begin with, we will fix k and focus on tuning over $M \in \mathbb{N}$. Recall the conditional risk $R(\widetilde{f}_{M,k}; \mathcal{D}_n, \{I_\ell\}_{\ell=1}^M)$ associated with an M-bagged predictor $\widetilde{f}_{M,k}$, as defined in (2.3). The subsequent proposition demonstrates that $R(\widetilde{f}_{M,k}; \mathcal{D}_n, \{I_\ell\}_{\ell=1}^M)$ can be expressed as a linear combination of the conditional risks for $M = 1$ and $M = 2$.

Proposition 3.1 (Squared conditional risk decomposition)**.** *The conditional prediction risk defined in* (2.3) *for a bagged predictor* $\widetilde{f}_{M,k}$ *decomposes into*

$$R(\widetilde{f}_{M,k}; \mathcal{D}_n, \{I_\ell\}_{\ell=1}^M) = -\left(1 - \frac{2}{M}\right) a_{1,M} + 2\left(1 - \frac{1}{M}\right) a_{2,M}, \tag{3.1}$$

where $$a_{1,M} = \frac{1}{M}\sum_{\ell=1}^{M} R(\widetilde{f}_{1,k}; \mathcal{D}_n, \{I_\ell\}),\ a_{2,M} = \frac{1}{M(M-1)} \sum_{\substack{\ell,m \in [M] \\ \ell \neq m}} R(\widetilde{f}_{2,k}; \mathcal{D}_n, \{I_\ell, I_m\}).$$

The proof of Proposition 3.1 can be found in Appendix S2.1. The statement follows due to a special decomposition that governs the squared risk of the M-bagged predictor. The components $a_{1,M}$ and $a_{2,M}$ in the decomposition (3.1) are the averages of the 1-bagged and 2-bagged conditional risks, respectively. Conditional on $\mathcal{D}_n$, note that $a_{1,M}$ and $a_{2,M}$ are U-statistic based on i.i.d. elements $\{I_\ell\}_{\ell=1}^M$ sampled from $\mathcal{I}_k$.

Towards motivating our ECV method, let us assume that there exist constants $\mathfrak{R}_{1,k}$ and $\mathfrak{R}_{2,k}$ such that as $n \to \infty$, both $a_{1,M} - \mathfrak{R}_{1,k}$ and $a_{2,M} - \mathfrak{R}_{2,k}$ converge almost surely to 0. Then Proposition 3.1 implies that the conditional prediction risk of $\widetilde{f}_{M,k}$ can be approximated asymptotically by $-(1-2/M)\mathfrak{R}_{1,k} + 2(1-1/M)\mathfrak{R}_{2,k}$. This serves as the basis for the concept of ECV, where consistent estimation of $\mathfrak{R}_{1,k}$ and $\mathfrak{R}_{2,k}$ allows for a consistent estimator of the risk of $\widetilde{f}_{M,k}$ for every $M \in \mathbb{N}$, hence justifying the name "extrapolated" cross-validation.

To consistently estimate the basic components $\mathfrak{R}_{1,k}$ and $\mathfrak{R}_{2,k}$ in (3.1), we first leverage the OOB risk estimator for an arbitrary predictor $\widehat{f}$ fitted on $\mathcal{D}_I$ based on the OOB test dataset $\mathcal{D}_{I^c} = \mathcal{D}_n \setminus \mathcal{D}_I$. One can simply consider the average of the squared loss (MEAN) on OOB observations in $\mathcal{D}_{I^c}$ as a risk estimator. This choice, however, is not suitable for heavy-tailed data and hence we also consider a median-of-means (MOM) estimator:

$$\widehat{R}(\widehat{f}, \mathcal{D}_{I^c}) = \begin{cases} \frac{1}{|I^c|}\sum_{i \in I^c}(y_i - \widehat{f}(\boldsymbol{x}_i))^2, & \text{if EST} = \text{MEAN}, \\ \text{median}\left\{\frac{1}{|I^{(b)}|}\sum_{i \in I^{(b)}}(y_i - \widehat{f}(\boldsymbol{x}_i))^2,\ b \in [B]\right\}, & \text{if EST} = \text{MOM}, \end{cases} \tag{3.2}$$

with $B = \lceil 8\log(1/\eta)\rceil$ and $I^{(1)}, \ldots, I^{(B)}$ being $B = \lceil 8\log(1/\eta)\rceil$ random splits of I^c for some $\eta > 0$. The median-of-means estimator was developed for heavy-tailed mean estimation and is commonly used in robust statistics (Lugosi and Mendelson, 2019).

We will provide a condition to certify the pointwise consistency of risk estimates $\widehat{R}$ under certain assumptions on the data distribution. Towards that end, for any non-negative loss function $\mathcal{L}$ and a given test observation $(\boldsymbol{x}_0, y_0)$ from P, define the conditional ψ_1-Orlicz norm of $\mathcal{L}(y_0, \widehat{f}(\boldsymbol{x}_0))$ given $\mathcal{D}_I$ as $\|\mathcal{L}(y_0, \widehat{f}(\boldsymbol{x}_0))\|_{\psi_1|\mathcal{D}_I} = \inf\big\{C > 0 : \mathbb{E}\big[\exp\big(\mathcal{L}(y_0, \widehat{f}(\boldsymbol{x}_0))C^{-1}\big) \mid \mathcal{D}_I\big] \leq 2\big\}$. Similarly, for $r \geq 1$, define the conditional L_r-norm as $\|\mathcal{L}(y_0, \widehat{f}(\boldsymbol{x}_0))\|_{L_r|\mathcal{D}_I} = \big(\mathbb{E}[\mathcal{L}(y_0, \widehat{f}(\boldsymbol{x}_0))^r \mid \mathcal{D}_I]\big)^{1/r}$. See Vershynin (2018, Chapter 2) for more details. The following proposition provides the condition for consistency.

Proposition 3.2 (Consistent risk estimators)**.** *Let $\widehat{f}(\cdot; \mathcal{D}_I)$ be any predictor trained on $\mathcal{D}_I \subset \mathcal{D}_n$, and* EST = MEAN *or* EST = MOM *with $\eta = n^{-A}$, where $A \in (0, \infty)$ is a fixed constant. Define $\widehat{\sigma}_I = \|(y_0 - \widehat{f}(\boldsymbol{x}_0; \mathcal{D}_I))^2\|_{\psi_1|\mathcal{D}_I}$. If $\widehat{\sigma}_I / \sqrt{|I^c|/\log n} \to 0$ in probability, then $|\widehat{R}(\widehat{f}, \mathcal{D}_{I^c}) - R(\widehat{f}; \mathcal{D}_I)| \to 0$ in probability. The conclusion remains true for* EST = MOM *even when the conditional ψ_1-Orlicz norm in the definition of $\widehat{\sigma}_I$ is replaced with $\|\cdot\|_{L_2|\mathcal{D}_I}$.*

The proof of Proposition 3.2 can be found in Appendix S2.2. It follows from the results in Patil et al. (2023, Lemma 2.9, Lemma 2.10), which are used to demonstrate the strong consistency of sample-split CV in (Patil et al., 2022, 2023) for subsample tuning. However, our analysis will utilize the finite-sample tail bounds that underlie Proposition 3.2 to obtain convergence rates.

Recall from (2.2) that $\widetilde{f}_{1,k}$ is computed from one dataset $\mathcal{D}_{I_1}$ and $\widetilde{f}_{2,k}$ is computed from two datasets $\mathcal{D}_{I_1}$ and $\mathcal{D}_{I_2}$ or equivalently that $\widetilde{f}_{2,k}$ is computed on $\mathcal{D}_{I_1 \cup I_2}$. Hence, Proposition 3.2 can be applied to $\widetilde{f}_{1,k}$ with $I = I_1$ and to $\widetilde{f}_{2,k}$ with $I = I_1 \cup I_2$ to consistently estimate the conditional prediction risks of $\widetilde{f}_{1,k}$ and $\widetilde{f}_{2,k}$. To obtain consistency, we need to ensure that the assumption on $\widehat{\sigma}_I$ becomes reasonable if $|I^c|/\log n \to \infty$ as $n \to \infty$. For subagging predictors $\widetilde{f}_{1,k}$ and $\widetilde{f}_{2,k}$, we have $|I_1^c| = n(1 - k/n)$ and by Lemma S5.2, $|(I_1 \cup I_2)^c| \approx n(1 - k/n)^2$, asymptotically. On the other hand, for bagging predictors $\widetilde{f}_{1,k}$ and $\widetilde{f}_{2,k}$, we have $|I_1^c| \approx n\exp(-k/n)$ and $|(I_1 \cup I_2)^c| \approx n\exp(-2k/n)$, asymptotically. Hence, collectively, assuming $k \leq n(1 - 1/\log n)$ implies that $|I^c| \gtrsim n/\log^2 n$, asymptotically.

Under the assumption $k \leq n(1 - 1/\log n)$, the risks of $\widetilde{f}_{1,k}$ and $\widetilde{f}_{2,k}$ computed on $\mathcal{D}_{I_1}$ and $\mathcal{D}_{I_1 \cup I_2}$, respectively, can be estimated consistently. Further, if we assume that the limiting conditional risks ($\mathfrak{R}_{1,k}$ and $\mathfrak{R}_{2,k}$) of $\widetilde{f}_{1,k}$ and $\widetilde{f}_{2,k}$ do not depend on the particular subsets I_1, I_2, then $\widehat{R}(\widetilde{f}_{1,k}, \mathcal{D}_{I_1^c})$ and $\widehat{R}(\widetilde{f}_{2,k}, \mathcal{D}_{(I_1 \cup I_2)^c})$ consistently estimate $\mathfrak{R}_{1,k}$ and $\mathfrak{R}_{2,k}$, as $n \to \infty$. Note, however, that because the limiting conditional risks do not depend on specific subsets I_1, I_2, we can reduce the variance in our estimates of $\mathfrak{R}_{1,k}$ and $\mathfrak{R}_{2,k}$ by averaging the estimated risks over several subsets I_ℓ's. This observation suggests the out-of-bag risk estimates for $M = 1, 2$ as

$$\widehat{R}^{\mathrm{ECV}}_{M,k} = \begin{cases} \frac{1}{M_0} \sum_{\ell=1}^{M_0} \widehat{R}(\widetilde{f}_{1,k}(\cdot; \mathcal{D}_n, \{I_\ell\}), \mathcal{D}_{I_\ell^c}), & M = 1, \\ \frac{1}{M_0(M_0-1)} \sum\limits_{\substack{\ell,m \in [M_0] \\ \ell \neq m}} \widehat{R}(\widetilde{f}_{2,k}(\cdot; \mathcal{D}_n, \{I_\ell, I_m\}), \mathcal{D}_{(I_\ell \cup I_m)^c}), & M = 2, \end{cases} \tag{3.3}$$

where $I_1, \ldots, I_{M_0}$ are i.i.d. samples from $\mathcal{I}_k$ and $M_0 \geq 2$ is a pre-specified natural number. Increasing M_0 improves estimates but also increases computation.

As hinted above, the risk decomposition (3.1), along with the component risk estimation (3.3), suggests a natural estimator for the M-bagged risk:

$$\widehat{R}^{\mathrm{ECV}}_{M,k} = -\left(1 - \frac{2}{M}\right)\widehat{R}^{\mathrm{ECV}}_{1,k} + 2\left(1 - \frac{1}{M}\right)\widehat{R}^{\mathrm{ECV}}_{2,k}, \quad M > 2. \tag{3.4}$$

We call this "extrapolated" risk estimation according to (3.1) because the M-bagged risk is *extrapolated* from the 1- and 2-bagged risks. If the prediction risk of $M = 1, 2$ can be consistently estimated, the extrapolated estimates (3.4) are also pointwise consistent over $M \in \mathbb{N}$ for k fixed.

3.2 Uniform risk consistency

The next step is then to tune both M and k. To tune k, we define $\mathcal{K}_n \subset [n]$ to be a grid of subsample sizes. In practice, we would like $\mathcal{K}_n$ to cover the full range of n asymptotically (in the sense that $\mathcal{K}_n/n \approx [0, 1]$), and one simple choice is to set $\mathcal{K}_n = \{0, k_0, 2k_0, \ldots, \lfloor n(1 - (\log n)^{-1})/k_0 \rfloor k_0\}$ where the minimum subsample size is $k_0 = \lfloor n^\nu \rfloor$ for some $\nu \in (0, 1)$. Here we adopt the convention that when $k = 0$, the ensemble predictor reduces to the *null predictor* that always outputs zero. To facilitate our theoretical results for general predictors, we make the following two assumptions. The results are stated asymptotically as n tends to infinity, where we view both k and p as sequences $\{k_n\}$ and $\{p_n\}$ indexed by n, and assume k diverges with the sample size n (except when $k \equiv 0$), but the feature dimension p may or may not diverge.

Assumption 1 (Variance proxy)**.** For $k \in \mathcal{K}_n$, assume for all $\{I_1, I_2\} \overset{\mathrm{SRS}}{\sim} \mathcal{I}_k$, as $n \to \infty$,

$$\frac{\log n}{\sqrt{n}}\widehat{\sigma}_{I_1} \to 0, \quad \text{and} \quad \frac{(\log n)^{3/2}}{\sqrt{n}}\widehat{\sigma}_{I_1 \cup I_2} \to 0,$$

in probability, where the variance proxies $\widehat{\sigma}_{I_1}$ and $\widehat{\sigma}_{I_1 \cup I_2}$ are defined in Proposition 3.2.

Assumption 2 (Convergence of asymptotic risks)**.** For $k \in \mathcal{K}_n$, assume for all $\{I_1, I_2\} \overset{\mathrm{SRS}}{\sim} \mathcal{I}_k$ and for $M = 1, 2$, there exist constants $\epsilon \in (0, 1)$, $C_0 > 0$, $\eta_0 \geq 1$, $\gamma_{M,n} = o(n^{-\epsilon})$, and $\mathfrak{R}_{M,k} \geq 0$, such that the following holds:

$$\limsup_{n \to \infty} \sup_{\eta \geq \eta_0} \eta^{1/\epsilon} \mathbb{P}(\gamma_{M,n}^{-1} |R_{M,k}(\widetilde{f}_{M,k};\, \mathcal{D}_n, \{I_\ell\}_{\ell=1}^M) - \mathfrak{R}_{M,k}| \leq \eta) \leq C_0. \tag{3.5}$$

Assumption 1 is used to show consistent risk estimation with $M = 1, 2$ in Appendix S2.2. Assumption 2 formalizes the assumption that the limiting values of the conditional risks $R_{M,k}(\widetilde{f}_{M,k}; \mathcal{D}_n, \{I_\ell\}_{\ell=1}^M)$ do not depend on $\{I_\ell\}_{\ell=1}^M$. This assumption also requires certain rate and tail assumptions, which are used to provide the rate of consistency of our ECV procedure. In Assumption 2, $\gamma_{M,n}$ for $M = 1, 2$ represent the lower bounds of the rates of convergence over $k \in \mathcal{K}_n$, which are typically on a scale of $n^{-\alpha}$ for some constant $\alpha > 0$. Condition (3.5) is also known as the weak moment norm condition (Rio, 2017; Guo and Peterson, 2019), which ensures that the expected differences between the risks and the limits also converge to zero in certain rates. Under classical linear models with fixed-X design

and Gaussian noises, the risk of linear predictor concentrates around its mean and satisfies (3.5); see, e.g., Bellec (2018, Lemma 3.1). Another sufficient condition for (3.5) is the strong moment condition that $\mathbb{E}(\gamma_{M,n}^{-1/\epsilon}|R_{M,k}(\widetilde{f}_{M,k}; \mathcal{D}_n, \{I_\ell\}_{\ell=1}^M) - \mathfrak{R}_{M,k}|^{1/\epsilon})$ is bounded.

From now on, we shall write $R_{M,k} = R(\widetilde{f}_{M,k}; \mathcal{D}_n, \{I_\ell\}_{\ell=1}^M)$ to indicate the dependency only on M and k and to simplify the notations. In Appendix S4.3, we present an example of ridge regression where the convergence is under the proportional asymptotics (i.e., both the *data aspect ratio* p/n and the *subsample aspect ratio* p/k converge to fixed constants), and Assumptions 1-2 are satisfied. The following theorem guarantees uniform consistency over both M and k.

Theorem 3.3 (Uniform consistency of risk extrapolation)**.** *Suppose Assumptions 1 and 2 hold for all $k \in \mathcal{K}_n$, then ECV estimates defined in (3.4) satisfy that*

$$\sup_{M\in\mathbb{N},k\in\mathcal{K}_n} \left|\widehat{R}_{M,k}^{\mathrm{ECV}} - R_{M,k}\right| = \mathcal{O}_p(\zeta_n),$$

where $\zeta_n = \widehat{\sigma}_n \log n/n^{1/2} + n^\epsilon(\gamma_{1,n} + \gamma_{2,n})$, and $\widehat{\sigma}_n = \max_{m,\ell\in[M_0],k\in\mathcal{K}_n} \widehat{\sigma}_{I_{k,\ell}\cup I_{k,m}}$.

The result in Theorem 3.3 is of paramount importance to establish the convergence rate of the CV-tuned estimator returned by our algorithm. In words, the theorem says that the extrapolated error depends on three factors: the cross-validation error and the rates for $M = 1, 2$. While the convergence rates of asymptotic risks of $M = 1, 2$ usually depend on the chosen predictor, the non-asymptotic analysis in Theorem 3.3 allows us to derive convergence rates even for general predictors. This is particularly advantageous when compared to the consistency results established in (Patil et al., 2023) for subagged ridge predictors.

The proof Theorem 3.3 is rather involved and can be found in Appendix S3. For the convenience of the readers, we provide a schematic of the whole proof in Figure S1. We will now explain the key ideas involved in the proof. The proof strategy relies on deriving concentration results of varied random quantities to their limits in a specific order. First, we establish the uniform consistency of the risk estimates $\widehat{R}_{M,k}^{\mathrm{ECV}}$ over $k \in \mathcal{K}_n$ to the risks $R_{M,k}$ for $M = 1, 2$ (see Proposition S3.1). Building upon this result and the risk decomposition presented in Proposition 3.1, we then derive the uniform consistency of the risk estimates $\widehat{R}_{M,k}^{\mathrm{ECV}}$ over $(M, k) \in \mathbb{N} \times \mathcal{K}_n$ to the deterministic limits $\mathfrak{R}_{M,k}$ (Proposition S3.2). On the other hand, to establish the concentration for subsample conditional risks over $M \in \mathbb{N}$ and $k \in \mathcal{K}_n$, we first establish the concentration for the expected conditional risk $\mathbb{E}[R_{M,k} \mid \mathcal{D}_n]$ over $k \in \mathcal{K}_n$ (Lemma S3.3). We then apply the reverse martingale concentration bound (Lemma S3.4).

4 Main proposal: Extrapolated cross-validation

Based on the previous discussion in Section 3, we present the proposed cross-validation algorithm for tuning the ensemble parameters without sample splitting in Algorithm 1. The procedure requires a dataset $\mathcal{D}_n$ of n observations, a base prediction procedure $\widehat{f}$, a natural number M_0 for risk estimation, and some other parameters. It first constructs the grid of subsample sizes $\mathcal{K}_n$ and fits only M_0 base predictors accordingly. Then, the prediction risk

for M-bagged predictors can be estimated based on the OOB observations through (3.3) and (3.4). Observe that the optimal risk of (3.4) for any k is obtained at $\widehat{R}^{\mathrm{ECV}}_{\infty,k} = 2\widehat{R}^{\mathrm{ECV}}_{2,k} - \widehat{R}^{\mathrm{ECV}}_{1,k}$ when $M = \infty$. Thus, to tune k, it suffices to perform a grid search to minimize $\widehat{R}^{\mathrm{ECV}}_{\infty,k}$ over $k \in \mathcal{K}_n$, because the optimal ensemble size happens to be infinity from the previous results (Lopes, 2019; Patil et al., 2023). However, calculating it is prohibitive in practice. Thus, we pick the smallest $\widehat{M}$ such that $\widehat{R}^{\mathrm{ECV}}_{\widehat{M},\widehat{k}}$ is close to $\widehat{R}^{\mathrm{ECV}}_{\infty,\widehat{k}}$ within δ error, where δ is the suboptimality parameter. Finally, Algorithm 1 returns a $\widehat{M}$-bagged predictor using subsample size $\widehat{k}$. Note that Algorithm 1 naturally applies to other randomized ensemble methods, such as random forests, when fixing other hyperparameters.

Algorithm 1 Tuning of ensemble and subsample sizes without sample splitting

Input: a dataset $\mathcal{D}_n = \{(\boldsymbol{x}_i, y_i) \in \mathbb{R}^p \times \mathbb{R} : 1 \le i \le n\}$, a base prediction procedure $\widehat{f}$, a real number $\nu \in (0,1)$ (subsample size unit parameter), a ensemble size $M_0 \ge 2$ for risk estimation, a centering procedure $\mathrm{EST} \in \{\mathrm{MEAN}, \mathrm{MOM}\}$, a real number A used to compute η when $\mathrm{EST} = \mathrm{MOM}$, and optimality tolerance parameter δ.

1: Construct a grid $\mathcal{K}_n = \{0, k_0, 2k_0, \ldots, \lfloor n(1 - 1/\log n)/k_0 \rfloor k_0\}$ where $k_0 = \lfloor n^\nu \rfloor$.

2: Build ensembles $\widetilde{f}_{M_0,k}(\cdot) = \widetilde{f}_{M_0}(\cdot; \{\mathcal{D}_{I_{k,\ell}}\}_{\ell=1}^{M_0})$ with M_0 base predictors, where $I_{k,1}, \ldots, I_{k,M_0} \overset{\mathrm{SRS}}{\sim} \mathcal{I}_k$ for each $k \in \mathcal{K}_n$.

3: Estimate the conditional prediction risk on OOB observations of $\widetilde{f}_{M_0,k}$ with $\widehat{R}^{\mathrm{ECV}}_{M,k}$ defined in (3.3) for $k \in \mathcal{K}_n$ and $M = 1, 2$.

4: Extrapolate the risk estimations $\widehat{R}^{\mathrm{ECV}}_{M,k}$ using (3.4).

5: Select a subsample size $\widehat{k} \in \mathcal{K}_n$ that minimizes the extrapolated estimates using

$$\widehat{k} \in \operatorname*{argmin}_{k \in \mathcal{K}_n} 2\widehat{R}^{\mathrm{ECV}}_{2,k} - \widehat{R}^{\mathrm{ECV}}_{1,k}.$$

6: Select an ensemble size $\widehat{M} \in \mathbb{N}$ for the δ-optimal risk with a plug-in estimator:

$$\widehat{M} = \left\lceil \frac{2}{\max\{\delta, n^{-1/2}\}} (\widehat{R}^{\mathrm{ECV}}_{1,\widehat{k}} - \widehat{R}^{\mathrm{ECV}}_{2,\widehat{k}}) \right\rceil.$$

7: If $\widehat{M} > M_0$, fit a $\widehat{M}$-bagged predictor $\widetilde{f}_{\widehat{M},\widehat{k}} = \widetilde{f}_{\widehat{M},\widehat{k}}(\cdot; \{\mathcal{D}_{I_{\widehat{k},\ell}}\}_{\ell=1}^{\widehat{M}})$.

Output: Return the ECV-tuned predictor $\widetilde{f}_{\widehat{M},\widehat{k}}$, and the risk estimators $\widehat{R}^{\mathrm{ECV}}_{M,k}$ for all M, k.

As a byproduct, Algorithm 1 also gives the ECV risk estimates $\widehat{R}^{\mathrm{ECV}}_{M,k}$ for all $M \in \mathbb{N}$ and $k \in \mathcal{K}_n$. This risk profile in (M, k) is helpful for users to investigate whether the given base predictor $\widehat{f}$ well fits the dataset $\mathcal{D}_n$ or not. For instance, one can tune the ensemble predictors under a computational budget on the maximum ensemble size $M_{\max}$. This gives rise to practical considerations presented later in Section 4.2.

4.1 Theoretical guarantees

By combining the ingredients in Section 3, our main theorem states that Algorithm 1 yields an ensemble predictor whose risk is at most δ away from the best ensemble predictor. Further, it provides an estimator of the risk of the selected ensemble predictor.

Theorem 4.1 (Optimality of OOB estimate and ECV-tuned risk)**.** *Under the assumed conditions in Theorem 3.3, for any $\delta > 0$, the OOB estimate and the ECV-tuned risk output by Algorithm 1 satisfy the following properties respectively:*

$$\left|\widehat{R}^{\mathrm{ECV}}_{\widehat{M},\widehat{k}} - R_{\widehat{M},\widehat{k}}\right| = \mathcal{O}_p(\zeta_n), \qquad \left|R_{\widehat{M},\widehat{k}} - \inf_{M\in\mathbb{N},k\in\mathcal{K}_n} R_{M,k}\right| = \delta + \mathcal{O}_p(\zeta_n), \tag{4.1}$$

where ζ_n is the quantity defined in Theorem 3.3.

Theorem 4.1 implies that the OOB estimate is close to the true risk because of the uniform consistency from Theorem 3.3. Furthermore, the ECV-tuned ensemble parameters $\widehat{M}$ and $\widehat{k}$ produce a bagged predictor with a risk δ-close to the optimal predictor. The optimality is model-agnostic because it does not directly depend on the feature and response models. On the other hand, in real-world applications, the infinite ensemble may not be of interest because of the computational cost. Because of the uniform consistency established in Theorem 3.3, one can naturally extend Theorem 4.1 to tuning with restriction on maximum ensemble sizes. In Appendix S4.3, we also provide a concrete example of the application of Theorem 4.1 to ridge predictors, which verifies Assumptions 1 and 2 under mild moment assumptions.

Remark 4.2 (From additive to multiplicative optimality)**.** Theorem 4.1 provides a guarantee of additive optimality for tuned predictor returned by Algorithm 1, while it is also useful to consider the multiplicative optimality. Towards that end, with the choice of $\widehat{k}$ in Step 5 of Algorithm 1 and change the choice of $\widehat{M}$ in Step 6 to

$$\widehat{M} = \left\lceil \frac{2}{\max\{\delta, n^{-1/2}\}} \frac{\widehat{R}^{\mathrm{ECV}}_{1,\widehat{k}} - \widehat{R}^{\mathrm{ECV}}_{2,\widehat{k}}}{2\widehat{R}^{\mathrm{ECV}}_{2,\widehat{k}} - \widehat{R}^{\mathrm{ECV}}_{1,\widehat{k}}} \right\rceil,$$

then, under the assumption that the irreducible risk $\int (y - \mathbb{E}(y \mid \boldsymbol{x}))^2 \, \mathrm{d}P(\boldsymbol{x}, y)$ is strictly positive, Proposition S4.1 guarantees

$$R_{\widehat{M},\widehat{k}} = (1+\delta) \inf_{M\in\mathbb{N},k\in\mathcal{K}_n} R_{M,k}(1 + \mathcal{O}_p(\zeta_n)). \tag{4.2}$$

Compared to (4.1), the optimality upper bound on the right-hand side of (4.2) depends on the scale of the optimal prediction risk.

4.2 Computational considerations

Algorithm 1 estimates the ensemble parameters $\widehat{k}$ and $\widehat{M}$ to derive a δ-optimal bagged predictor. Here, we compare the computational complexity of Algorithm 1 with other common

CV methods. For each $k \in \mathcal{K}_n$, suppose the computational complexity of fitting one base predictor on k subsampled observations and obtain their predicted values on $n-k$ OOB observations is $\mathcal{O}(C_n)$ (ignoring k). Then, the computational complexity of estimating $\widehat{M}$ and $\widehat{k}$ for all three validation methods are given below.

- ECV: For each $k \in \mathcal{K}_n$, we need to fit M_0 base predictors. Then we estimate $R_{1,k}$ and $R_{2,k}$ in $\mathcal{O}(M_0(n-k))$ and $\mathcal{O}(M_0^2(n-2k+i_0))$, respectively, where $i_0 = k^2/n$ is the intersect observations between two indices of a simple random sample. The computational complexity of risk extrapolation is negligible compared to the above time consumption. All in all, it takes $\mathcal{O}(C_n(|\mathcal{K}_n|M_0 + M_{\max}))$ to obtain tuned bagging parameter by ECV.

- sample-split CV: Suppose the ratio of training data is $\alpha \in (0,1)$. Similar to ECV, each base predictor is fitted and evaluated on $\lceil n\alpha \rceil$ observations and we need to fit $M_{\max}$ base predictors. We then compute the moving average of the predicted values for M varying from 1 to $M_{\max}$, which gives the predicted values of the M-bagged predictors, which takes $\mathcal{O}(M_{\max})$ operations. We note that one can alternatively fit one bagged predictor for each k and each M; however, this will cause much more time consumption compared to the simple matrix computation operations we used above. All in all, it takes $\mathcal{O}(|\mathcal{K}_n|M_{\max}(C_{n\alpha} + n))$ to obtain the tuned parameter.

- K-fold CV: We follow the same strategy for fitting base predictors so that K-fold CV has roughly K times of complexity as sample-split CV. Specifically, it takes $\mathcal{O}(K|\mathcal{K}_n|M_{\max}(C_{n/K} + n))$ to obtain the tuned parameter.

In general, we expect C_n to grow much faster than n, because fitting one base predictor may involve matrix multiplication operation, which takes $\mathcal{O}(n^2)$. Therefore, the computational complexity of the three methods has the ordering: ECV $\leq$ sample-split CV $\leq$ K-fold CV, provided that M_0 is much smaller than $M_{\max}$.

Besides the computational efficiency gained by ECV, we also discuss some considerations when the proposed method is used in practice.

1. Maximum ensemble size: Algorithm 1 determines $\widehat{k}$ and $\widehat{M}$ by minimizing the estimated risk (3.4) with the infinite ensemble. However, it may still be computationally infeasible if $\widehat{M}$ is too large. In such cases, based on the extrapolated risk estimation in Algorithm 1, we can also derive the δ-optimal bagged predictor whose ensemble size is no more than a pre-specified number $M_{\max}$, which we call the restricted oracle. That is, we choose subsample and ensemble size to restrict the computational cost: $\widehat{k} \in \operatorname{argmin}_{k \in \mathcal{K}_n} \widehat{R}^{\mathrm{ECV}}_{M_{\max},k}$ and $\widehat{M} = \left\lceil 2(\delta + \widehat{R}^{\mathrm{ECV}}_{M_{\max},\widehat{k}} - \widehat{R}^{\mathrm{ECV}}_{\infty,\widehat{k}})^{-1}(\widehat{R}^{\mathrm{ECV}}_{1,\widehat{k}} - \widehat{R}^{\mathrm{ECV}}_{2,\widehat{k}}) \right\rceil$. On the other hand, it also controls the suboptimality to the oracle:

$$\underbrace{R_{\widehat{M},\widehat{k}} - \min_{k \in \mathcal{K}_n} R_{\infty,k}}_{\text{suboptimality to the oracle}} = \underbrace{R_{\widehat{M},\widehat{k}} - \min_{k \in \mathcal{K}_n} R_{M_{\max},k}}_{\text{suboptimality to the restricted oracle}} + \underbrace{\min_{k \in \mathcal{K}_n} R_{M_{\max},k} - \min_{k \in \mathcal{K}_n} R_{\infty,k}}_{\text{unavoidable budget error}}.$$

   When $\delta = 0$, the suboptimality to the restricted oracle vanishes, and the tuned ensemble simply tracks the optimal $M_{\max}$-ensemble.

2. To bag or not to bag: The benefit of ensemble learning may be slight in some cases. For instance, when the number of samples is much larger than the feature dimensions and the signal-noise ratio is large, ensemble learning can only provide minor improvements over the non-ensemble predictor. Suppose that $\widehat{R}_0^{\mathrm{ECV}}$ is the estimated risk of the null predictor, $\min_{k\in\mathcal{K}_n} \widehat{R}_{1,k}^{\mathrm{ECV}}$ is the optimal estimated risk due to subsampling when $M = 1$, and $\min_{k\in\mathcal{K}_n} \widehat{R}_{M_{\max},k}^{\mathrm{ECV}}$ is the optimal ECV estimate with maximum ensemble size $M_{\max}$. Let $\zeta > 0$ be a user-specified improvement factor that encodes the desired excess risk improvement in a multiplicative sense. Then we can decide to bag if either the null risk is smaller in the sense that $\widehat{R}_0^{\mathrm{ECV}} < \min_{k\in\mathcal{K}_n} \widehat{R}_{1,k}^{\mathrm{ECV}}$, or the improvement due to ensemble exceeds ζ times the improvement due to subsampling:

$$\underbrace{\min_{k\in\mathcal{K}_n} \widehat{R}_{1,k}^{\mathrm{ECV}} - \min_{k\in\mathcal{K}_n} \widehat{R}_{M_{\max},k}^{\mathrm{ECV}}}_{\text{improvement due to ensemble}} > \underbrace{\zeta}_{\text{improvement factor}} \times \underbrace{\widehat{R}_0^{\mathrm{ECV}} - \min_{k\in\mathcal{K}_n} \widehat{R}_{1,k}^{\mathrm{ECV}}}_{\text{improvement due to subsampling}} . \tag{4.3}$$

3. Absolute versus normalized tolerances: The choice of the tolerance threshold δ is for controlling the absolute suboptimality, but the scale of the prediction risk may be different for different predictors and datasets. One can normalize the estimated risk by the null predictor's estimated risk to make the tolerance threshold comparable across different predictors and datasets, or tune based on the multiplicative guarantee (4.2).

5 Numerical illustrations

In this section, we evaluate Algorithm 1 on synthetic data. In Section 5.1, we inspect whether the extrapolated risk estimates $\widehat{R}_{M,k}^{\mathrm{ECV}}$ serve as reasonable proxies for the actual out-of-sample prediction errors for various base predictors on uncorrelated features. In Section 5.2, we further evaluate the risk minimization performance with tuned ensemble parameters $(\widehat{M}, \widehat{k})$. Finally, in Section 5.3, we consider tuning M for random forests on correlated features.

5.1 Validating extrapolated risk estimates

In this simulation, we examine whether our risk extrapolation strategy provides reasonable risk estimates for specific values of ensemble size M and subsample size k based on only the risk estimates for $M = 1$ and $M = 2$. We evaluate six base predictors (in fig. 2) on data models:

(M1) Linear: A linear model: $y = \boldsymbol{x}^\top\boldsymbol{\beta} + \epsilon$,

(M2) Quad: A polynomial regression model: $y = \boldsymbol{x}^\top\boldsymbol{\beta} + ((\boldsymbol{x}^\top\boldsymbol{\beta})^2 - \mathrm{tr}(\boldsymbol{\Sigma}_{\rho_{\mathrm{ar1}}})/p) + \epsilon$,

(M3) Tanh: A single-index regression model: $y = \tanh(\boldsymbol{x}^\top\boldsymbol{\beta}) + \epsilon$,

where the features and the coefficients are generated from $\boldsymbol{x} \sim \mathcal{N}_p(\boldsymbol{0}, \boldsymbol{\Sigma}_{\rho_{\mathrm{ar1}}})$ and $\boldsymbol{\beta}$ is the average of $\boldsymbol{\Sigma}_{\rho_{\mathrm{ar1}}}$'s eigenvectors associated with the top-5 eigenvalues. Here $\boldsymbol{\Sigma}_{\rho_{\mathrm{ar1}}} = (\rho_{\mathrm{ar1}}^{|i-j|})_{1\le i,j\le p}$ is the covariance matrix of an auto-regressive process of order 1 (AR(1)) with $\rho_{\mathrm{ar1}} = 0.5$. The additive noise is sampled from $\epsilon \sim \mathcal{N}(0, \sigma^2)$ with $\sigma = 0.5$. In this setup, (M1) has a

signal-noise ratio of 2.4. Here, the data aspect ratio $\phi = p/n$ varies from 0.1 (low-dimensional regime) to 10 (high-dimensional regime), and the subsample aspect ratio $\phi_s = p/k$ varies from 0.1 to 10 and from 10 to 100, respectively. The *null risk*, the risk of the null predictor that always outputs zero, can also be estimated at each ϕ. For ridgeless and lassoless predictors, we use rule (4.3) in Section 4.2 to exclude k with exploding risks more than 5 times the estimated null risk.

The ECV estimates and the corresponding prediction errors for different ensemble sizes M and subsample aspect ratios ϕ_s are then summarized in Appendix S6 for bagged and subagged predictors. As k decreases, the ensemble with subsample size k behaves more like the null predictor. As a result, the risk curves approach a particular value as the subsample aspect ratio p/k increases. From figs. S2 to S7, we observe a good match between the ECV estimates and the out-of-sample prediction errors. This suggests that Theorem 4.1 potentially applies to various types of predictors. Comparing the results of bagging and subagging, the risk estimates are very similar, especially in the overparameterized regime when $p > n$. Thus, we will only present the results using bagging for illustration purposes.

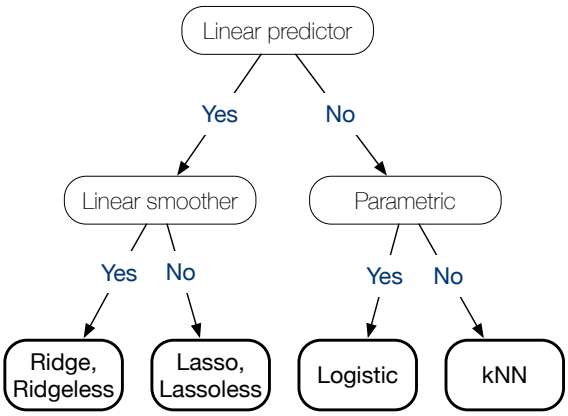

Figure 2: Predictors for regression tasks evaluated in Sections 5.1 and 5.2.

5.2 Tuning ensemble and subsample sizes

Next, we examine the performance of ECV on predictive risk minimization. More specifically, we apply Algorithm 1 using the rule (4.3) with $\zeta = 5$ to tune an ensemble that is close to the optimal $M_{\max}$-ensemble up to additive error δ, where the maximum ensemble size $M_{\max}$ is 50 and optimality threshold δ ranges from 0.01 to 1. With the same predictors and data used in Section 5.1, their out-of-sample mean squared errors are evaluated on the same test set.

As summarized in fig. 3, the thick dashed lines represent the non-ensemble prediction risk, and the thick solid lines represent the prediction risk of optimal 50-bagged predictors using a finer grid. Note that the former may be non-monotonic in the data aspect ratio ϕ, but the latter is increasing in ϕ. The finite-sample prediction errors of the ECV-tuned predictor are shown as thin lines. As we can see, when δ decreases, the prediction errors of ECV get closer to those of the optimal 50-bagged predictor. The slight discrepancy between the ECV-tuned risks with the least δ and the oracle risks comes from the fact that a coarser grid $\mathcal{K}_n$ is used for ECV tuning. Overall, the results suggest that the ECV-tuned ensemble parameters $(\widehat{M}, \widehat{k})$ give risks close to the oracle choices for various predictors within the desired optimality threshold δ in finite samples. Lastly, though ECV is proposed for regression tasks, the numerical results in Appendix S6.3 support its superiority over K-fold CV in imbalanced binary classification scenarios.

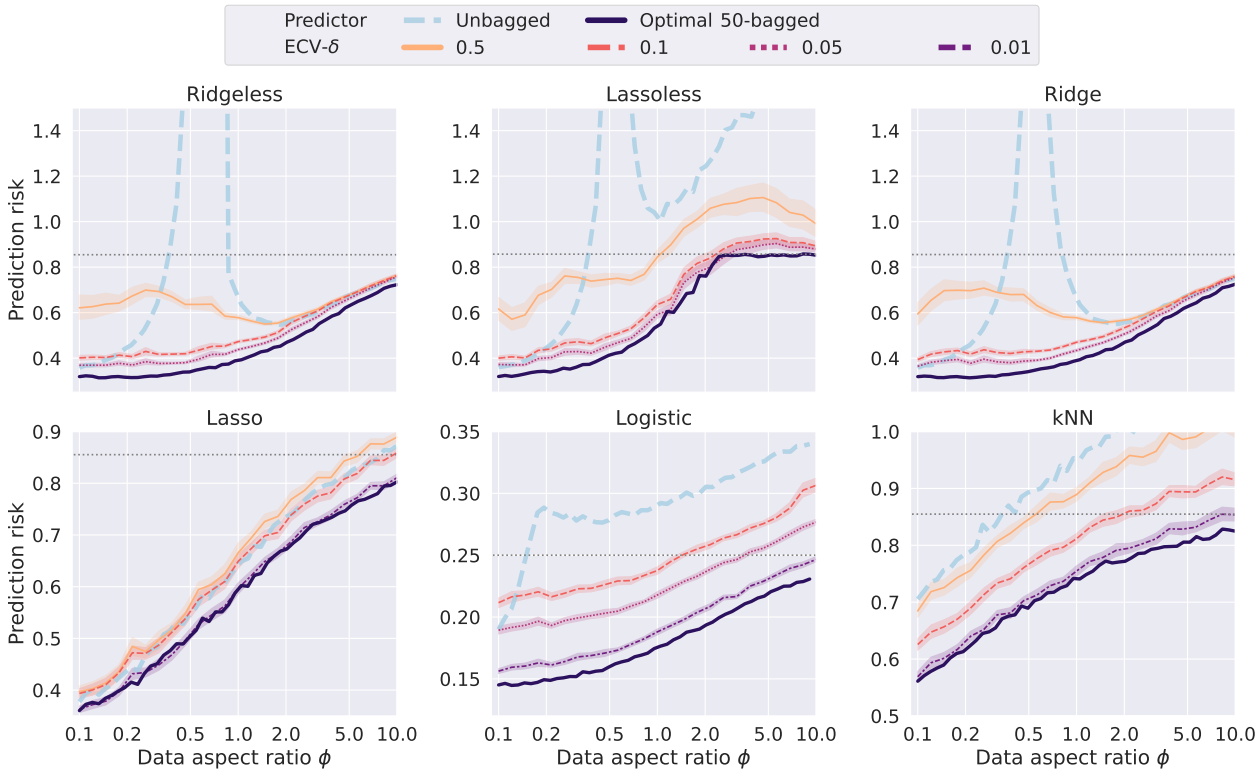

Figure 3: Prediction risk for different bagged predictors by ECV, under model (M2) with $\sigma = \rho_{\text{ar1}} = 0.5$, $M_0 = 10$, and $M_{\max} = 50$, for varying ϕ and tolerance threshold δ. An ensemble is fitted when (4.3) is satisfied with $\zeta = 5$. The null risks and the risks for the non-ensemble predictors are marked as gray dotted lines and blue thick dashed lines, respectively. The points denote finite-sample risks averaged over 100 dataset repetitions, and the shaded regions denote the values within one standard deviation, with $n = 1,000$ and $p = \lfloor n\phi \rfloor$.

5.3 Tuning ensemble sizes of random forests

When the data aspect ratio p/n is too small, tuning both the subsample size k and the ensemble size M may be unnecessary. In such cases, tuning the ensemble size M (in the sense that how large M is sufficient to have good performance) is a more substantial and practical consideration. In this experiment, we apply Algorithm 1 to tune only the ensemble size of random forests.

Since the most crucial advantage of the random forest model is its flexibility to incorporate highly correlated variables while avoiding multi-collinearity issues, we consider the nonlinear model (M2) with a non-isotropic AR(1) covariance matrix. For a given dataset, we examine two strategies to estimate the conditional prediction risks. The first utilizes the OOB observations according to Algorithm 1, while the other uses a hold-out subset to estimate the risks. Similar to Lopes (2019), $\lfloor n/6 \rfloor$ observations are randomly selected as the evaluation set for the hold-out estimates. As suggested by Hastie et al. (2009), each decision tree uses $\lfloor p/3 \rfloor$ randomly selected features with a minimum node size of 5 as the default without pruning. To build each tree, we fix the subsample size $k = n(1 - 1/\log n)$ observations for subagging. The results are shown in fig. 4, where the standard deviation of the estimates are also visualized as error bars. In the underparameterized regime when

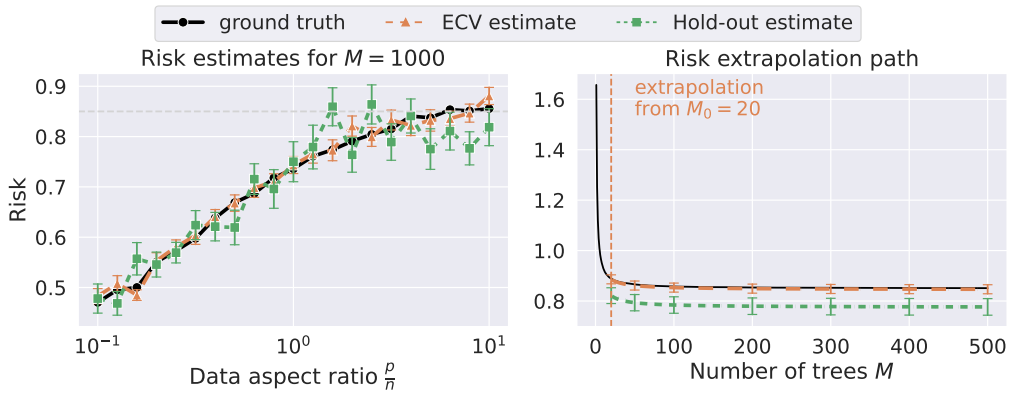

Figure 4: Risk extrapolation for random forest based on the first $M_0 = 20$ trees, under model (M2) with $\sigma = \rho_{\text{ar1}} = 0.5$. The left panel shows the risk estimates for random forests with $M = 500$ trees and varying data aspect ratios, and the gray dash line $y = 0.85$ denotes the risk of the null predictor; the right panel shows the full risk extrapolation path in M when $p/n = 8$. The error bar denotes one standard deviation across 100 simulations, and the ground truth is estimated from $2,000$ test observations, with $n = 500$.

$n > p$, we observe that ECV and hold-out estimates have similar performance. Both of them are close to the out-of-sample errors in this case. However, in the overparameterized regime when $n < p$, the hold-out estimates suffer from biases due to sample splitting. On the contrary, the ECV estimates are still accurate and have smaller variability compared to the hold-out estimates. In the right panel of fig. 4, we see that ECV estimates provide a valid extrapolation path from $M = 20$ to $M = 500$ in the high-dimensional scenarios.

Finally, we conduct a sensitivity analysis of the hyperparameter M_0, the correlative strength ρ_{ar1} of the feature covariance matrix, and the covariance structures in Appendices S6.4.1 to S6.4.3. The results suggest that ECV is relatively robust under various scenarios and various choices of hyperparameters. In Appendix S6.4.4, we illustrate the utility of ECV for tuning the number of features drawn when splitting each node of random forests.

6 Applications to single-cell multiomics

In genomics, cell surface proteins act as primary targets for therapeutic intervention and universal indicators of particular cellular processes. More importantly, immunophenotyping of cell surface proteins has become an essential tool in hematopoiesis, immunology, and cancer research over the past 30 years (Hao et al., 2021). However, most single-cell investigations only quantify the transcriptome without cell-matched measures of related surface proteins due to technical limitations and financial constraints (Zhou et al., 2020; Du et al., 2022). The specific cell types and differentially abundant surface proteins are determined after thoroughly analyzing the transcriptome. This has led researchers to investigate how to reliably predict protein abundances in individual cells using their gene expressions. Specifically, the effectiveness of ensemble methods has been illustrated by (Heckmann et al., 2018; Li et al., 2019; Xu et al., 2021) on the protein prediction problem. Yet, in practice, because of the lack of theoretical results and pragmatic guidelines, the ensemble and subsample sizes are

generally determined by ad hoc criteria.

In this section, we apply the proposed method to real datasets in single-cell multi-omics (Hao et al., 2021). See Appendix S6.5 for details on the datasets and the preprocessing steps. Based on these real-world datasets, we compare three different cross-validation methods for tuning both the ensemble size and the subsample size of random forests: (1) K-fold CV: the K-fold CV ($K = 5$); (2) sample-split CV: sample-split or holdout CV (the ratio of training to validation observations is 5:1); and (3) ECV: the proposed extrapolated CV. The grid of subsample sizes $\mathcal{K}_n$ is generated according to Algorithm 1. To ensure all the CV methods are comparable and fairly evaluated, we evaluate the three CV methods on the same grid $[M_{\max}] \times \mathcal{K}_n$ for $M_{\max} = 50$. Decision trees are used as the base predictors to predict the abundance of each protein based on the gene expressions of subsampled cells. After the tuning parameters $(\widehat{M}, \widehat{k})$ are obtained, we refit the ensemble on the entire training set and evaluate it on the test set. For each method, the $M_{\max}$ base predictors are fitted once so that the training costs are almost the same for all three. The computational complexity of the three methods is discussed in Section 4.2. Because different proteins may have different variances, we measure the overall protein prediction accuracy by the normalized mean squared error (NMSE), which is the ratio of the mean squared error to the empirical variance on the test set. Our target is to select a δ-optimal random forest so that its NMSE is no more than $\delta = 0.05$ away from the best random forest with 50 trees.

To illustrate our proposed method, we visualize the prediction risk estimate and out-of-sample error for surface protein CD103 in figs. S16 and S17. Because the response is centered, the empirical variance of the response serves as an estimate of the null risk, i.e., the risk of the null predictor that always outputs zeros. A value of NMSE less than one indicates that the predictor performs better than the null risk. From fig. S16, we see that the ECV extrapolated estimates in the left panel are largely consistent with the actual prediction errors in the right panel. The out-of-sample error can still be considerable when $M = 10$ as shown in fig. S16. As the ensemble size M increases, both become more stable for various subsample sizes. Further, we observe that the tuned ensemble and subsample sizes are close to the optimal ones on the test dataset in finite samples. The tune subsample size k is much smaller than the total sample size n. This indicates that our proposed method tracks the out-of-sample optimal parameter well.

We compare the performance of different CV methods on various cell types. As shown in fig. 5(a), K-fold CV is very sensitive to the choice of the number of fold K. sample-split CV and K-fold CV with $K = 3$ have the worst out-of-sample performance among all methods. In the low-dimensional dataset of the Mono cell type, all methods have similar performance. However, ECV significantly improves upon sample-split CV with insufficient sample sizes, as in the DC, B, and NK cell types. In all cases, the out-of-sample NMSEs of ECV are comparable to K-fold CV with $K = 5$. Yet, ECV is more flexible to tuning large ensemble sizes, such as $M_{\max} = 100$ and 250. This suggests that the extrapolated risk estimates of ECV based on only $M_0 = 20$ trees are accurate for tuning ensemble parameters on these datasets.

Regarding time consumption, recall that the base predictors are only fitted once (Section 4.2) so that the comparison is fair for all methods. From fig. 5(b), we see that tuning by ECV is significantly faster than K-fold CV because we are extrapolating the risks from 20 trees rather than estimating them for every $M > 20$. Though ECV has similar time

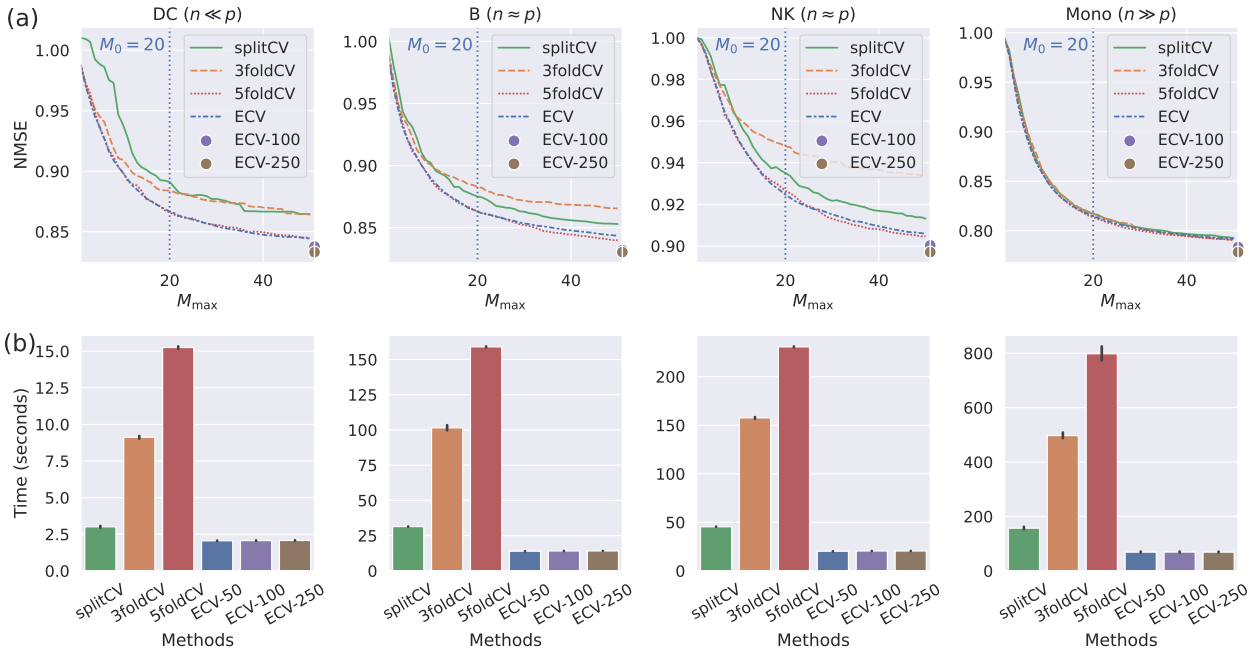

Figure 5: Performance of CV methods on predicting the protein abundances in different cell types. (a) The average NMSEs of the cross-validated predictors for different methods. ECV uses $M_0 = 20$ trees to extrapolate risk estimates, and the points correspond to $M_{\max} \in \{100, 250\}$. (b) The average CV time consumption in seconds.

complexity as sample-split CV when the sample sizes are small, it uses less than 50% of the time used by sample-split CV as the sample size increases. Thus, ECV achieves better computational efficiency than the alternative CV methods in a variety of settings.

7 Discussion

This paper addresses the challenge of tuning the ensemble and subsample sizes for randomized ensemble learning. While previous work has extensively focused on the statistical properties of bagging and random forests, ensemble tuning remains an area that has received comparatively less attention. To bridge this gap, this paper introduces a method that efficiently provides δ-optimal ensemble parameters (for the tuned risk) without requiring an exhaustive search over all feasible parameter combinations, a common limitation of conventional cross-validation methods.

Our method hinges on the extrapolation of risk estimates, which is afforded due to a special decomposition of the squared risk and the consistent risk estimators of each component. Our method is sample-efficient and can be naturally extended to tuning hyperparameters other than subsample sizes, as shown in Appendix S6.4.4. Furthermore, our algorithm guarantees the δ-optimality of the tuned ensemble risk in relation to the oracle ensemble risk.

To demonstrate the practical utility of ECV, we apply it to the task of predicting proteins in single-cell data. In scenarios with small sample sizes, ECV achieves smaller out-of-sample errors without sample splitting compared to traditional sample-split CV. On the other hand, it drastically reduces the time complexity compared to K-fold CV and maintains comparable

accuracy without repeatedly fitting numerous predictors. In summary, ECV exhibits both statistical and computational efficiency in protein prediction tasks.

We next point out some limitations of this work and propose possible future directions. Future directions for this work include extending the theoretical framework to other types of loss functions, such as smooth loss functions, by applying the "sandwich" sub-optimality gap approach (Patil et al., 2023, Proposition 3.6). Additionally, exploring variance estimation and extrapolation for the proposed risk estimators could provide a more comprehensive understanding of their uncertainty. Integrating the concept of performing CV with confidence (Lei, 2020) could address the important issue of quantifying confidence in a tuned model for model selection.

Acknowledgments We thank the editor, the associate editor, and two anonymous reviewers for their valuable and constructive comments, which led to various improvements in this paper. This work used the Bridges-2 system at the Pittsburgh Supercomputing Center (PSC) through allocations BIO220140 and MTH230020 from the Advanced Cyberinfrastructure Coordination Ecosystem: Services & Support (ACCESS) program. This project was partially funded by the National Institute of Mental Health (NIMH) grant R01MH123184.

References

Bellec, P. C. (2018). Optimal bounds for aggregation of affine estimators. *The Annals of Statistics*, 46(1):30–59.

Bickel, P. J., Götze, F., and van Zwet, W. R. (1997). Resampling fewer than n observations: gains, losses, and remedies for losses. *Statistica Sinica*, 7(1):1–31.

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

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

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

Chen, L., Min, Y., Belkin, M., and Karbasi, A. (2020). Multiple descent: Design your own generalization curve. *arXiv preprint arXiv:2008.01036*.

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

Guo, X. and Peterson, J. (2019). Berry–esseen estimates for regenerative processes under weak moment assumptions. *Stochastic Processes and their Applications*, 129(4):1379–1412.

Hao, Y., Hao, S., Andersen-Nissen, E., Mauck, W. M., Zheng, S., Butler, A., Lee, M. J., Wilk, A. J., Darby, C., Zager, M., et al. (2021). Integrated analysis of multimodal single-cell data. *Cell*, 184(13):3573–3587.

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

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

Heckmann, D., Lloyd, C. J., Mih, N., Ha, Y., Zielinski, D. C., Haiman, Z. B., Desouki, A. A., Lercher, M. J., and Palsson, B. O. (2018). Machine learning applied to enzyme turnover numbers reveals protein structural correlates and improves metabolic models. *Nature communications*, 9(1):1–10.

Lei, J. (2020). Cross-validation with confidence. *Journal of the American Statistical Association*, 115(532):1978–1997.

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

Li, H., Siddiqui, O., Zhang, H., and Guan, Y. (2019). Joint learning improves protein abundance prediction in cancers. *BMC Biology*, 17(1):1–14.

Liu, L., Chin, S. P., and Tran, T. D. (2019). Reducing sampling ratios and increasing number of estimates improve bagging in sparse regression. In *2019 53rd Annual Conference on Information Sciences and Systems (CISS)*, pages 1–5. IEEE.

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

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

Lugosi, G. and Mendelson, S. (2019). Mean estimation and regression under heavy-tailed distributions: A survey. *Foundations of Computational Mathematics*, 19(5):1145–1190.

Martínez-Muñoz, G. and Suárez, A. (2010). Out-of-bag estimation of the optimal sample size in bagging. *Pattern Recognition*, 43(1):143–152.

Mentch, L. and Hooker, G. (2016). Quantifying uncertainty in random forests via confidence intervals and hypothesis tests. *The Journal of Machine Learning Research*, 17(1):841–881.

Oshiro, T. M., Perez, P. S., and Baranauskas, J. A. (2012). How many trees in a random forest? In *International workshop on machine learning and data mining in pattern recognition*.

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

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

Peng, W., Coleman, T., and Mentch, L. (2022). Rates of convergence for random forests via generalized U-statistics. *Electronic Journal of Statistics*, 16(1):232–292.

Politis, D. N. (2023). Scalable subsampling: computation, aggregation and inference. *Biometrika*. asad021.

Politis, D. N. and Romano, J. P. (1994). Large sample confidence regions based on subsamples under minimal assumptions. *The Annals of Statistics*, pages 2031–2050.

Pugh, C. C. (2002). *Real Mathematical Analysis*. Springer.

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

Rio, E. (2017). About the constants in the fuk-nagaev inequalities. *Electronic Communications in Probability*, 22(28):12p.

Scornet, E., Biau, G., and Vert, J.-P. (2015). Consistency of random forests. *The Annals of Statistics*, pages 1716–1741.

Vershynin, R. (2018). *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge University Press.

Wager, S. and Athey, S. (2018). Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523):1228–1242.

Wager, S., Hastie, T., and Efron, B. (2014). Confidence intervals for random forests: The jackknife and the infinitesimal jackknife. *The Journal of Machine Learning Research*, 15(1):1625–1651.

Wang, S., Zhou, W., Lu, H., Maleki, A., and Mirrokni, V. (2018). Approximate leave-one-out for fast parameter tuning in high dimensions. In *International Conference on Machine Learning*.

Xu, F., Wang, S., Dai, X., Mundra, P. A., and Zheng, J. (2021). Ensemble learning models that predict surface protein abundance from single-cell multimodal omics data. *Methods*, 189:65–73.

Zhou, Z., Ye, C., Wang, J., and Zhang, N. R. (2020). Surface protein imputation from single cell transcriptomes by deep neural networks. *Nature communications*, 11(1):651.

Supplementary material for "Extrapolated cross-validation for randomized ensembles"

This document acts as a supplement to the paper "Extrapolated cross-validation for randomized ensembles." The section numbers in this supplement begin with the letter "S" and the equation numbers begin with the letter "S" to differentiate them from those appearing in the main paper.

Notation and organization

Notation

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

1. General notation: We denote scalars in non-bold lower or upper case (e.g., n, λ, C), vectors in lower case (e.g., $\boldsymbol{x}$, $\boldsymbol{\beta}$), and matrices in upper case (e.g., $\boldsymbol{X}$). For a real number x, $(x)_+$ denotes its positive part, $\lfloor x \rfloor$ its floor, and $\lceil x \rceil$ its ceiling. For a natural number n, $n! = \prod_{i=1}^{n} i$ denotes the n factorial. For a vector $\boldsymbol{\beta}$, $\|\boldsymbol{\beta}\|_2$ denotes its ℓ_2 norm. For a pair of vectors $\boldsymbol{v}$ and $\boldsymbol{w}$, $\langle \boldsymbol{v}, \boldsymbol{w} \rangle$ denotes their inner product. For an event A, $\mathbb{1}_A$ denotes the associated indicator random variable. We use $\mathcal{O}_p$ and o_p to denote probabilistic big-O and little-o notation, respectively.

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

3. Matrix notation: For a matrix $\boldsymbol{X} \in \mathbb{R}^{n \times p}$, $\boldsymbol{X}^\top \in \mathbb{R}^{p \times n}$ denotes its transpose. For a square matrix $\boldsymbol{A} \in \mathbb{R}^{p \times p}$, $\boldsymbol{A}^{-1} \in \mathbb{R}^{p \times p}$ denotes its inverse, provided it is invertible. For a positive semi-definite 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.

Organization

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

- In Appendix S2, we present proofs of results appearing in Section 3.
  - Appendix S2.1 proves Proposition 3.1.
  - Appendix S2.2 proves Proposition 3.2.
  - Appendix S2.3 proves Theorem 3.3, conditional on certain helper lemmas presented in the next section.
- Appendix S3 provides intermediate concentration results used in the proof of Theorem 3.3.

- In Appendix S4, we present proof of results in Section 4.
  - Appendix S4.1 proves Theorem 4.1.
  - Appendix S4.2 extends additive optimality in Theorem 4.1 to multiplicative optimality stated in Remark 4.2.
  - Appendix S4.3 specialize Theorem 4.1 to ridge predictor under weaker assumptions.
- In Appendix S5, we collect various technical helper lemmas related to concentrations and convergences along with their proofs that are used in various proofs in Appendices S2 to S4.
- In Appendix S6, we present additional numerical results for Section 5 and Section 6.
  - Appendix S6.1 presents additional illustrations for subagging and bagging in Section 5.1.
  - Appendix S6.2 presents additional illustrations for subagging and bagging in Section 5.2.
  - Appendix S6.3 presents results of ECV on imbalanced classification.
  - Appendix S6.4 presents additional illustrations for bagging in Section 5.3.
  - Appendix S6.5 presents additional illustrations for Section 6.

S1 Related work on cross-validation

Different cross-validation (CV) approaches have been proposed for parameter tuning and model selection (Allen, 1974; Stone, 1974, 1977; Geisser, 1975). We refer readers to Arlot and Celisse (2010); Zhang and Yang (2015) for a review of different CV variants used in practice. The simplest version of CV is the sample-split CV (Hastie et al., 2009), which holds out a specific portion of the data to evaluate models with different parameters. By repeated fitting of each candidate model on multiple subsets of the data, K-fold CV extends the idea of the sample splitting and reduces the estimation uncertainty. When K is small, the risk estimate may inherit more uncertainty; however, it can be computationally prohibitive when K is large. Asymptotic distributions of suitably normalized K-fold CV are obtained in Austern and Zhou (2020), under some stability conditions on the predictors.

In a high-dimensional regime where the number of variables is comparable to the number of observations, the commonly-used small values of K such as 5 or 10 suffer from bias issues in risk estimation (Rad and Maleki, 2020). Leave-one-out cross-validation (LOOCV), i.e., the case when $K = n$, alleviates the bias issues in risk estimation, whose theoretical properties have been analyzed in recent years by Kale et al. (2011); Kumar et al. (2013); Rad et al. (2020). However, LOOCV, in general, is computationally expensive to evaluate, and there has been some work on approximate LOOCV to address the computational issues (Wang et al., 2018; Stephenson and Broderick, 2020; Wilson et al., 2020; Rad and Maleki, 2020). Another line of research about CV is on statistical inference; see, for example, Wager et al. (2014); Lei (2020); Bates et al. (2021). Central limit theorems for CV error and a consistent

estimator of its variance are derived in Bayle et al. (2020), which assumes certain stability assumptions, similar to Kumar et al. (2013); Celisse and Guedj (2016). Their results yield asymptotic confidence intervals for the prediction error and apply to K-fold CV and LOOCV. A naive application of these traditional CV methods for ensemble learning to tune M and k requires fitting the ensembles of arbitrary sizes M, leading to a higher computational cost.

S2 Proofs of results in Section 3

S2.1 Proof of Proposition 3.1 (Squared risk decomposition)

Proof of Proposition 3.1. We start by expanding the squared risk as:

$$
\begin{aligned}
&R(\widetilde{f}_{M,k};\, \mathcal{D}_n, \{I_\ell\}_{\ell=1}^M) \\
&= \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,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,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} \Big\{ 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 \\
&\qquad\qquad - \big(y - \widehat{f}(\boldsymbol{x}; \mathcal{D}_{I_i}) \big)^2 - \big(y - \widehat{f}(\boldsymbol{x}; \mathcal{D}_{I_j}) \big)^2 \Big\} \, \mathrm{d}P(\boldsymbol{x}, y) \\
&= \frac{1}{M^2} \sum_{\ell=1}^{M} R(\widetilde{f}_{1,k}; \mathcal{D}_n, I_\ell) \\
&\quad + \frac{1}{M^2} \sum_{i=1}^{M} \sum_{\substack{j=1 \\ j \neq i}}^{M} \frac{1}{2} \Big\{ 4R(\widehat{f}_{2,k}; \mathcal{D}_n; I_i, I_j) - R(\widetilde{f}_{1,k}; \mathcal{D}_n; I_i) - R(\widetilde{f}_{1,k}; \mathcal{D}_n; I_j) \Big\} \\
&= \frac{1}{M^2} \sum_{\ell=1}^{M} R(\widetilde{f}_{1,k}; \mathcal{D}_n, I_\ell)
\end{aligned}
$$

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

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

S2.2 Proof of Proposition 3.2 (Consistent component risk estimation)

Proof of Proposition 3.2. Let $\Delta_n = |\widehat{R}(f, \mathcal{D}_I) - R(\widehat{f}; \mathcal{D}_I)|$. We will view p, $|I|$, and $|I^c|$ as sequences indexed by n. Define $\widehat{\sigma}_I = \|(y_0 - \widehat{f}(\boldsymbol{x}_0; \mathcal{D}_I))^2\|_{\psi_1|\mathcal{D}_I}$. From Patil et al. (2023, Lemma 2.9, Lemma 2.10), we have

$$
\mathbb{P}\left(\Delta_n \geq C\widehat{\sigma}_I \max\left\{\sqrt{\frac{A\log n}{|I^c|}}, \frac{A\log n}{|I^c|}\right\}\right) \leq n^{-A}, \tag{S2.1}
$$

for some positive constant C. Let $\kappa_n = C\widehat{\sigma}_I \max\left\{\sqrt{\frac{A\log n}{|I^c|}}, \frac{A\log n}{|I^c|}\right\}$.

Since $\widehat{\sigma}_I = o_p(\sqrt{|I^c|/\log n})$ as $n \to \infty$, we have that $\kappa_n = o_p(1)$. Let $A > 0$ be fixed, For all $\epsilon > 0$, we have that

$$
\begin{aligned}
\mathbb{P}(\Delta_n > \epsilon) &= \mathbb{P}(\Delta_n > \epsilon \geq \kappa_n) + \mathbb{P}(\Delta_n > \epsilon, \kappa_n > \epsilon) \\
&\leq n^{-A} + \mathbb{P}(\kappa_n > \epsilon) \to 0,
\end{aligned}
$$

which implies that $\Delta_n = o_p(1)$. The proof for $\|\cdot\|_{L_2|\mathcal{D}_I}$ follows analogously. □

S2.3 Proof of Theorem 3.3 (Uniform risk estimation over (M, k))

Proof of Theorem 3.3. Before we prove Theorem 3.3, we present the proof strategy in fig. S1. In fig. S1, four important auxiliary lemmas and propositions are deferred to the next section.

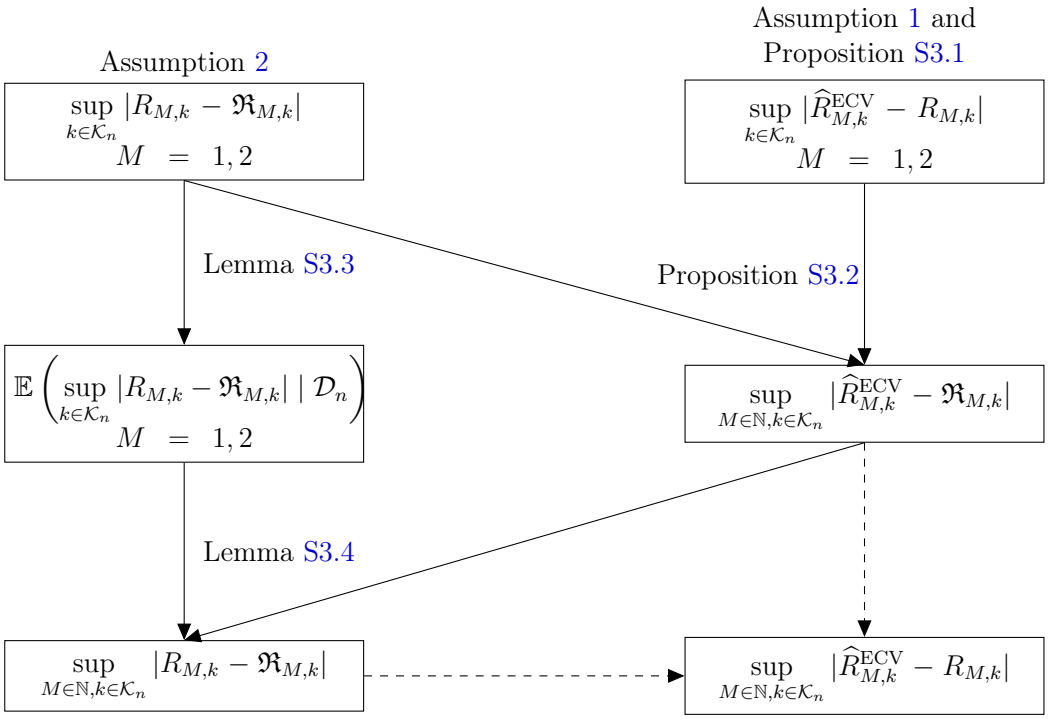

Figure S1: Reduction strategy for obtaining concentration results in Theorem 3.3. The solid lines indicate basic components presented later in Appendix S3 and the dashed lines indicate the proof strategy in Appendix S2.3. Here, $\mathfrak{R}_{M,k} = 2\mathfrak{R}_{2,k} - \mathfrak{R}_{1,k} + 2(\mathfrak{R}_{1,k} - \mathfrak{R}_{2,k})/M$.

For the asymptotic results, we will let k, p be sequences of integers $\{k_n\}_{n=1}^{\infty}$, $\{p_n\}_{n=1}^{\infty}$ indexed by n but drop the subscripts n.

From Lemma S3.4 we have

$$\sup_{M\in\mathbb{N},k\in\mathcal{K}_n} |R_{M,k} - \mathfrak{R}_{M,k}| = \mathcal{O}_p(n^{\epsilon}(\gamma_{1,n} + \gamma_{2,n})).$$

On the other hand, from Proposition S3.2 we have

$$\sup_{M\in\mathbb{N},k\in\mathcal{K}_n} |\widehat{R}^{\mathrm{ECV}}_{M,k} - \mathfrak{R}_{M,k}| = \mathcal{O}_p(\widehat{\sigma}_n \log n/n^{1/2} + n^{\epsilon}(\gamma_{1,n} + \gamma_{2,n})),$$

where $\widehat{\sigma}_n = \max_{m,\ell\in[M_0],k\in\mathcal{K}_n} \widehat{\sigma}_{I_{k,\ell}\cup I_{k,m}}$. By the triangle inequality, we have

$$\sup_{M\in\mathbb{N},k\in\mathcal{K}_n} |\widehat{R}^{\mathrm{ECV}}_{M,k} - R_{M,k}| = \mathcal{O}_p(\widehat{\sigma}_n \log n/n^{1/2} + n^{\epsilon}(\gamma_{1,n} + \gamma_{2,n})),$$

which completes the proof. □

S3 Intermediate concentration results for Theorem 3.3

In this section, we show the concentration of conditional prediction risks to their limits. Proposition S3.1 derives the uniform consistency over $k \in \mathcal{K}_n$ of the cross-validated risk estimates to the risk $R_{M,k}$ for $M = 1, 2$. Proposition S3.2 derives the uniform consistency

over $(M,k) \in \mathbb{N} \times \mathcal{K}_n$ of the cross-validated risk estimates to the deterministic limits $\mathfrak{R}_{M,k}$. Lemma S3.3 establishes the concentration for the expected (with respect to sampling) conditional risk over $k \in \mathcal{K}_n$. Lemma S3.4 establishes the concentration for subsample conditional risks over $M \in \mathbb{N}$ and $k \in \mathcal{K}_n$.

S3.1 Uniform consistency of cross-validated risk over k for $M = 1, 2$

In what follows, we derive the uniform consistency over $k \in \mathcal{K}_n$ of the cross-validated risk estimates for $M = 1, 2$ in Appendix S3.1.

Proposition S3.1 (Uniform consistency over k for $M = 1, 2$)**.** *Suppose that Assumption 2 holds, then ECV estimates defined in (3.3) satisfy that for $M = 1, 2$,*

$$\sup_{k \in \mathcal{K}_n} \left| \widehat{R}^{\mathrm{ECV}}_{M,k} - R_{M,k} \right| = \mathcal{O}_p(\widehat{\sigma}_n \log(n) n^{-1/2} + n^{\epsilon}(\gamma_{1,n} + \gamma_{2,n})),$$

where $\widehat{\sigma}_n = \max_{m,\ell \in [M_0], k \in \mathcal{K}_n} \widehat{\sigma}_{I_{k,\ell} \cup I_{k,m}}$ and $R_{M,k} = R_{M,k}(\widetilde{f}_{M,k}; \mathcal{D}_n, \{I_{k,\ell}\}_{\ell=1}^M)$ for $\{I_{k,\ell}\}_{\ell=1}^M \overset{\mathrm{SRS}}{\sim} \mathcal{I}_k$.

Proof of Proposition S3.1. Note that

$$\begin{aligned}
\widehat{R}^{\mathrm{ECV}}_{1,k} &= \frac{1}{M_0} \sum_{\ell=1}^{M_0} \widehat{R}(\widehat{f}(\cdot; \{\mathcal{D}_{I_{k,\ell}}\}), \mathcal{D}_{I^c_{k,\ell}}) \\
\widehat{R}^{\mathrm{ECV}}_{2,k} &= \frac{1}{M_0(M_0-1)} \sum_{\substack{\ell,m \in [M_0] \\ \ell \neq m}} \widehat{R}(\widehat{f}(\cdot; \{\mathcal{D}_{I_{k,\ell}}, \mathcal{D}_{I_{k,m}}\}), \mathcal{D}_{(I_{k,\ell} \cup I_{k,m})^c}).
\end{aligned}$$

Define $\widehat{\sigma}_I = \|(y_0 - \widehat{f}(\boldsymbol{x}_0; \mathcal{D}_I))^2\|_{\psi_1 | \mathcal{D}_I}$ and recall $\Delta_{M,k}$ for $M = 1, 2$ are defined as follows:

$$\begin{aligned}
\Delta_{1,k} &= \widehat{R}^{\mathrm{ECV}}_{1,k} - \frac{1}{M_0} \sum_{\ell=1}^{M_0} R_{1,k}(\widehat{f}; \mathcal{D}_n, \{I_{k,\ell}\}) \\
\Delta_{2,k} &= \widehat{R}^{\mathrm{ECV}}_{2,k} - \frac{1}{M_0(M_0-1)} \sum_{\substack{\ell,m \in [M_0] \\ \ell \neq m}} R_{2,k}(\widehat{f}_{2,k}; \mathcal{D}_n, \{I_{k,\ell}, I_{k,m}\}).
\end{aligned} \tag{S3.1}$$

By the triangle inequality, for $M = 1, 2$, it holds that

$$\sup_{k \in \mathcal{K}_n} |\widehat{R}^{\mathrm{ECV}}_{1,k} - \mathfrak{R}_{1,k}| \leq \sup_{k \in \mathcal{K}_n} |\Delta_{1,k}| + \sup_{k \in \mathcal{K}_n, \ell \in [M_0]} |R_{1,k}(\widehat{f}; \mathcal{D}_n, \{I_{k,\ell}\}) - \mathfrak{R}_{1,k}| \tag{S3.2}$$

$$\sup_{k \in \mathcal{K}_n} |\widehat{R}^{\mathrm{ECV}}_{2,k} - \mathfrak{R}_{1,k}| \leq \sup_{k \in \mathcal{K}_n} |\Delta_{2,k}| + \sup_{k \in \mathcal{K}_n, m, \ell \in [M_0]} |R_{2,k}(\widehat{f}; \mathcal{D}_n, \{I_{k,m}, I_{k,\ell}\}) - \mathfrak{R}_{1,k}| \tag{S3.3}$$

where $\Delta_{M,k}$ are defined in (S3.1). Next we analyze each term separately for $M = 1, 2$.

Part (1). For the first term, from (S2.1) in Proposition 3.2 and applying union bound over $k \in \mathcal{K}_n$, we have that

$$\mathbb{P}\left(\sup_{k\in\mathcal{K}_n}|\Delta_{1,k}| \geq C \max_{k\in\mathcal{K}_n,\ell\in[M_0]} \widehat{\sigma}_{I_{k,\ell}} \max\left\{\sqrt{\frac{\log(M_0|\mathcal{K}_n|\eta)}{n/\log n}}, \frac{\log(M_0|\mathcal{K}_n|\eta)}{n/\log n}\right\}\right) \leq \frac{1}{\eta},$$
$$\mathbb{P}\left(\sup_{k\in\mathcal{K}_n}|\Delta_{2,k}| \geq C \max_{k\in\mathcal{K}_n,\ell\neq m\in[M_0]} \widehat{\sigma}_{I_{k,\ell}\cup I_{k,m}} \max\left\{\sqrt{\frac{\log(M_0|\mathcal{K}_n|\eta)}{n/\log n}}, \frac{\log(M_0|\mathcal{K}_n|\eta)}{n/\log n}\right\}\right) \leq \frac{1}{\eta}, \tag{S3.4}$$

for some constant $C > 0$. This implies that

$$\sup_{k\in\mathcal{K}_n}|\Delta_{1,k}| = \mathcal{O}_p\left(\max_{k\in\mathcal{K}_n,\ell\in[M_0]} \widehat{\sigma}_{I_{k,\ell}}\sqrt{\frac{\log(|\mathcal{K}_n|)\log n}{n}}\right)$$
$$\sup_{k\in\mathcal{K}_n}|\Delta_{2,k}| = \mathcal{O}_p\left(\max_{k\in\mathcal{K}_n,\ell\neq m\in[M_0]} \widehat{\sigma}_{I_{k,\ell}\cup I_{k,m}}\sqrt{\frac{\log(|\mathcal{K}_n|)\log n}{n}}\right).$$

To summarize, $\sup_{k\in\mathcal{K}_n}|\Delta_{M,k}|$ for $M = 1, 2$ can be bounded as:

$$\sup_{k\in\mathcal{K}_n}|\Delta_{M,k}| = \mathcal{O}_p\left(\widehat{\sigma}_n\sqrt{\frac{\log^2 n}{n}}\right), \tag{S3.5}$$

where $\widehat{\sigma}_n = \max_{m,\ell\in[M_0],k\in\mathcal{K}_n} \widehat{\sigma}_{I_{k,\ell}\cup I_{k,m}}$.

Part (2). For the second term, from (3.5) in Assumption 2 we have that when n is large enough, for all $\eta \geq 1$,

$$\mathbb{P}\left(n^{-\epsilon}\gamma_{1,n}^{-1}|R(\widetilde{f}_{1,k};\mathcal{D}_{I_{k,\ell}}) - \mathfrak{R}_{1,k}| \geq \eta\right) \leq \frac{C_0}{n\eta^{1/\epsilon}}.$$

Let $\gamma'_{1,n} = n^{\epsilon}\gamma_{1,n}$. Taking a union bound over $k \in \mathcal{K}_n$ and $\ell \in [M_0]$, we have

$$\mathbb{P}\left(\gamma'^{-1}_{1,n}\sup_{k\in\mathcal{K}_n,\ell\in[M_0]}|R_{1,k}(\widehat{f};\mathcal{D}_n,\{I_{k,\ell}\}) - \mathfrak{R}_{1,k}| \geq \eta\right) \leq \frac{C_0 M_0}{\eta^{1/\epsilon}},$$

which implies that

$$\sup_{k\in\mathcal{K}_n,\ell\in[M_0]}|R_{1,k}(\widehat{f};\mathcal{D}_n,\{I_{k,\ell}\}) - \mathfrak{R}_{1,k}| = \mathcal{O}_p(n^{\epsilon}\gamma_{1,n}). \tag{S3.6}$$

Analogously, for $M = 2$, we also have

$$\sup_{k\in\mathcal{K}_n,m\neq\ell\in[M_0]}|R_{2,k}(\widehat{f};\mathcal{D}_n,\{I_{k,m},I_{k,\ell}\}) - \mathfrak{R}_{1,k}| = \mathcal{O}_p(n^{\epsilon}\gamma_{2,n}). \tag{S3.7}$$

Part (3). Combining (S3.3), (S3.5), (S3.6) and (S3.7) yields that

$$\sup_{k\in\mathcal{K}_n}|\widehat{R}^{\mathrm{ECV}}_{M,k} - \mathfrak{R}_{M,k}| = \mathcal{O}_p\left(\widehat{\sigma}_n\sqrt{\frac{\log^2 n}{n}} + n^{\epsilon}\gamma_{M,n}\right), \qquad M = 1, 2,$$

ignoring constant factors. □

S3.2 Uniform consistency of cross-validated risk to $\mathfrak{R}_{M,k}$ over (M,k)

Proposition S3.2 (Uniform consistency to $\mathfrak{R}_{M,k}$ over (M,k))**.** *Suppose that Assumptions 1 and 2 hold, then ECV estimates defined in (3.3) satisfy that*

$$\sup_{M\in\mathbb{N},k\in\mathcal{K}_n}\left|\widehat{R}^{\mathrm{ECV}}_{M,k}-\mathfrak{R}_{M,k}\right|=\mathcal{O}_p(\widehat{\sigma}_n\log(n)n^{-\frac{1}{2}}+n^{\epsilon}(\gamma_{1,n}+\gamma_{2,n})),$$

where $\widehat{\sigma}_n=\max_{m,\ell\in[M_0],k\in\mathcal{K}_n}\widehat{\sigma}_{I_{k,\ell}\cup I_{k,m}}$.

Proof of Proposition S3.2. Note that $\widehat{R}^{\mathrm{ECV}}_{M,k}$ satisfies the squared risk decomposition and the randomness of $\widehat{R}^{\mathrm{ECV}}_{M,k}$ also due to both the full data $\mathcal{D}_n$ and random sampling.

From Proposition S3.1, we have

$$\sup_{k\in\mathcal{K}_n}|\widehat{R}^{\mathrm{ECV}}_{M,k}-\mathfrak{R}_{M,k}|=\mathcal{O}_p\left(\widehat{\sigma}_n\sqrt{\frac{\log^2 n}{n}}+n^{\epsilon}\gamma_{M,n}\right),\qquad M=1,2.$$

On the other hand, by the definition of $\widehat{R}^{\mathrm{ECV}}_{M,k}$ for $M\in\mathbb{N}$ in (3.4), taking the supremum over both M and k yields that

$$\begin{aligned}\sup_{M\in\mathbb{N},k\in\mathcal{K}_n}\left|\widehat{R}^{\mathrm{ECV}}_{M,k}-\mathfrak{R}_{M,k}\right|&\le\sup_{k\in\mathcal{K}_n}\left|\widehat{R}^{\mathrm{ECV}}_{1,k}-\mathfrak{R}_{1,k}\right|+2\sup_{k\in\mathcal{K}_n}\left|\widehat{R}^{\mathrm{ECV}}_{2,k}-\mathfrak{R}_{2,k}\right|\\&=\mathcal{O}_p\left(\zeta_n\right),\end{aligned}$$

where $\zeta_n=\widehat{\sigma}_n\sqrt{\log^2 n/n}+n^{\epsilon}(\gamma_{1,n}+\gamma_{2,n})$. □

S3.3 Proof of Lemma S3.3 (Concentration of expected risk over $k\in\mathcal{K}_n$)

To obtain tail bounds for the subsample conditional risk defined in (2.3), we need to analyze its conditional expectation. Here the expectation is taken with respect to only the randomness due to sampling and conditioned on $\mathcal{D}_n$. For example, we define the data conditional (on $\mathcal{D}_n$) risks as:

$$R(\widetilde{f}_{M,k}(\cdot;\{\mathcal{D}_{I_\ell}\}_{\ell=1}^M);\,\mathcal{D}_n)=\int\mathbb{E}\left[\left(y-\widetilde{f}_{M,k}(\boldsymbol{x};\{\mathcal{D}_{I_{k,\ell}}\}_{\ell=1}^M)\right)^2\mid\mathcal{D}_n\right]\mathrm{d}P(\boldsymbol{x},y).\tag{S3.8}$$

Observe that the conditional (on $\mathcal{D}_n$) risk of the bagged predictor $\widetilde{f}_{M,k}(\cdot;\{\mathcal{D}_{I_{k,\ell}}\}_{\ell=1}^M)$ integrates over the randomness of the future observation $(\boldsymbol{x},y)$ as well as the randomness due the simple random sampling of $I_{k,\ell}$, $\ell=1,\ldots,M$. Nevertheless, the subsample conditional risk ignores the expectation over the simple random sample. Considering a more general setup, when we need to obtain tail bounds for $\sup_{k\in\mathcal{K}}|R(\widetilde{f}_M;\mathcal{D}_n,\{I_{k,\ell}\}_{\ell=1}^M)-\mathfrak{R}_{M,k}|$, we again need to control its expectation conditional on $\mathcal{D}_n$. The result is summarized as in the following lemma. Note that when $|\mathcal{K}_n|=1$, it simply reduces to controlling the data conditional risk.

Lemma S3.3 (Concentration of expected risk). *Consider a dataset $\mathcal{D}_n$ with n observations, a subsample grid $\mathcal{K}_n \subset [n]$, and a base predictor $\widehat{f}$. Suppose Assumptions 1 and 2 hold, then it holds for $M = 1, 2$ that*

$$\mathbb{E}\left(\sup_{k\in\mathcal{K}_n} |R(\widetilde{f}_M; \mathcal{D}_n, \{I_{k,\ell}\}_{\ell=1}^M) - \mathfrak{R}_{M,k}|\right) = \mathcal{O}(n^\epsilon \gamma_{M,n}),$$

where $\{I_{k,\ell}\}_{\ell=1}^M \overset{\text{SRS}}{\sim} \mathcal{I}_k$.

Proof of Lemma S3.3. Define

$$B_{1,n} = \sup_{k\in\mathcal{K}_n} |R(\widetilde{f}_{1,k}; \mathcal{D}_n, \{I_{k,1}\}) - \mathfrak{R}_{1,k}|. \tag{S3.9}$$

$$B_{2,n} = \sup_{k\in\mathcal{K}_n} |R(\widetilde{f}_{2,k}; \mathcal{D}_n, \{I_{k,1}, I_{k,2}\}) - \mathfrak{R}_{2,k}|. \tag{S3.10}$$

We start with the $M = 1$ case by bounding the expectation of (S3.9). Note that from (3.5), we have that for all $\eta \geq \eta_0$,

$$\mathbb{P}\left(n^{-\epsilon}\gamma_{1,n}^{-1}|R(\widetilde{f}_{1,k}; \mathcal{D}_{I_1}) - \mathfrak{R}_{1,k}| \geq \eta\right) \leq \frac{1}{n\eta^{1/\epsilon}}$$

Let $\gamma'_{1,n} = n^\epsilon \gamma_{1,n}$. Taking a union bound over $k \in \mathcal{K}_n$, we have

$$\mathbb{P}(\gamma'^{-1}_{1,n} B_{1,n} \geq \eta) \leq \frac{1}{\eta^{1/\epsilon}}. \tag{S3.11}$$

Then, it follows that

$$\begin{aligned}
\mathbb{E}(B_{1,n}) &= \gamma'_{1,n} \int_0^\infty \mathbb{P}(\gamma'^{-1}_{1,n} B_{1,n} \geq \eta)\, \mathrm{d}\eta \\
&\leq \gamma'_{1,n} \left(\int_0^{\eta_0} 1\, \mathrm{d}\epsilon + \int_{\eta_0}^\infty \frac{C_0}{\eta^{1/\epsilon}}\, \mathrm{d}\epsilon\right) \\
&\leq \gamma'_{1,n} \left(\eta_0 - C_0 \frac{\epsilon - 1}{\epsilon} \eta^{1-1/\epsilon}\Big|_{\eta_0}^\infty\right) \\
&= \left(\eta_0 + C_0 \frac{\epsilon - 1}{\epsilon} \eta_0^{1-1/\epsilon}\right) \gamma'_{1,n}.
\end{aligned}$$

The proof for $M = 2$ follows analogously. □

S3.4 Proof of Lemma S3.4 (Concentration of conditional risk)

Lemma S3.4 (Concentration of conditional risk). *Consider a dataset $\mathcal{D}_n$ with n observations and a base predictor $\widehat{f}$. Under Assumption 2, it holds that,*

$$\sup_{M\in\mathbb{N}, k\in\mathcal{K}_n} |R_{M,k} - \mathfrak{R}_{M,k}| = \mathcal{O}_p\left(C_1(|\mathcal{K}_n|)\gamma_{1,n} + C_2(|\mathcal{K}_n|)\gamma_{1,n}\right), \tag{S3.12}$$

where $\mathfrak{R}_{M,k} = 2\mathfrak{R}_{2,k} - \mathfrak{R}_{1,k} + 2(\mathfrak{R}_{1,k} - \mathfrak{R}_{2,k})/M$.

Proof of Lemma S3.4. By Proposition 3.1, we have for $\{I_{k,\ell}\}_{\ell=1}^{M} \overset{\text{SRS}}{\sim} \mathcal{I}_k$

$$\begin{aligned}
&\left|R(\widetilde{f}_{M,k}; \mathcal{D}_n, \{I_{k,\ell}\}_{\ell=1}^{M}) - \mathfrak{R}_{M,k}\right| \\
&= \left|R(\widetilde{f}_{M,k}; \mathcal{D}_n, \{I_{k,\ell}\}_{\ell=1}^{M}) - \left[(2\mathfrak{R}_{2,k} - \mathfrak{R}_{1,k}) + \frac{2(\mathfrak{R}_{1,k} - \mathfrak{R}_{2,k})}{M}\right]\right| \\
&= \left|-\left(\frac{1}{M} - \frac{2}{M^2}\right)\sum_{\ell=1}^{M}\left(R(\widetilde{f}_{1,k};\, \mathcal{D}_n, \{I_{k,\ell}\}) - \mathfrak{R}_{1,k}\right)\right| \\
&\qquad + \left|\frac{2}{M^2}\sum_{\substack{i,j\in[M]\\ i\neq j}}\left(R(\widetilde{f}_{2,k};\, \mathcal{D}_n, \{I_{k,i}, I_{k,j}\}) - \mathfrak{R}_{2,k}\right)\right|.
\end{aligned}$$

Then, it follows that

$$\begin{aligned}
&\sup_{M\in\mathbb{N}, k\in\mathcal{K}_n}\left|R(\widetilde{f}_{M,k}; \mathcal{D}_n, \{I_{k,\ell}\}_{\ell=1}^{M}) - \mathfrak{R}_{M,k}\right| \\
&\leq \left|1 - \frac{2}{M}\right| \sup_{M\geq 1}\left|\frac{1}{M}\sum_{\ell\in[M]}\sup_{k\in\mathcal{K}_n}|R(\widetilde{f}_{1,k};\, \mathcal{D}_n, \{I_{k,\ell}\}) - \mathfrak{R}_{1,k}|\right| \\
&\qquad + 2\sup_{M\geq 2}\left|\frac{1}{M(M-1)}\sum_{i,j\in[M], i\neq j}\sup_{k\in\mathcal{K}_n}|R(\widetilde{f}_{2,k};\, \mathcal{D}_n, \{I_i, I_j\}) - \mathfrak{R}_{2,k}|\right| \\
&\leq \sup_{M\geq 1}\left|\frac{1}{M}\sum_{\ell\in[M]}\sup_{k\in\mathcal{K}_n}|R(\widetilde{f}_{1,k};\, \mathcal{D}_n, \{I_{k,\ell}\}) - \mathfrak{R}_{1,k}|\right| \\
&\qquad + 2\sup_{M\geq 2}\left|\frac{1}{M(M-1)}\sum_{i,j\in[M], i\neq j}\sup_{k\in\mathcal{K}_n}|R(\widetilde{f}_{2,k};\, \mathcal{D}_n, \{I_{k,i}, I_{k,j}\}) - \mathfrak{R}_{2,k}|\right|. \qquad \text{(S3.13)}
\end{aligned}$$

We start by observing that the two terms

$$\begin{aligned}
U_M &= \frac{1}{M}\sum_{\ell\in[M]}\sup_{k\in\mathcal{K}_n}|R(\widetilde{f}_{1,k}; \mathcal{D}_n, \{I_{k,\ell}\}) - \mathfrak{R}_{1,k}|, \\
U'_M &= \frac{1}{M(M-1)}\sum_{i,j\in[M], i\neq j}\sup_{k\in\mathcal{K}_n}|R(\widetilde{f}_{2,k}; \mathcal{D}_n, \{I_{k,i}, I_{k,j}\}) - \mathfrak{R}_{2,k}|
\end{aligned}$$

are U-statistics based on sample $\mathcal{D}_{I_1}, \ldots, \mathcal{D}_{I_M}$. Theorem 2 in Section 3.4.2 of Lee (1990) implies that $\{U_M\}_{M\geq 1}$ and $\{U'_M\}_{M\geq 2}$ are a reverse martingale conditional on $\mathcal{D}_n$ with respect to some filtration. This combined with Theorem 3 (maximal inequality for reverse martingales) in Section 3.4.1 of Lee (1990) (for $r = 1$) yields

$$\mathbb{P}\left(\sup_{M\geq 1}|U_M| \geq \delta\right) \leq \frac{1}{\delta}\mathbb{E}\left[|U_1|\right] = \frac{1}{\delta}\mathbb{E}\left[\sup_{k\in\mathcal{K}_n}|R(\widetilde{f}_{1,k}; \mathcal{D}_n, \{I_{k,1}\}) - \mathfrak{R}_{1,k}|\right].$$

On the other hand, from Lemma S3.3, the expectations are bounded as

$$\mathbb{E}\left[\sup_{k\in\mathcal{K}_n}\left|R(\widetilde{f}_{1,k};\mathcal{D}_n,\{I_{k,1}\})-\mathfrak{R}_{1,k}\right|\right]=\mathcal{O}(n^{\epsilon}\gamma_{1,n}).$$

It follows that

$$\sup_{M\geq 1}|U_M|=\mathcal{O}_p(n^{\epsilon}\gamma_{1,n}).$$

Analogously, for the second U-statistic we also have

$$\sup_{M\geq 2}|U'_M|=\mathcal{O}_p(n^{\epsilon}\gamma_{2,n}).$$

From (S3.13), it follows that

$$\begin{aligned}&\sup_{M\in\mathbb{N},k\in\mathcal{K}_n}|R(\widetilde{f}_M;\mathcal{D}_n,\{I_{k,\ell}\}_{\ell=1}^M)-\mathfrak{R}_{M,k}|\\&=\sup_{M\geq 1}|U_M|+2\sup_{M\geq 2}|U'_M|\\&=\mathcal{O}_p(n^{\epsilon}(\gamma_{1,n}+\gamma_{2,n})).\end{aligned}\tag{S3.14}$$

This completes the proof. □

S4 Proofs of results in Section 4

S4.1 Proof of Theorem 4.1 (δ-optimality of ECV)

Proof of Theorem 4.1. For simplicity, we denote $R(\widetilde{f}_{M,k};\mathcal{D}_n,\{I_{k,\ell}\}_{\ell=1}^M)$ by $R_{M,k}$ as a function of M and k. We split the proof for the two parts below.

Part (1) Error bound on the estimated risk. From Theorem 3.3, we have

$$\sup_{M\in\mathbb{N},k\in\mathcal{K}_n}|\widehat{R}^{\mathrm{ECV}}_{M,k}-R_{M,k}|=\mathcal{O}_p(\zeta_n),$$

where $\zeta_n=\widehat{\sigma}_n\log n/n^{1/2}+n^{\epsilon}(\gamma_{1,n}+\gamma_{2,n})$ and $\widehat{\sigma}_n=\max_{m,\ell\in[M_0],k\in\mathcal{K}_n}\widehat{\sigma}_{I_{k,\ell}\cup I_{k,m}}$. Then the conditional risk of the ECV-tuned predictor $R_{\widehat{M},\widehat{k}}$ admits

$$|\widehat{R}^{\mathrm{ECV}}_{\widehat{M},\widehat{k}}-R_{\widehat{M},\widehat{k}}|\leq\sup_{M\in\mathbb{N},k\in\mathcal{K}_n}|\widehat{R}^{\mathrm{ECV}}_{M,k}-R_{M,k}|=\mathcal{O}_p(\zeta_n).$$

Part (2) Additive suboptimality. The proof proceeds in two steps.

Step 1: Bounding the difference between $\widehat{R}^{\mathbf{ECV}}_{\infty}(\widetilde{f}_{M_0,\widehat{k}})$ and the oracle-tuned risk. Let

$$(M^*,k^*)\in\operatorname*{arginf}_{M\in\mathbb{N},k\in\mathcal{K}_n}R(\widetilde{f}_{M,k};\mathcal{D}_n,\{I_\ell\}_{\ell=1}^M),$$

which is a tuple of random variables and also functions of n. For any $k \in \mathcal{K}_n$, by the risk decomposition (3.1) we have that

$$\inf_{M \in \mathbb{N}} R_{M,k} = R_{\infty,k}.$$

That is, $M^* = \infty$ is one minimizer for any $k \in \mathcal{K}_n$. Then it follows that

$$\widehat{R}^{\mathrm{ECV}}_{\infty,\widehat{k}} = \inf_{k \in \mathcal{K}_n} R_{\infty,k} + \mathcal{O}_p(\zeta_n) = \inf_{M \in \mathbb{N}, k \in \mathcal{K}_n} R_{M,k} + \mathcal{O}_p(\zeta_n) \tag{S4.1}$$

where the first equality is due to Theorem 3.3.

Step 2: δ optimality. Next, we bound the suboptimality by the triangle inequality:

$$\begin{aligned} &|R_{\widehat{M},\widehat{k}} - \inf_{M \in \mathbb{N}, k \in \mathcal{K}_n} R_{M,k}| \\ &= |R_{\widehat{M},\widehat{k}} - \widehat{R}^{\mathrm{ECV}}_{\widehat{M},\widehat{k}} + \widehat{R}^{\mathrm{ECV}}_{\widehat{M},\widehat{k}} - \widehat{R}^{\mathrm{ECV}}_{\infty,\widehat{k}} + \widehat{R}^{\mathrm{ECV}}_{\infty,\widehat{k}} - R_{\infty,k^*}| \\ &\leq |R_{\widehat{M},\widehat{k}} - \widehat{R}^{\mathrm{ECV}}_{\widehat{M},\widehat{k}}| + |\widehat{R}^{\mathrm{ECV}}_{\widehat{M},\widehat{k}} - \widehat{R}^{\mathrm{ECV}}_{\infty,\widehat{k}}| + |\widehat{R}^{\mathrm{ECV}}_{\infty,\widehat{k}} - R_{\infty,k^*}|. \end{aligned} \tag{S4.2}$$

From Part (1) we know that the first term in (S4.2) can be bounded as $|R_{\widehat{M},\widehat{k}} - \widehat{R}^{\mathrm{ECV}}_{\widehat{M},\widehat{k}}| = \mathcal{O}_p(\zeta_n)$. From (S4.1) we know that the last term in (S4.2) can also be bounded as $|\widehat{R}^{\mathrm{ECV}}_{\infty,\widehat{k}} - R_{\infty,k^*}| = \mathcal{O}_p(\zeta_n)$. It remains to bound the second term in (S4.2) . By the definition of $\widehat{R}^{\mathrm{ECV}}_{M,k}$ in (3.4), we have that

$$\widehat{R}^{\mathrm{ECV}}_{\infty,k} = 2\widehat{R}^{\mathrm{ECV}}_{2,k} - \widehat{R}^{\mathrm{ECV}}_{1,k}.$$

Since $\widehat{M} = \lceil 2/\delta \cdot \widehat{R}^{\mathrm{ECV}}_{1,\widehat{k}} - \widehat{R}^{\mathrm{ECV}}_{2,\widehat{k}} \rceil \geq 2/\delta \cdot \widehat{R}^{\mathrm{ECV}}_{1,\widehat{k}} - \widehat{R}^{\mathrm{ECV}}_{2,\widehat{k}}$, the second term in (S4.2) is bounded by

$$|\widehat{R}^{\mathrm{ECV}}_{\widehat{M},\widehat{k}} - \widehat{R}^{\mathrm{ECV}}_{\infty,\widehat{k}}| = \frac{2}{\widehat{M}} \left| \widehat{R}^{\mathrm{ECV}}_{1,\widehat{k}} - \widehat{R}^{\mathrm{ECV}}_{2,\widehat{k}} \right| \leq \delta. \tag{S4.3}$$

Therefore, the δ-optimality conclusion on $\widehat{R}^{\mathrm{ECV}}_{M,k}$ follows. □

S4.2 Proof of multiplicative optimality

Proposition S4.1 (Multiplicative optimality)**.** *Under the same conditions in Theorem 4.1, if $\int \{y - \mathbb{E}(y \mid \boldsymbol{x})\}^2 \, \mathrm{d}P(\boldsymbol{x}, y)$ is lower bounded away from zero and $(\widehat{M}, \widehat{k})$ is defined for relative optimality such that*

$$\widehat{R}^{\mathrm{ECV}}_{\widehat{M},\widehat{k}} \leq (1+\delta) \inf_{M \in \mathbb{N}, k \in \mathcal{K}_n} \widehat{R}^{\mathrm{ECV}}_{M,k}, \tag{S4.4}$$

then it holds that

$$R_{\widehat{M},\widehat{k}} \leq (1+\delta) \inf_{M \in \mathbb{N}, k \in \mathcal{K}_n} R_{M,k}(1 + \mathcal{O}_p(\zeta_n)).$$

Proof of Proposition S4.1. By definition of (S4.4), we have

$$\begin{aligned}\widehat{R}^{\mathrm{ECV}}_{\widehat{M},\widehat{k}} &= (1+\delta)\inf_{M\in\mathbb{N},k\in\mathcal{K}_n}\widehat{R}^{\mathrm{ECV}}_{M,k}\\ &\le (1+\delta)\left(\inf_{M\in\mathbb{N},k\in\mathcal{K}_n} R_{M,k}+\mathcal{O}_p(\zeta_n)\right)\\ &= (1+\delta)\inf_{M\in\mathbb{N},k\in\mathcal{K}_n} R_{M,k}(1+\mathcal{O}_p(\zeta_n)).\end{aligned}$$

where the inequality is from Theorem 4.1 and the last equality is from the assumption that the risks are lower bounded. Further, since from Theorem 3.3, $\sup_{M\in\mathbb{N},k\in\mathcal{K}_n}|R_{M,k}-\widehat{R}^{\mathrm{ECV}}_{M,k}| = \mathcal{O}_p(\zeta_n)$, we have

$$R_{\widehat{M},\widehat{k}} \le \widehat{R}^{\mathrm{ECV}}_{\widehat{M},\widehat{k}}+\mathcal{O}_p(\zeta_n) \le (1+\delta)\inf_{M\in\mathbb{N},k\in\mathcal{K}_n} R_{M,k}(1+\mathcal{O}_p(\zeta_n)).$$

This finishes the proof. □

S4.3 Concrete example: bagged ridge predictors

Application of Theorem 4.1 to a specific data model and a base predictor requires verification of Assumptions 1 and 2. As an illustration, we verify all the conditions for ridge predictors. Consider a dataset $\mathcal{D}_n = \{(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_n, y_n)\}$ 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. Recall that the *ridge* estimator with regularization parameter $\lambda > 0$ fitted on $\mathcal{D}_I$ for $I \subseteq [n]$ is defined as $\widehat{\boldsymbol{\beta}}_\lambda(\mathcal{D}_I) = \operatorname{argmin}_{\boldsymbol{\beta}\in\mathbb{R}^p} \sum_{j\in I}(y_j - \boldsymbol{x}_j^\top\boldsymbol{\beta})^2/|I| + \lambda\|\boldsymbol{\beta}\|_2^2$. The associated ridge base predictor and the subagged predictor are given by $\widehat{f}_\lambda(\boldsymbol{x};\mathcal{D}_I) = \boldsymbol{x}^\top\widehat{\boldsymbol{\beta}}_\lambda(\mathcal{D}_I)$ and $\widetilde{f}_{M,k}(\boldsymbol{x};\mathcal{D}_n) = \boldsymbol{x}^\top\widetilde{\boldsymbol{\beta}}_{\lambda,M}(\{\mathcal{D}_{I_\ell}\}_{\ell=1}^M)$, where $I \in \mathcal{I}_k$ and $\{I_\ell\}_{\ell=1}^M \overset{\mathrm{SRS}}{\sim} \mathcal{I}_k$.

We consider Assumptions 3-4 on the dataset $\mathcal{D}_n$ to characterize the risk, which are standard in the study of the ridge and ridgeless regression under proportional asymptotics; see, e.g., Hastie et al. (2022); Patil et al. (2023).

Assumption 3 (Feature model)**.** The feature vectors $\boldsymbol{x}_i \in \mathbb{R}^p$, $i = 1, \ldots, n$, 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 semi-definite 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 kth moment for $k \ge 2$. Let $\boldsymbol{\Sigma} = \boldsymbol{W}\boldsymbol{R}\boldsymbol{W}^\top$ denote the eigenvalue decomposition of the covariance matrix $\boldsymbol{\Sigma}$, where $\boldsymbol{R} \in \mathbb{R}^{p\times p}$ is a diagonal matrix containing eigenvalues (in non-increasing order) $r_1 \ge r_2 \ge \cdots \ge r_p \ge 0$, 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 $\boldsymbol{\Sigma}$ (supposed on $\mathbb{R}_{>0}$) whose value at any $r \in \mathbb{R}$ is given by

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

Assume there exists $0 < r_{\min} \le r_{\max} < \infty$ such that $r_{\min} \le r_1 \le r_p \le r_{\max}$, and there exists a fixed distribution H such that $H_p \to H$ in distribution as $p \to \infty$.

Assumption 4 (Response model)**.** The response variables $y_i \in \mathbb{R}$, $i = 1, \ldots, n$, additively decompose as $y_i = \boldsymbol{x}_i^\top \boldsymbol{\beta}_0 + \varepsilon_i$, 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 $\delta > 0$. The ℓ_2-norm of the signal vector $\|\boldsymbol{\beta}_0\|_2$ is uniformly bounded in p, and $\lim_{p\to\infty} \|\boldsymbol{\beta}_0\|_2^2 = \rho^2 < \infty$. 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 $\boldsymbol{\Sigma}$ via the distribution of (squared) projection of $\boldsymbol{\beta}_0$ along the eigenvectors $\boldsymbol{w}_j, 1 \le j \le p$, whose value at any $r \in \mathbb{R}$ is given by

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

Assume there exists a fixed distribution G such that $G_p \to G$ in distribution as $p \to \infty$.

Assumptions 3-4 provide risk characterization for subagged ridge predictors (Patil et al., 2023), which establishes the consistency of sample-split CV for any fixed ensemble size M, without obtaining the convergence rate. In Proposition S4.2, Assumptions 3-4 assume a linear model $y_i = \boldsymbol{x}_i^\top \boldsymbol{\beta}_0 + \epsilon_i$ for observation i, where the feature vector is generated by $\boldsymbol{x}_i = \boldsymbol{\Sigma}^{1/2} \boldsymbol{z}_i$, $\boldsymbol{\Sigma}$ is the covariance matrix and $\boldsymbol{z}_i$ contains i.i.d. entries with zero mean and unit variance. We make mild assumptions about such a data-generating process: (1) the noise ϵ_i and the entries of $\boldsymbol{z}_i$ have bounded moments, and (2) the covariance and signal-weighted spectrums converge weakly to some distributions as $n, p \to \infty$.

Under Assumptions 3-4, we will show that $\widehat{\sigma}_n = o(1)$ when using with MOM. To prove the result for MEAN, we need the modified assumptions by replacing the bounded moment conditions on z_{ij} in Assumption 3 and ϵ_i in Assumption 3 by bounded ψ_2-norm conditions.

We will analyze the bagged predictors (with M bags) in the proportional asymptotics regime, where 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 $k \le n$, ϕ_s is always no less than ϕ.

Under these assumptions, the results for ridge predictors are summarized in Proposition S4.2.

Proposition S4.2 (ECV for ridge predictors)**.** *Suppose Assumptions 3-4 in Appendix S4.3 hold. Then, the ridge predictors with $\lambda > 0$ using subagging satisfy Assumption 1 with $\widehat{\sigma}_n = \mathcal{O}_p(1)$ for* MOM *and Assumption 2 for any $\psi \in (0, 1)$ such that $p/n \to \phi \in [\psi, \psi^{-1}]$ and $p/k \to \phi_s \in [\phi, \psi^{-1}]$ as $k, n, p \to \infty$. Consequently, the conclusions in Theorem 4.1 hold.*

We remark that Proposition S4.2 verifies Theorem 4.1 for EST =MOM, but one can also verify for EST =MEAN under slightly different assumptions; see Appendix S4.3 for more details. It is worth mentioning that Assumption 2 only requires the conditional risks for the ridge ensemble with $M = 1, 2$ converge to their respective conditional (on $\mathcal{D}_n$) limits, while the limiting forms of them are not required. In this regard, Assumptions 3-4 can be further relaxed with some efforts, but we do not pursue fine-tuning of assumptions as our intent is only to illustrate the end-to-end applicability of Theorem 4.1. The generality of Theorem 4.1 to general predictors is illustrated empirically in the next section.

Proof of Proposition S4.2. We split the proof into different parts.
Part (1) Bounded moments of the risk for $M = 1$. For $I \in \mathcal{I}_k$, define $\boldsymbol{L}_I$ be the diagnoal matrix whose ith diagonal entry is one if $i \in I$ and zero otherwise, let $\widehat{\boldsymbol{\Sigma}}_I = \boldsymbol{X}^\top \boldsymbol{L}_I \boldsymbol{X}/|I|$. We begin with analyzing the risk for $M = 1$:

$$\begin{aligned} & R(\widetilde{f}_{1,k}; \mathcal{D}_I) \\ &= \mathbb{E}\{(y_0 - \boldsymbol{x}_0^\top \boldsymbol{\beta}_0)^2\} + \|\widehat{\boldsymbol{\beta}}_1(\mathcal{D}_I) - \boldsymbol{\beta}_0\|^2_{\boldsymbol{\Sigma}^{1/2}} \\ &= \mathbb{E}\{(y_0 - \boldsymbol{x}_0^\top \boldsymbol{\beta}_0)^2\} + \boldsymbol{\beta}_0^\top \widehat{\boldsymbol{\Sigma}}_I (\widehat{\boldsymbol{\Sigma}}_I + \lambda \boldsymbol{I}_p)^{-2} \widehat{\boldsymbol{\Sigma}}_I \boldsymbol{\beta}_0 + \frac{1}{k^2} \boldsymbol{\varepsilon}^\top \boldsymbol{L}_I \boldsymbol{X} (\widehat{\boldsymbol{\Sigma}}_I + \lambda \boldsymbol{I}_p)^{-2} \boldsymbol{X}^\top \boldsymbol{L}_I \boldsymbol{\varepsilon}. \end{aligned} \tag{S4.5}$$

Note that the first term is just a constant, and the last two terms can be bounded as:

$$\begin{aligned} \boldsymbol{\beta}_0^\top \widehat{\boldsymbol{\Sigma}}_I (\widehat{\boldsymbol{\Sigma}}_I + \lambda \boldsymbol{I}_p)^{-2} \widehat{\boldsymbol{\Sigma}}_I \boldsymbol{\beta}_0 &\le \|\widehat{\boldsymbol{\Sigma}}_I (\widehat{\boldsymbol{\Sigma}}_I + \lambda \boldsymbol{I}_p)^{-2} \widehat{\boldsymbol{\Sigma}}_I\|_{\mathrm{op}} \|\boldsymbol{\beta}_0\|_2^2 \\ &\le \rho^2 \max_{j \in [p]} \frac{s_j(\widehat{\boldsymbol{\Sigma}}_I)^2}{(\lambda + s_j(\widehat{\boldsymbol{\Sigma}}_I))^2} \\ &\le \rho^2 \\ \frac{1}{k^2} \boldsymbol{\varepsilon}^\top \boldsymbol{L}_I \boldsymbol{X} (\widehat{\boldsymbol{\Sigma}}_I + \lambda \boldsymbol{I}_p)^{-2} \boldsymbol{X}^\top \boldsymbol{L}_I \boldsymbol{\varepsilon} &\le \frac{1}{k^2} \|\boldsymbol{L}_I \boldsymbol{X} (\widehat{\boldsymbol{\Sigma}}_I + \lambda \boldsymbol{I}_p)^{-2} \boldsymbol{X}^\top \boldsymbol{L}_I\|_{\mathrm{op}} \|\boldsymbol{L}_I \boldsymbol{\varepsilon}\|_2^2 \\ &\le \frac{\|\boldsymbol{L}_I \boldsymbol{\varepsilon}\|_2^2}{k} \max_{j \in [p]} \frac{s_j(\widehat{\boldsymbol{\Sigma}}_I)}{(\lambda + s_j(\widehat{\boldsymbol{\Sigma}}_I))^2} \\ &\le \frac{\|\boldsymbol{L}_I \boldsymbol{\varepsilon}\|_2^2}{k\lambda}, \end{aligned}$$

where $s_j(\widehat{\boldsymbol{\Sigma}}_I) \ge 0$ is the jth eigenvalue of $\widehat{\boldsymbol{\Sigma}}_I$. For all $q \le 2 + \delta/2$, it follows that

$$\begin{aligned} \|R(\widetilde{f}_{1,k}; \mathcal{D}_I)\|_{L_q} &\le \mathbb{E}\{(y_0 - \boldsymbol{x}_0^\top \boldsymbol{\beta}_0)^2\} + \rho^2 + \frac{1}{k\lambda} \|\|\boldsymbol{L}_I \boldsymbol{\varepsilon}\|_2^2\|_{L_q} \\ &\le \mathbb{E}\{(y_0 - \boldsymbol{x}_0^\top \boldsymbol{\beta}_0)^2\} + \rho^2 + \frac{1}{\lambda} \max_{j \in [n]} \|\epsilon_j^2\|_{L_q}, \end{aligned}$$

which we denote as $C_{0,q}$. From the bounded moment assumption 4, we know that $\|\epsilon_j^2\|_{L_q} < \infty$ for all j. Thus, $\|R(\widetilde{f}_{1,k}; \mathcal{D}_I)\|_{L_q}$ is upper bounded by constant $C_{0,q}$ for all $n \in \mathbb{N}$.
Part (2) Uniform integrability of the risk for $M = 1$. Under Assumptions 3-4, from Patil et al. (2023, Theorem 4.1) we have that there exist deterministic functions $\mathfrak{R}_1$ and $\mathfrak{R}_2$ such that, for all $I \in \mathcal{I}_k$ and $\{I_1, I_2\} \overset{\text{SRS}}{\sim} \mathcal{I}_k$,

$$R(\widetilde{f}_{1,k}; \mathcal{D}_I) - \mathfrak{R}_1(\phi, \phi_s) \to 0, \qquad R(\widetilde{f}_{2,k}; \mathcal{D}_n, \{I_1, I_2\}) - \mathfrak{R}_2(\phi, \phi_s) \to 0, \tag{S4.6}$$

with probability tending to one, as $k, n, p \to \infty$, $p/n \to \phi \in [\psi, \psi^{-1}]$ and $p/k \to \phi_s \in [\phi, \psi^{-1}]$. Furthermore, $\mathfrak{R}_1$ and $\mathfrak{R}_2$ are continuous functions on ϕ_s for any ϕ, which are bounded in the domain. To verify the tail bound condition for $M = 1$, Hastie et al. (2022, Theorem 5) shows that

$$\mathbb{P}(n^{(1-\epsilon')/2} |R(\widetilde{f}_{1,k}; \mathcal{D}_I) - \mathfrak{R}_1(\phi, \phi_s)| \ge C_1) \le C n^{-D},$$

where $C_1 = C(\rho^2 + \lambda^{-1})\lambda^{-1}$ for any $D, \epsilon' > 0$ and the constant C depends on ϵ', D and other model parameters. For all $q \leq 1 + \delta/4$, let $\gamma_{1,n} = n^{-(1-\epsilon')/(2q)}$ and $Z_n = \gamma_{1,n}^{-1}|R(\widetilde{f}_{1,k}; \mathcal{D}_I) - \mathfrak{R}_1(\phi, \phi_s)|$. Then, we have $\mathbb{P}(Z_n^q \geq C_1^q) \leq 1 - Cn^{-D}$, and

$$\begin{aligned}
\mathbb{E}(Z_n^q) &= \mathbb{E}(Z_n^q \mathbb{1}\{Z_n < C_1\}) + \mathbb{E}(Z_n^q \mathbb{1}\{Z_n \geq C_1\}) \\
&\leq C_1^q + \{\mathbb{E}(Z_n^{2q})\}^{\frac{1}{2}} \mathbb{P}(Z_n \geq C_1)^{\frac{1}{2}} \\
&\leq C_1^q + (C_{0,2q}^q + \mathfrak{R}_1(\phi, \phi_s)^{\frac{q}{2}}) \gamma_{1,n}^{-q} 2^{q-1} C^{\frac{1}{2}} n^{-\frac{D}{2}},
\end{aligned}$$

where the first inequality is from Holder's inequality. For $\epsilon' \in (0, 1)$ fixed, setting $D = (1 - \epsilon')/2$ yields that

$$\mathbb{E}(Z_n^q) \leq C_1^q + (C_{0,2q}^q + \mathfrak{R}_1(\phi, \phi_s)^{\frac{q}{2}}) 2^{q-1} C^{\frac{1}{2}},$$

Therefore, Z_n^q is uniformly integrable. By Chebyshev's inequality, we further have

$$\limsup_{n \to \infty} \sup_{\eta > 0} \eta^q \mathbb{P}(Z_n \geq \eta) \leq C_1^q + (C_{0,2q}^q + \mathfrak{R}_1(\phi, \phi_s)^{\frac{q}{2}}) 2^{q-1} C^{\frac{1}{2}}, \tag{S4.7}$$

which implies Assumption 2 for $\gamma_{1,n} = n^{-(1-\epsilon')/2}$ and for any $\epsilon = 1/q$ where $q \leq 1 + \delta/4$.
Part (3) Concentration of the risk for $M = 2$. To show the existence of $\gamma_{2,n}$, from Patil et al. (2023, Lemma 3.8.), we have that with probability tending to one,

$$R(\widetilde{f}_{2,k}; \mathcal{D}_n, \{I_1, I_2\}) - \mathfrak{R}_2(\phi, \phi_s) \to 0.$$

By convexity, we have

$$0 \leq R(\widetilde{f}_{2,k}; \mathcal{D}_n, \{I_1, I_2\}) \leq 2^{-1}(R(\widetilde{f}_{1,k}; \mathcal{D}_n, \{I_1\}) + R(\widetilde{f}_{1,k}; \mathcal{D}_n, \{I_2\})),$$

which implies that

$$0 \leq R(\widetilde{f}_{2,k}; \mathcal{D}_n, \{I_1, I_2\})^{1/\epsilon} \leq 2^{-1/\epsilon}(R(\widetilde{f}_{1,k}; \mathcal{D}_n, \{I_1\}) + R(\widetilde{f}_{1,k}; \mathcal{D}_n, \{I_2\}))^{1/\epsilon}. \tag{S4.8}$$

We next apply Pratt's lemma to show $L^{1/\epsilon}$-convergence for $M = 2$. Note that the tail bound condition (S4.7) implies $L^{1/\epsilon}$-convergence:

$$\gamma_{1,n}^{-1/\epsilon} \mathbb{E}\{(R(\widetilde{f}_{1,k}; \mathcal{D}_n, \{I_j\}) - \mathfrak{R}_1(\phi, \phi_s))^{1/\epsilon}\} \to 0, \qquad j = 1, 2$$

and the $1/\epsilon$-th moment converges:

$$\gamma_{1,n}^{-1/\epsilon} \mathbb{E}(R(\widetilde{f}_{1,k}; \mathcal{D}_n, \{I_j\})^{1/\epsilon} - \mathfrak{R}_1(\phi, \phi_s)^{1/\epsilon}) \to 0, \qquad j = 1, 2.$$

By Pratt's lemma (see, e.g., Gut, 2005, Theorem 5.5), we have the $1/\epsilon$-th moment for $M = 2$ also converges:

$$\gamma_{1,n}^{-1/\epsilon} \mathbb{E}(R(\widetilde{f}_{2,k}; \mathcal{D}_n, \{I_1, I_2\})^{1/\epsilon} - \mathfrak{R}_2(\phi, \phi_s)^{1/\epsilon}) \to 0.$$

From Gut (2005, Theorem 5.2), we further have $L^{1/\epsilon}$-convergence for $M = 2$:

$$\gamma_{1,n}^{-1/\epsilon} \mathbb{E}\{(R(\widetilde{f}_{2,k}; \mathcal{D}_n, \{I_1, I_2\}) - \mathfrak{R}_2(\phi, \phi_s))^{1/\epsilon}\} \to 0.$$

This implies that there exists a constant sequence of $\gamma'_{2,n}$ such that $\mathbb{E}\{(R(\widetilde{f}_{2,k}; \mathcal{D}_n, \{I_1, I_2\}) - \mathfrak{R}_2(\phi, \phi_s))^{1/\epsilon}\} \leq \gamma'_{2,n}$ and $\gamma'_{2,n} \geq \gamma_{1,n}^{1/\epsilon}$. Hence, we can simply pick $\gamma'_{2,n} = \gamma_{1,n}^{1/\epsilon}$. Then, by Markov's inequality, we have that for all $\eta' > 0$

$$\mathbb{P}(|R(\widetilde{f}_{2,k}; \mathcal{D}_n, \{I_1, I_2\}) - \mathfrak{R}_2(\phi, \phi_s)|^{1/\epsilon} > \eta') \leq \gamma'_{2,n}/\eta',$$

or equivalently for all $\eta > 1$,

$$\mathbb{P}(|R(\widetilde{f}_{2,k}; \mathcal{D}_n, \{I_1, I_2\}) - \mathfrak{R}_2(\phi, \phi_s)| > \eta\gamma_{2,n}) \leq 1/\eta^{1/\epsilon},$$

where $\gamma_{2,n} = \gamma'^{\epsilon}_{2,n}$ and $\gamma_{2,n} = o(n^{-\epsilon})$. Thus, Assumption 2 is satisfied under proportional asymptotics.

Part (4) Bounded variance proxy for CV estimates. From results by Patil et al. (2022, Remark 2.19), Assumption 3 in random matrix theory implies $L_4 - L_2$ norm equivalence, since the components of $\boldsymbol{Z}$ are independent and have bounded kurtosis. Invoking Proposition 2.16 of Patil et al. (2022), there exists $\tau > 0$ such that

$$\widehat{\sigma}_I \leq \tau \inf_{\boldsymbol{\beta} \in \mathbb{R}^p} (\|y_0 - \boldsymbol{x}_0^\top \boldsymbol{\beta}\|_{L_2} + \|\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}_\lambda(\mathcal{D}_I)\|_{\boldsymbol{\Sigma}})^2.$$

When $\boldsymbol{\beta} = \boldsymbol{\beta}_0$, $\|y_0 - \boldsymbol{x}_0^\top \boldsymbol{\beta}_0\|_{L_2} = \sigma$ and $\|\boldsymbol{\beta}_0 - \widehat{\boldsymbol{\beta}}_\lambda(\mathcal{D}_I)\|_{\boldsymbol{\Sigma}} \to \mathfrak{R}_1(\phi, \phi_s)$ with probability tending to one, which is continuous and bounded in $\phi \in [\psi, \psi^{-1}]$ and $\phi_s \in [\phi, \psi^{-1}]$. This implies that $\widehat{\sigma}_I = \mathcal{O}_p(1)$. Analogously, $\widehat{\sigma}_{I_1 \cup I_2} = \mathcal{O}_p(1)$ and hence it holds for EST = MOM.

Combining the parts above finishes the proof. □

S5 Helper concentration results

S5.1 Size of the intersection of randomly sampled datasets

In this section, we collect a helper result concerned with convergences that are used in the proofs of Theorem 4.1. Before stating the lemma, we recall the definition of a hypergeometric random variable along with its mean and variance; see Greene and Wellner (2017) for more related details.

Definition S5.1 (Hypergeometric random variable)**.** *A random variable X follows the hypergeometric distribution $X \sim$ Hypergeometric(n, K, N) if its probability mass function is given by*

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

The expectation and variance of X are given by

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

The following lemma characterizes the limiting proportions of shared observations in two simple random samples when both the subsample size k and the full data size n tend to infinity.

Lemma S5.2 (Asymptotic proportions of shared observations, adapted from Patil et al. (2023)). *For $n \in \mathbb{N}$, define $\mathcal{I}_k = \{\{i_1, i_2, \ldots, i_k\} : 1 \le i_1 < i_2 < \ldots < i_k \le n\}$. Let $I_1, I_2 \overset{\texttt{SRSWR}}{\sim} \mathcal{I}_k$, define the random variable $i_0^{\texttt{SRSWR}} = |I_1 \cap I_2|$ to be the number of shared samples, and define $i_0^{\texttt{SRSWOR}}$ accordingly. Then $i_0^{\texttt{SRSWR}} \sim \text{Binomial}(k, k/n)$ and $i_0^{\texttt{SRSWOR}} \sim \text{Hypergeometric}(k, k, n)$. Let $\{k_m\}_{m=1}^{\infty}$ and $\{n_m\}_{m=1}^{\infty}$ be two sequences of positive integers such that n_m is strictly increasing in m, $n_m^{\nu} \le k_m \le n_m$ for some constant $\nu \in (0, 1)$. Then, $i_0^{\texttt{SRSWR}}/k_m - k_m/n_m \to 0$, and $i_0^{\texttt{SRSWOR}}/k_m - k_m/n_m \to 0$ with probability tending to one.*

S6 Additional experimental details and results

S6.1 Risk estimation and extrapolation in Section 5.1

In these experiments, we use $n = 500, p = 5,000$ ($\phi = 0.1$) and $n = 5,000, p = 500$ ($\phi = 10$) for underparameterized and overparameterized regimes, where the subsample aspect ratio ϕ_s varies from 0.1 to 10, and from 10 to 100, respectively. The out-of-sample prediction errors are computed on $n_{\text{te}} = 2,000$ samples, and the results are averaged over 50 dataset repetitions. The ECV cross-validation estimates (3.4) are computed on $M_0 = 10$ base predictors. For the kNN predictor, we use 5 nearest neighbors. For the logistic predictor, we further binarize the response at the median with the null risk of a predictor always outputs 0.5 being 0.25.

Table S1: Summary of experimental results in Section 4.

Predictors	ridgeless	ridge	lassoless	lasso	logistic	kNN
Figure	Figure S3	Figure S2	Figure S5	Figure S4	Figure S6	Figure S7

S6.2 Tuning ensemble and subsample sizes in Section 5.2

In these experiments, we use $n = 1,000$ and $p = \lfloor n\phi \rfloor$ with data aspect ratio ϕ varying from 0.1 to 10. The ECV is performed on a grid of subsample aspect ratios ϕ_s given in Algorithm 1, with $M_0 = 10$ and $M_{\max} = 50$. The out-of-sample prediction errors are computed on $n_{\text{te}} = 2,000$ samples, and the results are averaged over 50 dataset repetitions.

Table S2: Summary of experimental results in Section 5.2.

Procedure	bagging			subagging		
Model	(M1)	(M2)	(M3)	(M1)	(M2)	(M3)
Figure	Figure S8	Figure 3	Figure S9	Figure S10	Figure S11	Figure S12

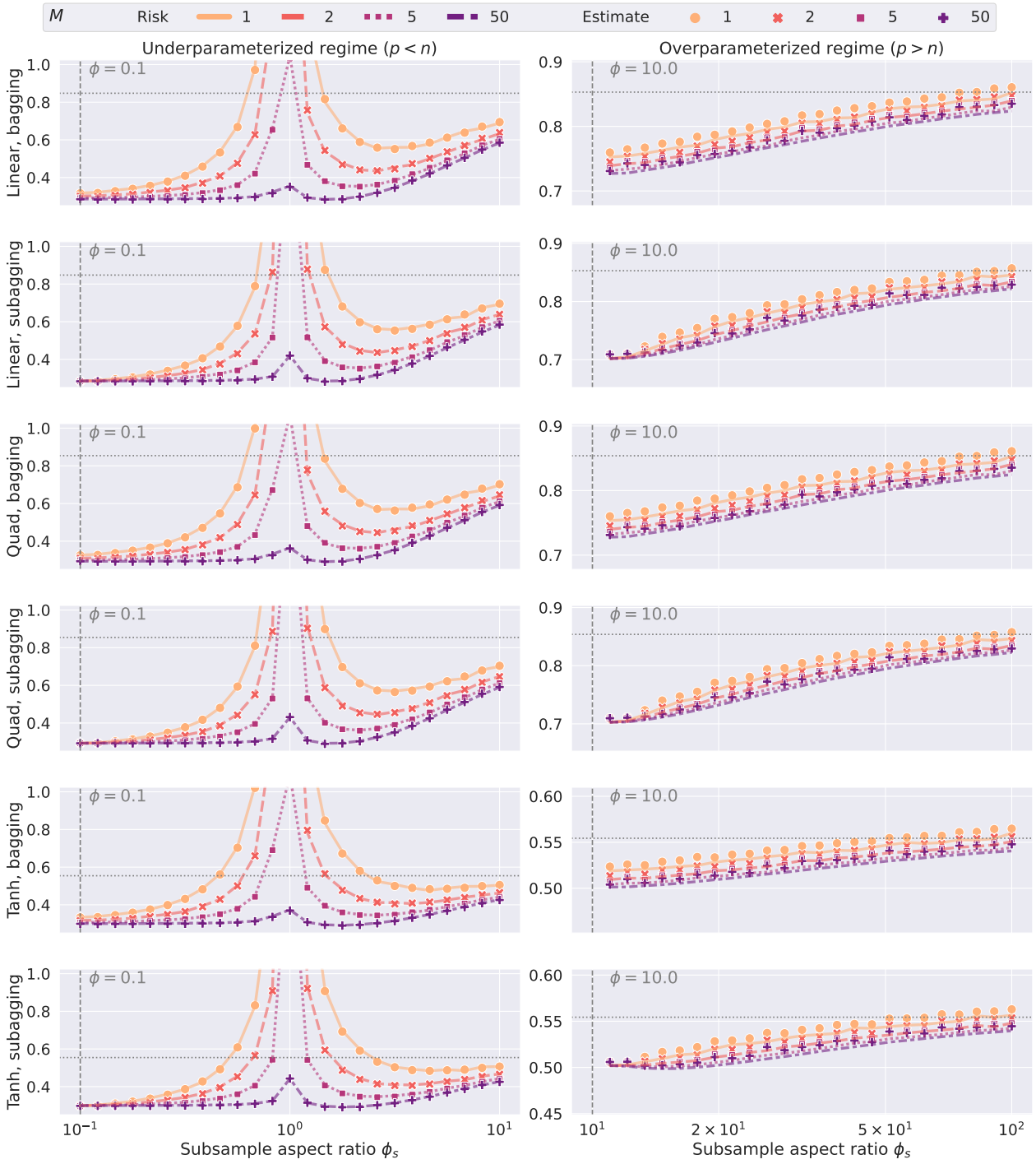

Figure S2: Finite-sample ECV estimate (points) and prediction risk (lines) for ridge predictors ($\lambda = 0.1$) using bagging and subagging, under nonlinear quadratic and tanh models (Section 5.1) for varying subsample size $k = \lfloor p/\phi_s \rfloor$, and the ensemble size M. For each value of M, the points denote the finite-sample ECV cross-validation estimates (3.4), and the lines denote the out-of-sample prediction error. See Appendix S6.1 for more details.

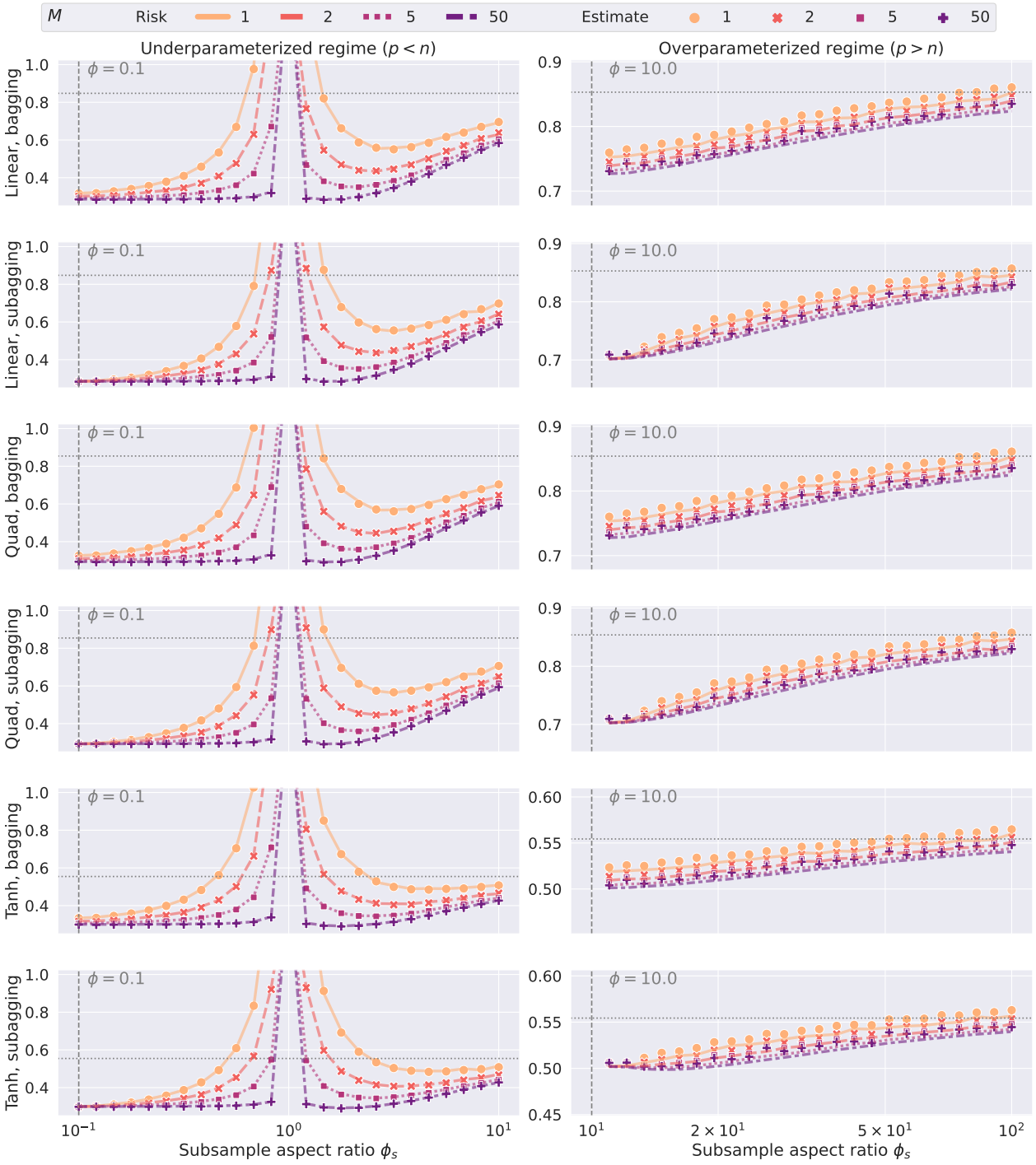

Figure S3: Finite-sample ECV estimate (points) and prediction risk (lines) for ridgeless predictors using bagging and subagging, under nonlinear quadratic and tanh models (Section 5.1) for varying subsample size $k = \lfloor p/\phi_s \rfloor$, and the ensemble size M. For each value of M, the points denote the finite-sample ECV cross-validation estimates (3.4), and the lines denote the out-of-sample prediction error. See Appendix S6.1 for more details.

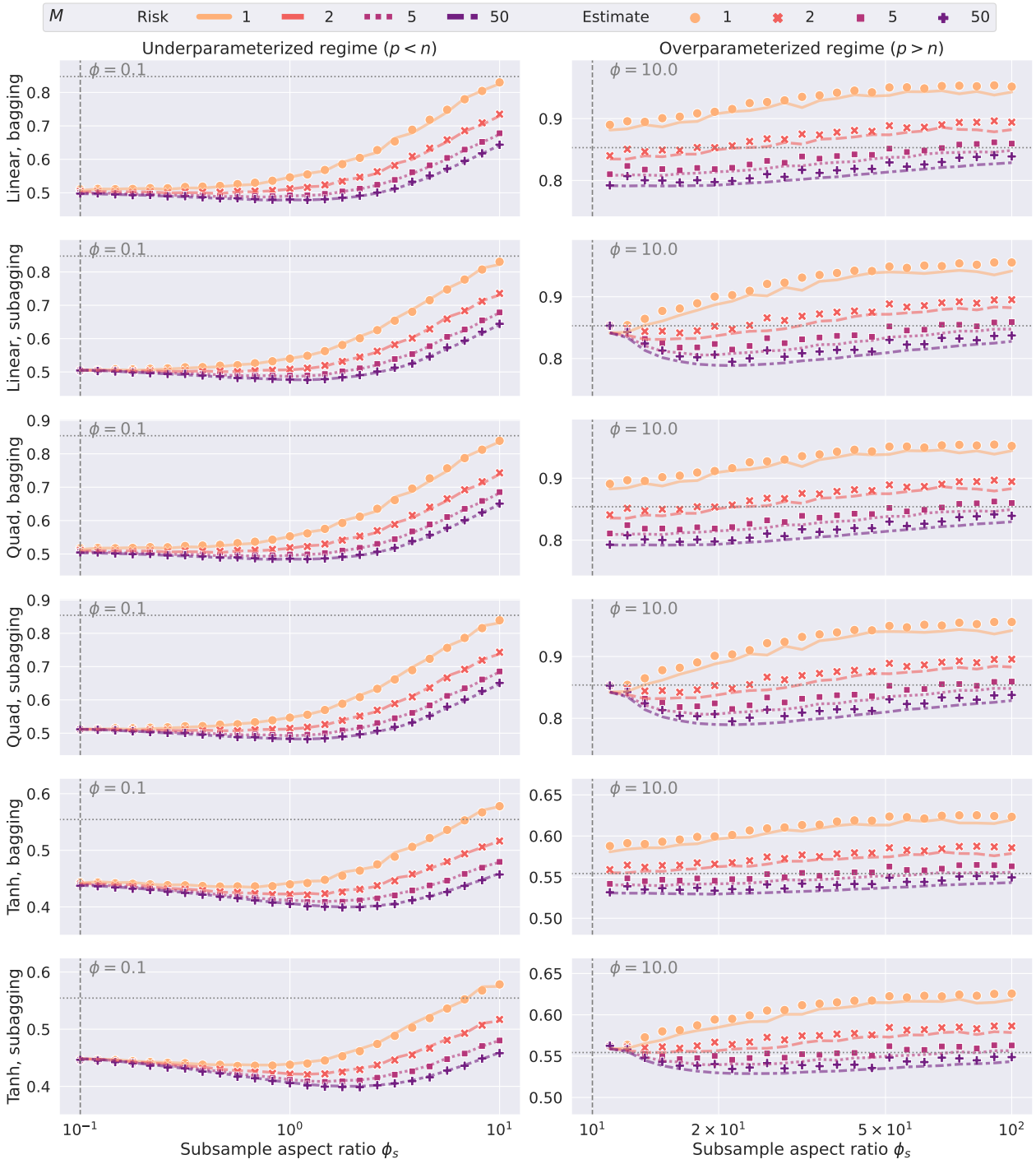

Figure S4: Finite-sample ECV estimate (points) and prediction risk (lines) for lasso predictors ($\lambda = 0.1$) using bagging and subagging, under nonlinear quadratic and tanh models (Section 5.1) for varying subsample size $k = \lfloor p/\phi_s \rfloor$, and the ensemble size M. For each value of M, the points denote the finite-sample ECV cross-validation estimates (3.4), and the lines denote the out-of-sample prediction error. See Appendix S6.1 for more details.

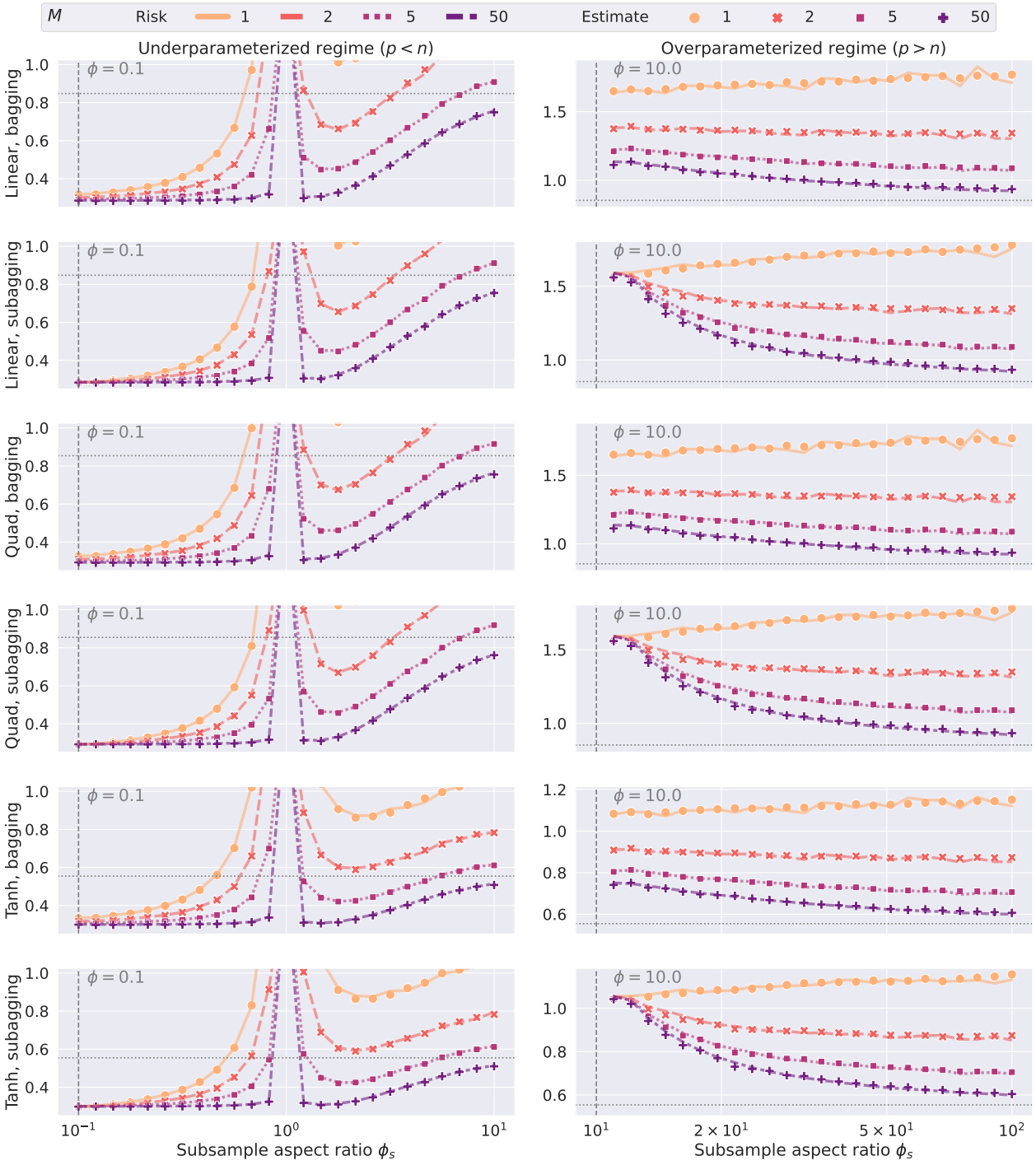

Figure S5: Finite-sample ECV estimate (points) and prediction risk (lines) for lassoless predictors using bagging and subagging, under nonlinear quadratic and tanh models (Section 5.1) for varying subsample size $k = \lfloor p/\phi_s \rfloor$, and the ensemble size M. For each value of M, the points denote the finite-sample ECV cross-validation estimates (3.4) computed on $M_0 = 10$ base predictors, and the lines denote the out-of-sample prediction error computed on $n_{\text{te}} = 2,000$ samples, averaged over 50 dataset repetitions, with $n = 1,000$ and $p = \lfloor n\phi \rfloor$, and $\phi = 0.1$ and 10 for underparameterized ($p < n$) and overparameterized ($p > n$) regimes.

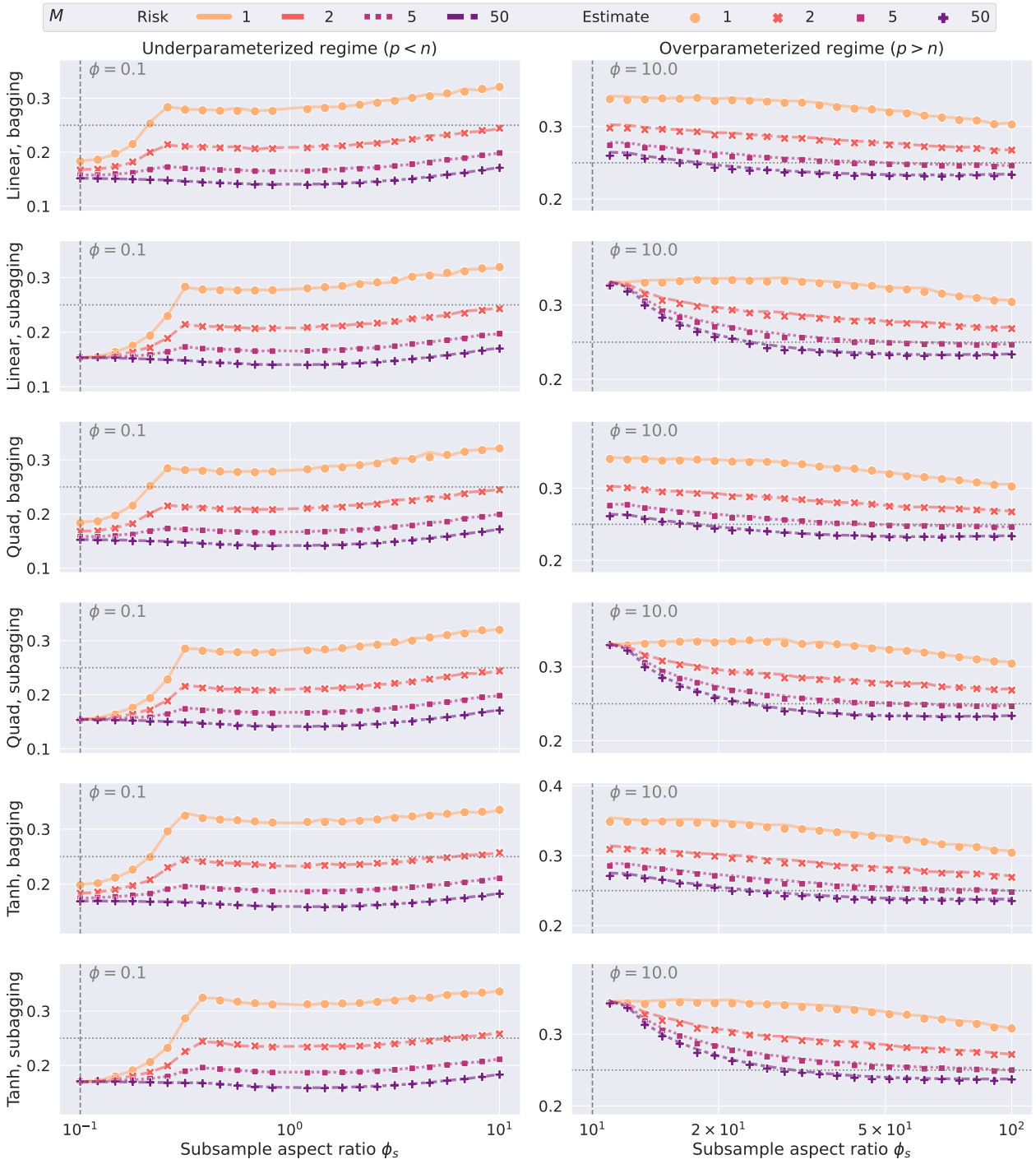

Figure S6: Finite-sample ECV estimate (points) and prediction risk (lines) for logistic predictors using bagging and subagging, under nonlinear quadratic and tanh models (Section 5.1) for varying subsample size $k = \lfloor p/\phi_s \rfloor$, and the ensemble size M. For each value of M, the points denote the finite-sample ECV cross-validation estimates (3.4), and the lines denote the out-of-sample prediction error. See Appendix S6.1 for more details.

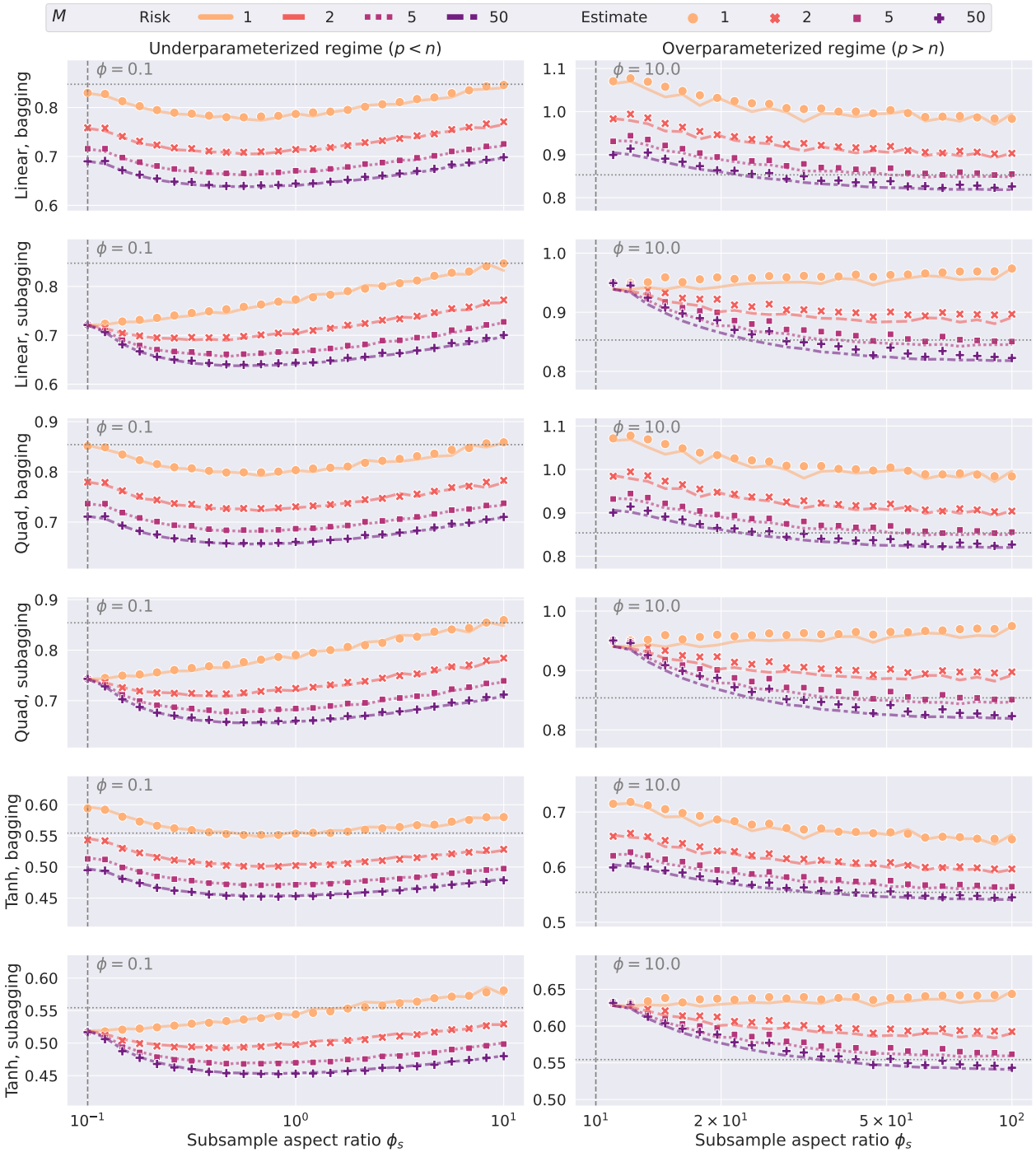

Figure S7: Finite-sample ECV estimate (points) and prediction risk (lines) for kNN predictors using bagging and subagging, under nonlinear quadratic and tanh models (Section 5.1) for varying subsample size $k = \lfloor p/\phi_s \rfloor$, and the ensemble size M. For each value of M, the points denote the finite-sample ECV cross-validation estimates (3.4), and the lines denote the out-of-sample prediction error. See Appendix S6.1 for more details.

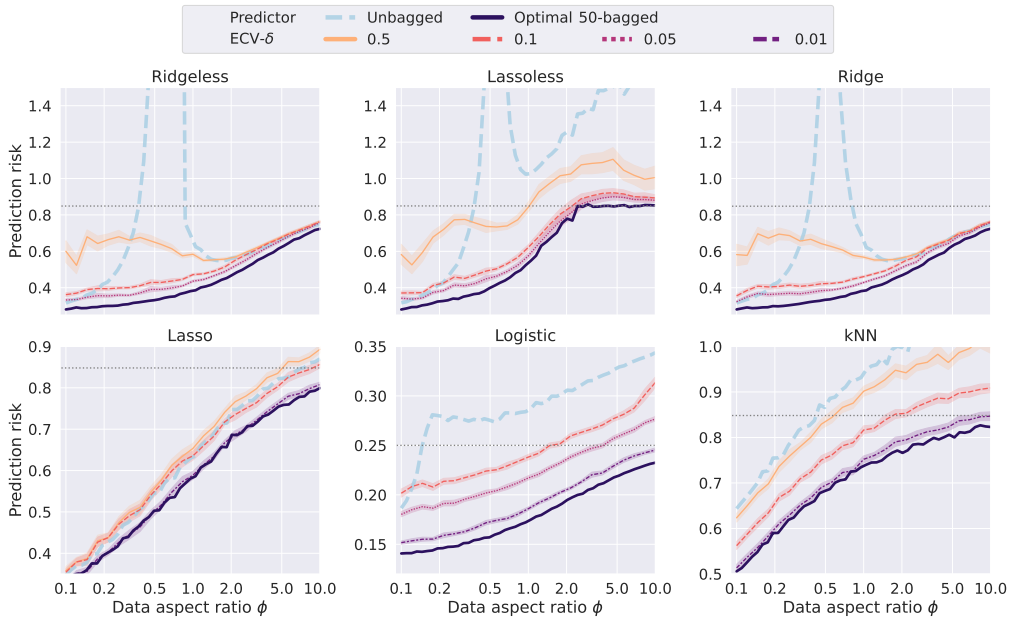

Figure S8: Prediction risk for different bagged predictors by ECV, under model (M1), for varying ϕ and tolerance threshold δ. An ensemble is fitted when (4.3) is satisfied with $\zeta = 5$. The null risks and the risks for the non-ensemble predictors are marked as gray dotted lines and blue dashed lines, respectively. The points denote finite-sample risks, and the shaded regions denote the values within one standard deviation. See Appendix S6.2 for more details.

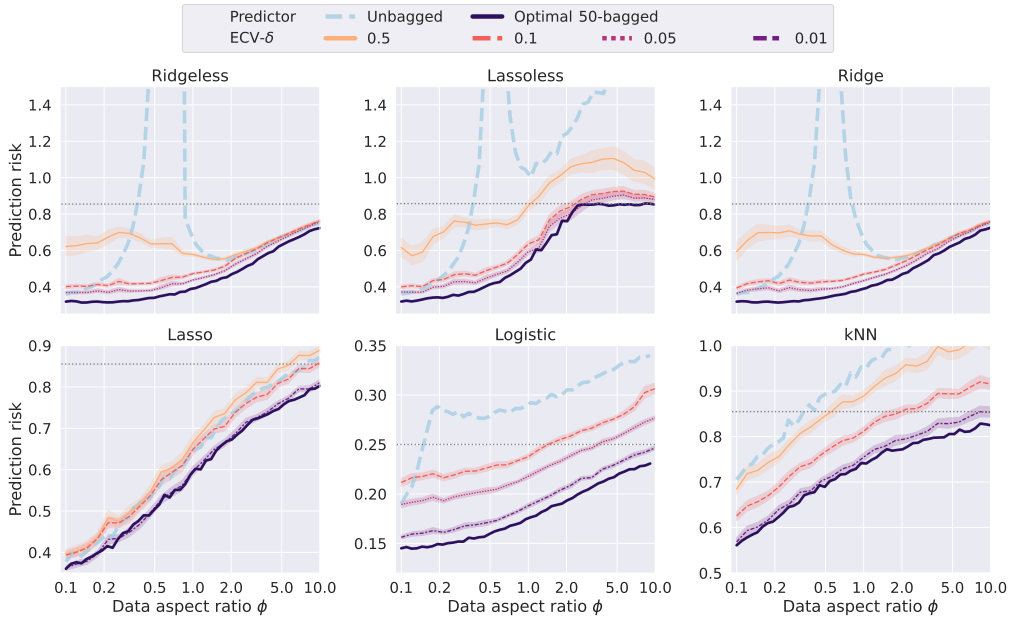

Figure S9: Prediction risk for different bagged predictors by ECV, under model (M3), for varying ϕ and tolerance threshold δ. An ensemble is fitted when (4.3) is satisfied with $\zeta = 5$. The null risks and the risks for the non-ensemble predictors are marked as gray dotted lines and blue dashed lines, respectively. The points denote finite-sample risks, and the shaded regions denote the values within one standard deviation. See Appendix S6.2 for more details.

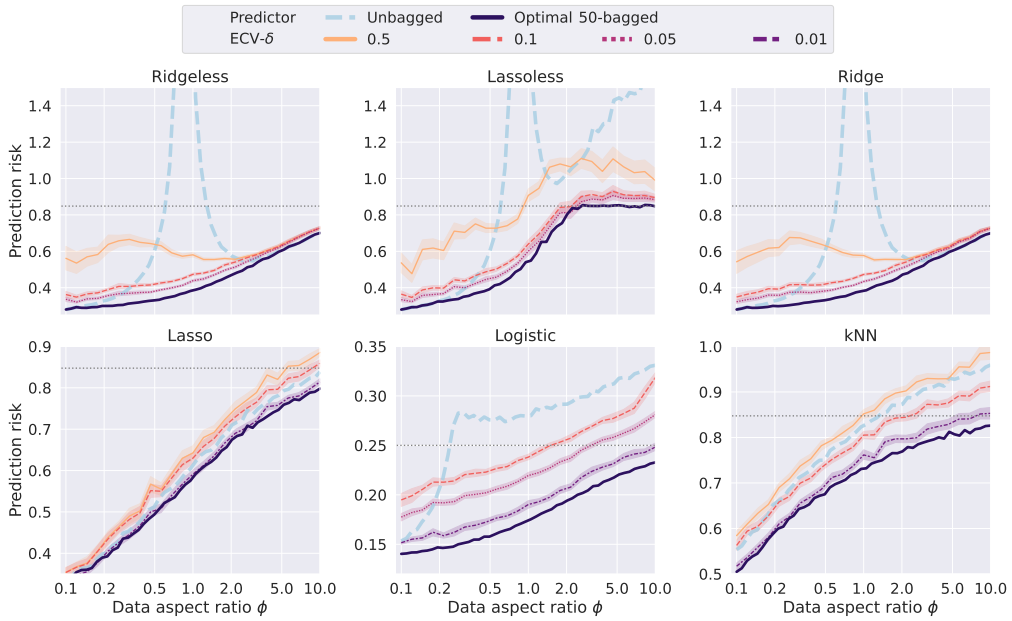

Figure S10: Prediction risk for different subagged predictors by ECV, under model (M1), for varying ϕ and tolerance threshold δ. An ensemble is fitted when (4.3) is satisfied with $\zeta = 5$. The null risks and the risks for the non-ensemble predictors are marked as gray dotted lines and blue dashed lines, respectively. The points denote finite-sample risks, and the shaded regions denote the values within one standard deviation. See Appendix S6.2 for more details.

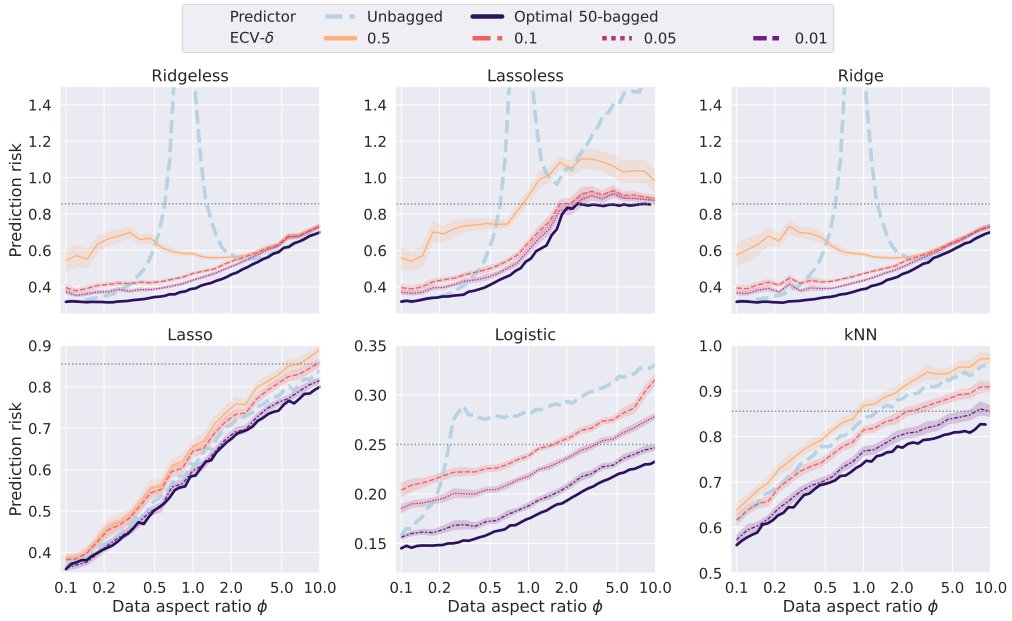

Figure S11: Prediction risk for different subagged predictors by ECV, under model (M2), for varying ϕ and tolerance threshold δ. An ensemble is fitted when (4.3) is satisfied with $\zeta = 5$. The null risks and the risks for the non-ensemble predictors are marked as gray dotted lines and blue dashed lines, respectively. The points denote finite-sample risks, and the shaded regions denote the values within one standard deviation. See Appendix S6.2 for more details.

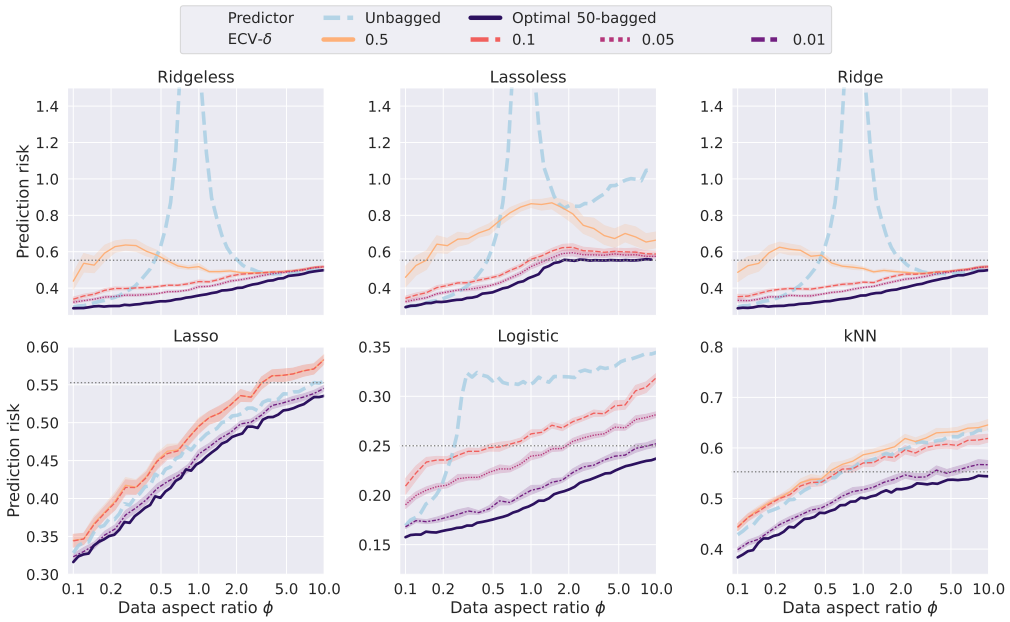

Figure S12: Prediction risk for different subagged predictors by ECV, under model (M3), for varying ϕ and tolerance threshold δ. An ensemble is fitted when (4.3) is satisfied with $\zeta = 5$. The null risks and the risks for the non-ensemble predictors are marked as gray dotted lines and blue dashed lines, respectively. The points denote finite-sample risks, and the shaded regions denote the values within one standard deviation. See Appendix S6.2 for more details.

S6.3 Imbalanced classification

For classification tasks, K-fold cross-validation is recognized to have suboptimal performance in the case of imbalanced data. In this subsection, we evaluate the performance of ECV and K-fold cross-validation under varying degrees of class imbalance. When K is large, K-fold cross-validation also behaves similarly to the leave-one-out cross-validation. However, due to the increased computational complexity, we only examine $K = 3$, 5, and 10. We evaluate these methods using the following two metrics:

- Pointwise prediction error: the absolute error between the risk estimate and the true prediction risk of each ensemble predictor.
- Tuned prediction error: the absolute error between the prediction risk of the tuned ensemble predictor and the one of the optimal ensemble predictors fitted on all training data.

From Figure S13, we can see that when M is small, the pointwise prediction error of K-fold CV increases in both the number of fold K and the class proportion. This indicates that K-fold CV is unstable in the case of unbalanced data. As M increases, the prediction risk gets stabilized and all methods have similar performance. On the other hand, ECV has stable prediction errors across different class proportions and much smaller computational complexity than K-fold CV, as pointed out in Section 4.2.

In terms of tuned prediction errors, when M is small, the optimal risks are obtained at either the smallest subsample size in the overparameterized regime or the largest one in the overparameterized regime, as shown in Figure S6. Thus, it is easier to achieve good predictive performance once the subsample size is close to the endpoint of the grid. As a result, we see that the tuned prediction errors of all methods are smaller than their pointwise prediction errors. However, K-fold CV has increasing prediction error and variability as the class proportion increases, even when M is large.

Overall, the ECV method is more accurate, robust, and efficient than the K-fold CV method for binary classification tasks in the presence of class imbalance.

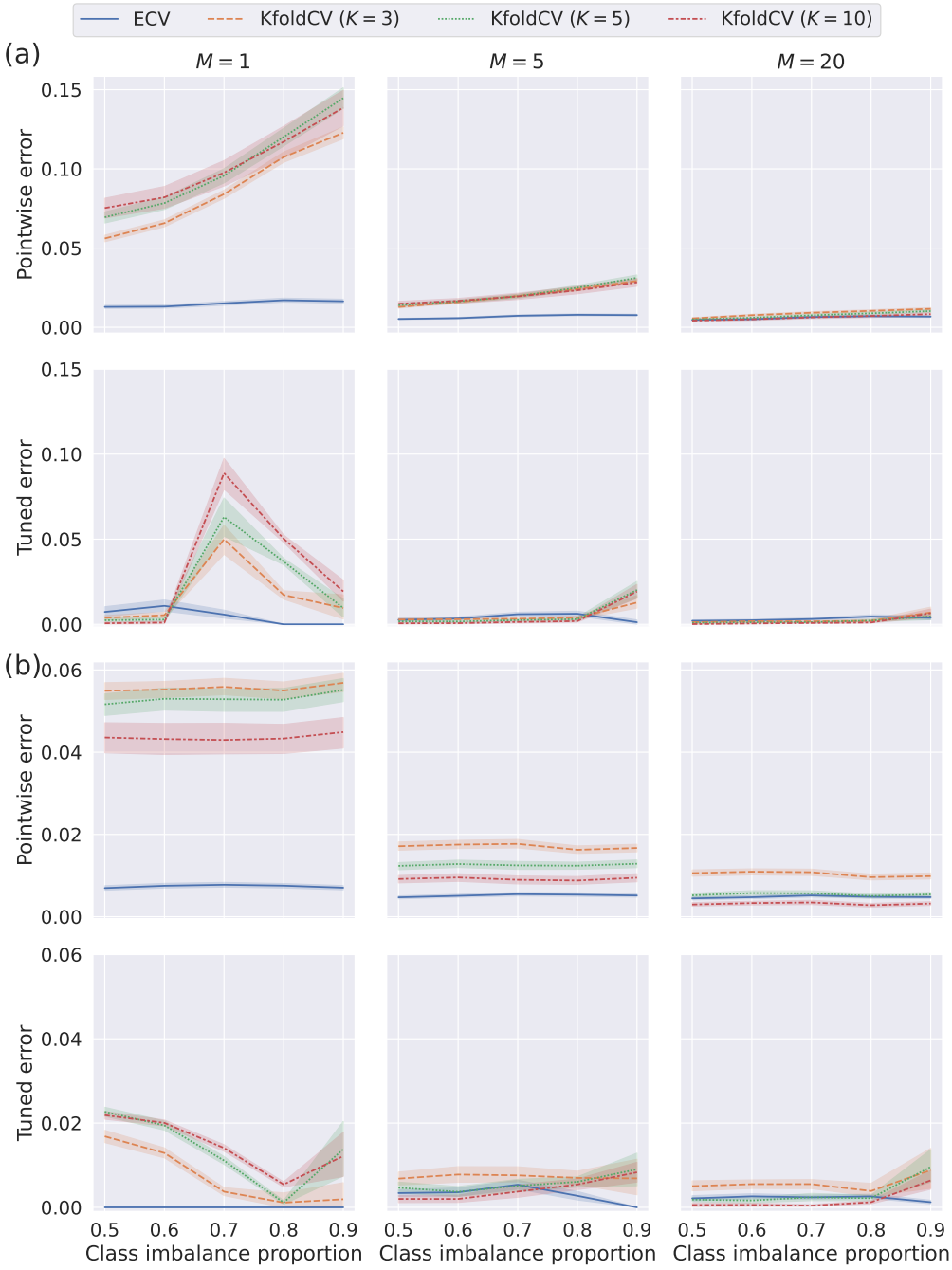

Figure S13: The pointwise and tuned errors for Logistic predictors with varying negative class proportions, in (a) the underparameterized regime ($\phi = 0.1$) and (b) the overparameterized ($\phi = 10$) regimes, with $n = 1,000$ and $p = \lfloor n\phi \rfloor$ across 50 repetitions. The setup is under model (M2) with $\rho_{\text{ar1}} = 0.5$, where the response is further binarized by thresholding at the specified quantile from 0.5 to 0.9 on the training data. ECV uses $M_0 = 10$ to extrapolate the estimates.

S6.4 Risk extrapolation in Section 5.3

S6.4.1 Sensitivity analysis of M_0

To inspect the performance of ECV on different values M_0, the number of base predictors used for estimation, we conduct sensitivity analysis of M_0 in terms of relative pointwise error (defined as the ratio of pointwise error to the null risk) and time complexity. Because the scale of the (absolute) prediction errors may be quite different in the underparameterized and overparameterized regimes, we normalize it by the null risk so that it is more informative when comparing the two scenarios. The results are shown in Table S3. We see that the relative error decreases as M_0 increases initially; however, it gets stable when $M_0 \geq 20$ in both the underparameterized and overparameterized scenarios. As anticipated, the errors are slightly greater in the overparameterized regime, reflecting the inherent difficulty of estimation in this scenario. On the other hand, the time complexity increases in M_0 sublinearly due to the parallel computation. We also observe an increase in the time complexity in the overparameterized regime, which is unavoidable because of fitting deeper random forests. This indicates that one could choose a relatively large M_0, e.g. $M_0 = 20$ to obtain good accuracy with a small computation cost.

One can also determine M_0 adaptively when building the ensembles in an online learning fashion. Since the risk estimate as a function M_0 converges to a certain limit, we can stop increasing M_0 when the risk estimate changes slowly enough. There are various criteria one may consider for early stopping. For instance, comparing the variance of risk estimate to a pre-specified threshold, as in Lopes (2019); Lopes et al. (2020).

ϕ			M_0				
			5	10	15	20	25
Pointwise error	0.1	mean	0.0874	0.0731	0.0668	0.0638	0.0661
		sd	0.0691	0.0570	0.0628	0.0881	0.0696
	10	mean	0.1133	0.1311	0.1156	0.1058	0.0945
		sd	0.1090	0.1272	0.0882	0.0939	0.0812
Tuned error	0.1	mean	0.0002	0.0006	0.0004	0.0004	0.0001
		sd	0.0006	0.0012	0.0011	0.0011	0.0004
	10	mean	0.0002	0.0003	0.0002	0.0002	0.0002
		sd	0.0006	0.0011	0.0007	0.0007	0.0005
Time	0.1	mean	0.0718	0.0870	0.1067	0.1388	0.1728
		sd	0.0295	0.0346	0.0292	0.0381	0.0301
	10	mean	0.9249	1.2543	1.4492	1.9112	1.9652
		sd	0.2702	0.3360	0.3746	0.4383	0.4035

Table S3: The effect of M_0 on the relative prediction errors and computational time for random forests under model (M2) with $n = 500$ across 50 repetitions.

S6.4.2 Sensitivity analysis of ρ_{ar1}

To inspect the impact of ρ_{ar1} related to feature correlation, we conduct sensitivity analysis of ρ_{ar1} and the results of the relative prediction error (defined as the ratio of prediction error to the null risk) are shown in Table S4. When the magnitude of ρ_{ar1} is small, the variability in the pointwise prediction errors of ECV tends to increase. This phenomenon occurs because when ρ_{ar1} approaches zero, the features exhibit almost complete independence. In such cases, random forests struggle to leverage collinearity for mitigating prediction risk. Overall, the relative error lies between 0.05 and 0.15 in various settings.

	ϕ		ρ_{ar1}						
			-0.75	-0.5	-0.25	0	0.25	0.50	0.75
Pointwise error	0.1	mean	0.0663	0.0822	0.1119	0.1158	0.1053	0.0916	0.0552
		sd	0.0471	0.0803	0.0803	0.0948	0.0819	0.0816	0.0714
	10	mean	0.1322	0.1348	0.1524	0.1277	0.1348	0.1143	0.1154
		sd	0.1080	0.0927	0.1145	0.1102	0.1145	0.1107	0.1004
Tuned error	0.1	mean	0.0004	0.0002	0.0007	0.0006	0.0003	0.0003	0.0004
		sd	0.0008	0.0007	0.0018	0.0014	0.0009	0.0008	0.0007
	10	mean	0.0004	0.0004	0.0010	0.0004	0.0005	0.0007	0.0004
		sd	0.0009	0.0009	0.0022	0.0011	0.0013	0.0016	0.0008

Table S4: The effect of ρ_{ar1} on the relative prediction errors for random forests under model (M2) with $n = 500$ and $M_0 = 20$ with varying values of ρ_{ar1} across 50 repetitions.

S6.4.3 Sensitivity analysis of covariance structures

To inspect the performance of ECV on more complex covariance structures, we test it on the following three block-diagonal or graph-based correlation structures under data model (M2):

- A usual AR1 covariance matrix $\mathbf{\Sigma}_0$.
- A block-diagonal covariance matrix with 2 blocks $(\mathbf{\Sigma}_0^{(1)}, \mathbf{\Sigma}_{0.5}^{(2)})$ and $\mathbf{\Sigma}_0^{(1)}, \mathbf{\Sigma}_{0.5}^{(2)} \in \mathbb{R}^{p/2}$.
- A block-diagonal covariance matrix with 3 blocks $(\mathbf{\Sigma}_0^{(1)}, \mathbf{\Sigma}_{0.5}^{(2)}, \mathbf{\Sigma}_{-0.5}^{(3)})$ and $\mathbf{\Sigma}_0^{(1)}, \mathbf{\Sigma}_{0.5}^{(2)}, \mathbf{\Sigma}_{-0.5}^{(3)} \in \mathbb{R}^{p/3}$.

The results of the relative prediction error (defined as the ratio of prediction error to the null risk) are shown in Table S4. In these more complex situations, the relative pointwise prediction errors are also between 0.05 and 0.15, as we observed in Appendix S6.4.2. Thus, we conclude that ECV is robust across different feature correlation structures as tested in the current experiment.

	ϕ		Type of block structure		
			1	2	3
Pointwise error	0.1	mean	0.1158	0.0827	0.0725
		sd	0.0729	0.0513	0.0736
	10	mean	0.1277	0.1287	0.1497
		sd	0.1190	0.0988	0.1003
Tuned error	0.1	mean	0.0006	0.0004	0.0004
		sd	0.0014	0.0009	0.0009
	10	mean	0.0004	0.0005	0.0004
		sd	0.0011	0.0010	0.0011

Table S5: The effect of covariance structure on the relative prediction errors under model (M2) with $n = 500$ and $M_0 = 20$ across 50 repetitions.

S6.4.4 Tuning feature subsampling ratios

In the paper, we describe how to tune the ensemble and subsample sizes. In practice, there are also other hyperparameters that people would like to tune. For instance, the faction of features to consider when looking for the best split of random forests, related to parameter `max_features` in Python package `sckit-learn` (Pedregosa et al., 2011) or parameter `mtry` in R package `randomForest` (Liaw et al., 2002). In practice, there are a few common heuristics for choosing a value for `mtry`. For example, `mtry` can be set as $\sqrt{p}$, $p/3$, or $\log_2(p)$, where p is the total number of features. In the regression setting, `mtry` $= p/3$ is suggested by Breiman (2001), while `mtry` $= p$ was more recently justified empirically in Geurts et al. (2006). These heuristics are a good place to start when determining what value to use for `mtry`, before doing a grid search on `mtry`. Below, we illustrate the utility of ECV in tuning factions of features for random forests.

We consider data model (M2) with $\sigma = 5$, $n = 100$ and $p = 1,000$ (such that data aspect ratio is $\phi = 10$). We set the maximum ensemble size M_{max} to be 100, which is the default in `sckit-learn`, and set $M_0 = 25$. We first fix the subsample size as $k = 90$, and tune over factions of features $r \in \mathcal{R} = \{0.1, 0.2, \ldots, 0.9, 1\}$. After $\widehat{r}$ is obtained based on the ECV estimates, we further tune $k \in \mathcal{K} = \{10, 20, \ldots, 90\}$ by fixing the factions of features as $\widehat{r}$.

The results are shown in Figure S14. For tuning factions of features, we see a huge gain of time-saving from extrapolation of the risk estimate of the ensemble size M. This suggests that ECV is also beneficial in tuning other parameters other than the subsample size. On the other hand, even after fractions of features are tuned, we observed improvement with further tuning of the subsample size. When the ensemble size is $M = 10$, the improvement is significant. When M is large, we still see consistent slight improvement compared to the one with only fractions of features tuned.

S6.5 Risk profile in Section 6

Hao et al. (2021) collected samples of 50,781 human peripheral blood mononuclear cells

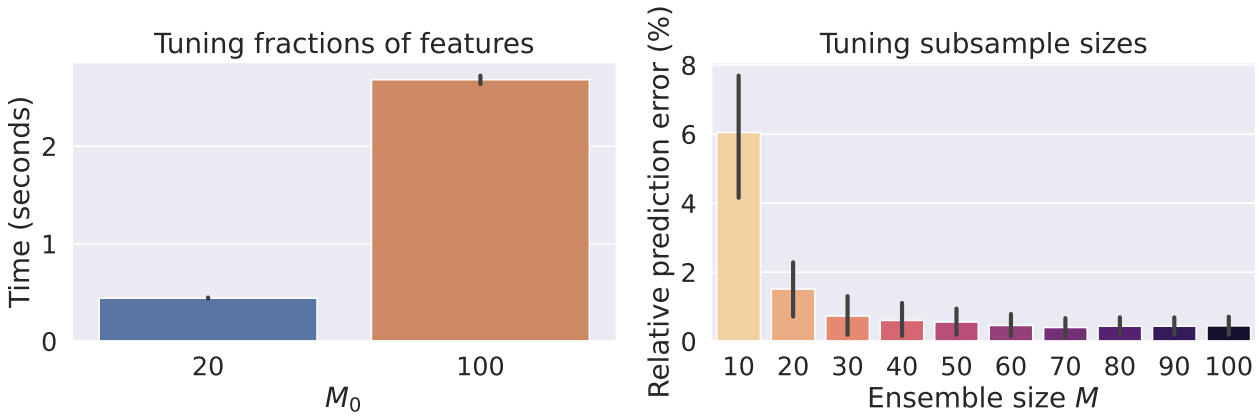

Figure S14: Tuning factions of features, subsample size by ECV. The left panel shows that computational time with extrapolation ($M_0 = 20$) and without extrapolation ($M_0 = 100$). The right panel shows the relative prediction error (the difference between the prediction errors without and with subsample size tunings, normalized by the null risk) in ensemble size M.

(PBMCs) originating from eight volunteers post-vaccination (day 3) of an HIV vaccine. The readers can follow the Seurat tutorial: `https://satijalab.org/seurat/articles/multimodal_reference_mapping.html` to download the preprocessed dataset. This single-cell CITE-seq dataset simultaneously measures 20,729 genes and 228 proteins in individual cells. As most genes are not variable across the dataset, we further subset the top 5,000 highly variable genes that exhibit high cell-to-cell variation in the dataset and the top 50 highly abundant surface proteins. Then, the genes and proteins are size-normalized to have total counts 10^4 and log-normalized. In Figure S15, we visualize the low-dimensional cell embeddings as well as the histograms of gene expressions and protein abundances in the Mono cell type. Overall, the gene expressions are extraordinarily sparse and zero-inflated distributed. On the contrary, the distribution of protein abundances is close to normal distributions.

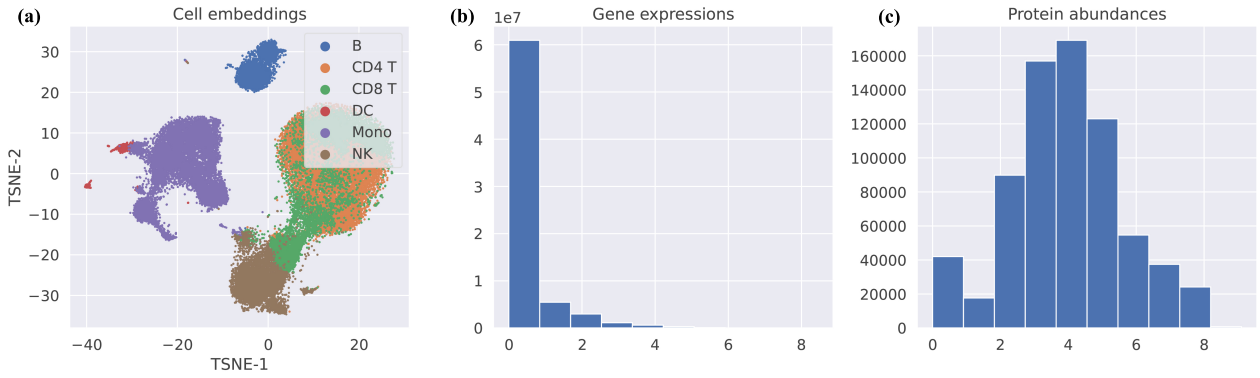

Figure S15: Overview of the single-cell CITE-seq dataset from Hao et al. (2021). (a) The 2-dimensional tSNE cell embeddings. (b-c) The histogram of overall log-normalized gene expressions and protein abundances in the Mono cell type.

As there is the most outstanding level of heterogeneity within T cell subsets, we only consider the non-T cell types for the experiments and randomly split cells in each cell type into the training and test sets with equal probability. Different cell types consist of different

Table S6: Description of different cell types in the CITE-sep dataset (Hao et al., 2021). The samples of PBMCs originate from eight volunteers post-vaccination (day 3) of an HIV vaccine.

Cell type	Training size n	Test size n_{te}	Data aspect ratio p/n
DC	516	515	9.71
B	2279	2279	2.19
NK	3152	3152	1.59
Mono	7156	7156	0.70

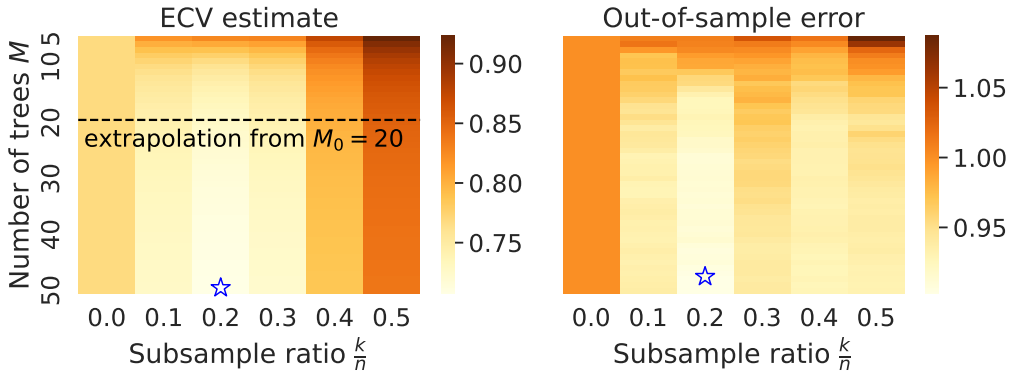

Figure S16: Heatmap of ECV performance on predicting the abundances of surface protein CD103 in the DC cell type by random forests (subagging). The left and right panels show the NMSEs of OOB risk estimates and out-of-sample prediction risk, respectively. The values are normalized by the empirical variance of the response estimated from the test set; the darker, the larger value of NMSE. The extrapolated risk estimates are based on $M_0 = 20$ trees.

numbers of cells and have various data aspect ratios. As summarized in Table S6, the four cell types cover both low-dimensional ($n > p$) and high-dimensional ($n < p$) datasets. Because the gene expressions exhibit high dimensionality, sparsity, and heterogeneity, predicting the protein abundances based on the transcriptome is thus a challenging problem.

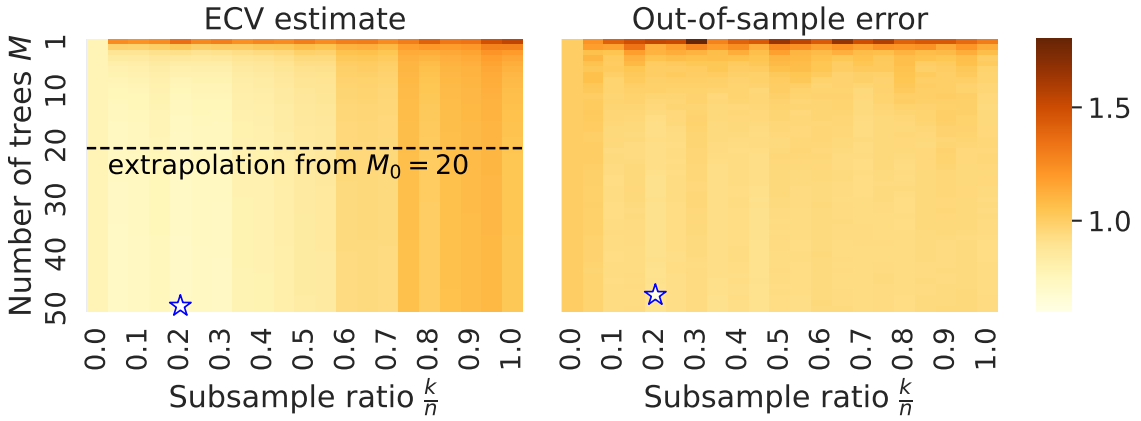

Figure S17: Heatmap of ECV performance on predicting the abundances of surface protein CD103 in the DC cell type by random forests (subagging) as in fig. S16 with $M \in \{1, \ldots, 50\}$ and k/n ranging from 0 to 1 displayed.

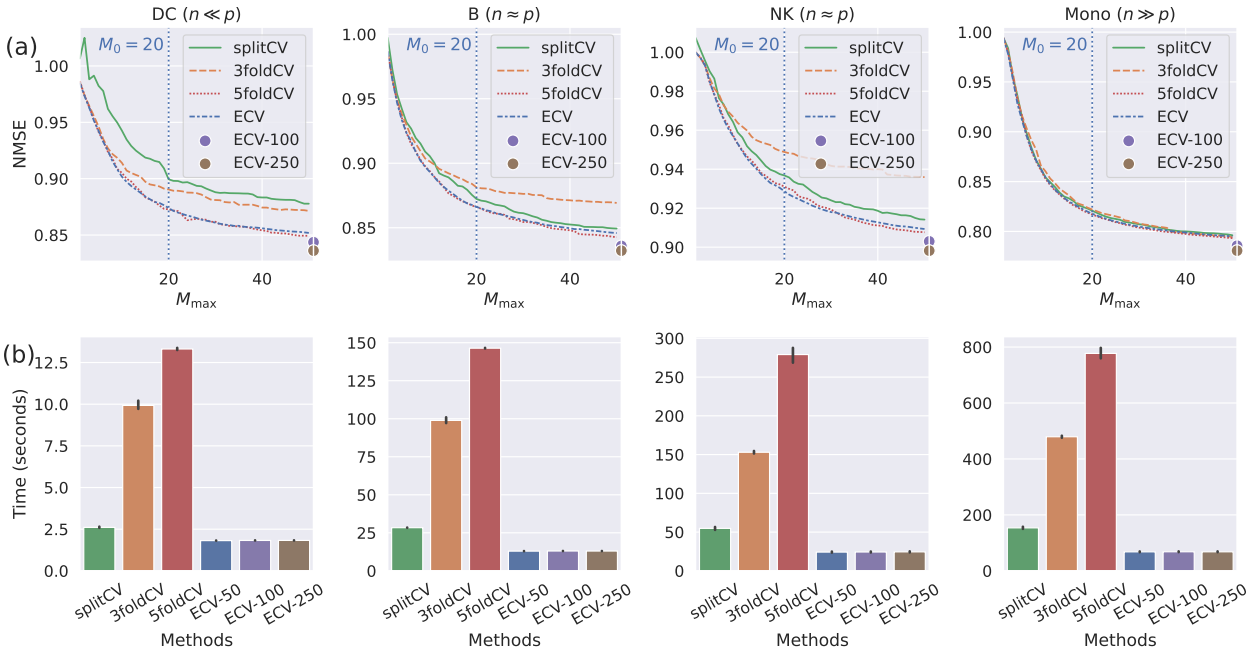

Figure S18: Performance of cross-validation methods on predicting the protein abundances in different cell types. (a) The average normalized mean squared error (NMSE) of the cross-validated predictors for different methods. ECV (bagging) uses $M_0 = 20$ trees to extrapolate risk estimates, and the dashed lines represent the performance using $M_{\max} \in \{100, 250\}$. (b) The average time consumption for cross-validation in seconds.

References

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

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

Austern, M. and Zhou, W. (2020). Asymptotics of cross-validation. *arXiv preprint arXiv:2001.11111*.

Bates, S., Hastie, T., and Tibshirani, R. (2021). Cross-validation: what does it estimate and how well does it do it? *arXiv preprint arXiv:2104.00673*.

Bayle, P., Bayle, A., Janson, L., and Mackey, L. (2020). Cross-validation confidence intervals for test error. *Advances in Neural Information Processing Systems*, 33:16339–16350.

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

Celisse, A. and Guedj, B. (2016). Stability revisited: new generalisation bounds for the leave-one-out. *arXiv preprint arXiv:1608.06412*.

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

Geurts, P., Ernst, D., and Wehenkel, L. (2006). Extremely randomized trees. *Machine learning*, 63:3–42.

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.

Hao, Y., Hao, S., Andersen-Nissen, E., Mauck, W. M., Zheng, S., Butler, A., Lee, M. J., Wilk, A. J., Darby, C., Zager, M., et al. (2021). Integrated analysis of multimodal single-cell data. *Cell*, 184(13):3573–3587.

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

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

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

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

Lee, A. J. (1990). *U-statistics: Theory and Practice*. Routledge. First edition.

Lei, J. (2020). Cross-validation with confidence. *Journal of the American Statistical Association*, 115(532):1978–1997.

Liaw, A., Wiener, M., et al. (2002). Classification and regression by randomforest. *R news*, 2(3):18–22.

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

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

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

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

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

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

Rad, K. R., Zhou, W., and Maleki, A. (2020). Error bounds in estimating the out-of-sample prediction error using leave-one-out cross validation in high-dimensions. In *International Conference on Artificial Intelligence and Statistics*, pages 4067–4077. PMLR.

Stephenson, W. and Broderick, T. (2020). Approximate cross-validation in high dimensions with guarantees. In *International Conference on Artificial Intelligence and Statistics*.

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

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

Wager, S., Hastie, T., and Efron, B. (2014). Confidence intervals for random forests: The jackknife and the infinitesimal jackknife. *The Journal of Machine Learning Research*, 15(1):1625–1651.

Wang, S., Zhou, W., Lu, H., Maleki, A., and Mirrokni, V. (2018). Approximate leave-one-out for fast parameter tuning in high dimensions. In *International Conference on Machine Learning*.

Wilson, A., Kasy, M., and Mackey, L. (2020). Approximate cross-validation: Guarantees for model assessment and selection. In *International Conference on Artificial Intelligence and Statistics*.

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