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

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

[†]Department of Statistics and Data Science, Carnegie Mellon University

[♯]Machine Learning Department, Carnegie Mellon University

[‡]Department of Statistics, University of California, Berkeley

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 1.
  - Appendix S2.2 proves Proposition 2.
  - Appendix S2.3 proves Theorem 1, conditional on certain helper lemmas presented in the next section.
- Appendix S3 provides intermediate concentration results used in the proof of Theorem 1.
- In Appendix S4, we present proof of results in Section 4.
  - Appendix S4.1 proves Theorem 2.
  - Appendix S4.2 extends additive optimality in Theorem 2 to multiplicative optimality stated in Remark 1.
  - Appendix S4.3 specialize Theorem 2 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; Geisser, 1975; Stone, 1974, 1977). 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 (Rad and Maleki, 2020; Stephenson and Broderick, 2020; Wang et al., 2018; Wilson et al., 2020). Another line of research about CV is on statistical inference; see, for example, Bates et al. (2021); Lei (2020); Wager et al. (2014). 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 Celisse and Guedj (2016); Kumar et al. (2013). 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 1 (Squared risk decomposition)

Proof of Proposition 1. We start by expanding the squared risk as:

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

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

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 2 (Consistent component risk estimation)

Proof of Proposition 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

$$\mathbb{P}\left(\Delta_n \geq C\widehat{\sigma}_I \max\left\{\sqrt{\frac{A\log n}{|I^c|}}, \frac{A\log n}{|I^c|}\right\}\right) \leq n^{-A}, \tag{12}$$

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 1 (Uniform risk estimation over (M, k))

Proof of Theorem 1. Before we prove Theorem 1, we present the proof strategy in Fig. S1. In Fig. S1, four important auxiliary lemmas and propositions are deferred to the next section. 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 S2 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 S4 we have

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

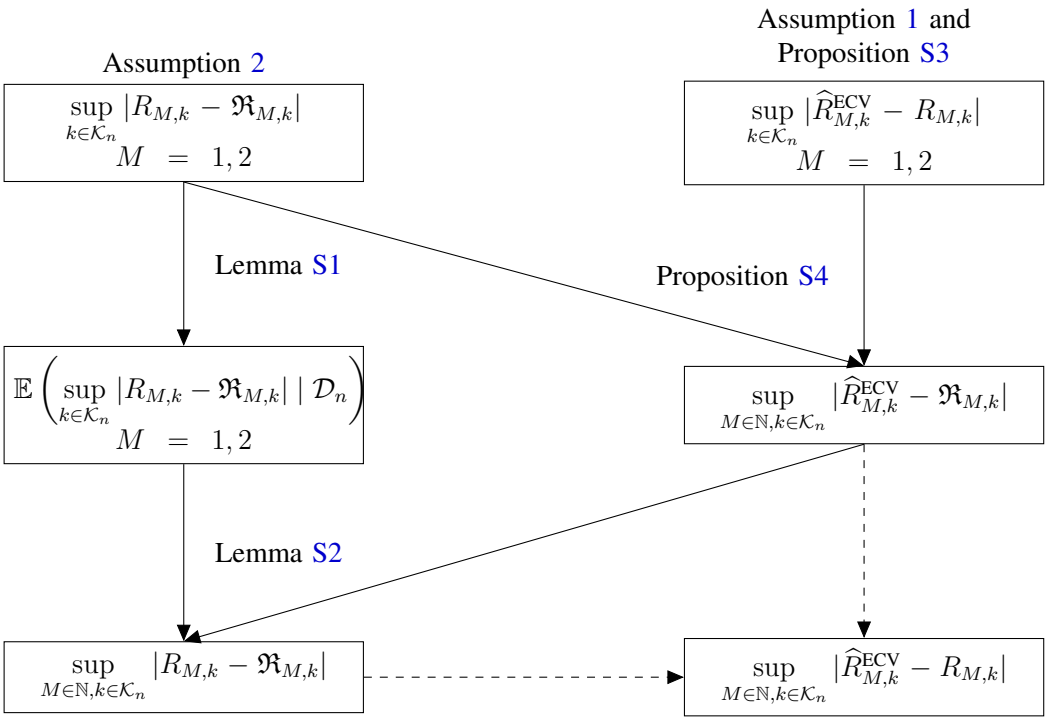

Figure S1: Reduction strategy for obtaining concentration results in Theorem 1. The solid lines indicate basic components presented later in Appendix S3 and the dashed lines indicate the proof strategy in Appendix S2.3. Here, $\mathfrak{R}_{M,k} = 2\mathfrak{R}_{2,k} - \mathfrak{R}_{1,k} + 2(\mathfrak{R}_{1,k}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

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 1

In this section, we show the concentration of conditional prediction risks to their limits. Proposition S3 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 S4 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 S1 establishes the concentration for the expected (with respect to sampling) conditional risk over $k \in \mathcal{K}_n$. Lemma S2

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 (Uniform consistency over k for $M = 1, 2$)**.** *Suppose that Assumption 2 holds, then ECV estimates defined in* (6) *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. 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:

$$\begin{aligned}
\Delta_{1,k} &= \widehat{R}^{\mathrm{ECV}}_{1,k} - \frac{1}{M_0} \sum_{\ell=1}^{M_0} R_{1,k}(\widehat{f}; \mathcal{D}_n, \{I_{k,\ell}\}) \\
\Delta_{2,k} &= \widehat{R}^{\mathrm{ECV}}_{2,k} - \frac{1}{M_0(M_0 - 1)} \sum_{\substack{\ell, m \in [M_0] \\ \ell \neq m}} R_{2,k}(\widehat{f}_{2,k}; \mathcal{D}_n, \{I_{k,\ell}, I_{k,m}\}).
\end{aligned} \tag{13}$$

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

$$\sup_{k \in \mathcal{K}_n} |\widehat{R}^{\mathrm{ECV}}_{1,k} - \mathfrak{R}_{1,k}| \leq \sup_{k \in \mathcal{K}_n} |\Delta_{1,k}| + \sup_{k \in \mathcal{K}_n, \ell \in [M_0]} |R_{1,k}(\widehat{f}; \mathcal{D}_n, \{I_{k,\ell}\}) - \mathfrak{R}_{1,k}| \tag{14}$$

$$\sup_{k\in\mathcal{K}_n}|\widehat{R}^{\mathrm{ECV}}_{2,k}-\mathfrak{R}_{1,k}|\le\sup_{k\in\mathcal{K}_n}|\Delta_{2,k}|\ +\sup_{k\in\mathcal{K}_n,m,\ell\in[M_0]}|R_{2,k}(\widehat{f};\mathcal{D}_n,\{I_{k,m},I_{k,\ell}\})-\mathfrak{R}_{1,k}| \tag{15}$$

where $\Delta_{M,k}$ are defined in (13). Next we analyze each term separately for $M=1,2$.

Part (1). For the first term, from (12) in Proposition 2 and applying union bound over $k\in\mathcal{K}_n$, we have that

$$\begin{aligned}
&\mathbb{P}\left(\sup_{k\in\mathcal{K}_n}|\Delta_{1,k}|\ge C\max_{k\in\mathcal{K}_n,\ell\in[M_0]}\widehat{\sigma}_{I_{k,\ell}}\max\left\{\sqrt{\frac{\log(M_0|\mathcal{K}_n|\eta)}{n/\log n}},\frac{\log(M_0|\mathcal{K}_n|\eta)}{n/\log n}\right\}\right)\le\frac{1}{\eta},\\
&\mathbb{P}\left(\sup_{k\in\mathcal{K}_n}|\Delta_{2,k}|\ge C\max_{k\in\mathcal{K}_n,\ell\neq m\in[M_0]}\widehat{\sigma}_{I_{k,\ell}\cup I_{k,m}}\max\left\{\sqrt{\frac{\log(M_0|\mathcal{K}_n|\eta)}{n/\log n}},\frac{\log(M_0|\mathcal{K}_n|\eta)}{n/\log n}\right\}\right)\le\frac{1}{\eta},
\end{aligned} \tag{16}$$

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

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

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

$$\sup_{k\in\mathcal{K}_n}|\Delta_{M,k}|=\mathcal{O}_p\left(\widehat{\sigma}_n\sqrt{\frac{\log^2 n}{n}}\right), \tag{17}$$

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 (8) in Assumption 2 we have that when n is large enough, for all $\eta\ge 1$,

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

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

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

which implies that

$$\sup_{k\in\mathcal{K}_n, \ell\in[M_0]} |R_{1,k}(\widehat{f}; \mathcal{D}_n, \{I_{k,\ell}\}) - \mathfrak{R}_{1,k}| = \mathcal{O}_p(n^{\epsilon}\gamma_{1,n}). \tag{18}$$

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

$$\sup_{k\in\mathcal{K}_n, m\neq\ell\in[M_0]} |R_{2,k}(\widehat{f}; \mathcal{D}_n, \{I_{k,m}, I_{k,\ell}\}) - \mathfrak{R}_{1,k}| = \mathcal{O}_p(n^{\epsilon}\gamma_{2,n}). \tag{19}$$

Part (3). Combining (15), (17), (18) and (19) yields that

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

ignoring constant factors. □

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

Proposition S4 (Uniform consistency to $\mathfrak{R}_{M,k}$ over (M, k))**.** *Suppose that Assumptions 1 and 2 hold, then ECV estimates defined in* (6) *satisfy that*

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

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 S4. Note that $\widehat{R}^{\text{ECV}}_{M,k}$ satisfies the squared risk decomposition and the randomness of $\widehat{R}^{\text{ECV}}_{M,k}$ also due to both the full data $\mathcal{D}_n$ and random sampling.

From Proposition S3, 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 (7), 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 S1 (Concentration of expected risk over $k \in \mathcal{K}_n$)

To obtain tail bounds for the subsample conditional risk defined in (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:

$$R(\widetilde{f}_{M,k}(\cdot; \{\mathcal{D}_{I_\ell}\}_{\ell=1}^M);\, \mathcal{D}_n) = \int \mathbb{E}\left[\left(y - \widetilde{f}_{M,k}(\boldsymbol{x}; \{\mathcal{D}_{I_{k,\ell}}\}_{\ell=1}^M)\right)^2 \mid \mathcal{D}_n\right] \mathrm{d}P(\boldsymbol{x}, y). \tag{20}$$

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 S1 (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 S1. Define

$$B_{1,n} = \sup_{k\in\mathcal{K}_n} |R(\widetilde{f}_{1,k}; \mathcal{D}_n, \{I_{k,1}\}) - \mathfrak{R}_{1,k}|. \tag{21}$$

$$B_{2,n} = \sup_{k\in\mathcal{K}_n} |R(\widetilde{f}_{2,k}; \mathcal{D}_n, \{I_{k,1}, I_{k,2}\}) - \mathfrak{R}_{2,k}|. \tag{22}$$

We start with the $M = 1$ case by bounding the expectation of (21). Note that from (8), 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

$$\mathbb{P}(\gamma_{1,n}'^{-1} B_{1,n} \geq \eta) \leq \frac{1}{\eta^{1/\epsilon}}. \tag{23}$$

Then, it follows that

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

$$= \left(\eta_0 + C_0 \frac{\epsilon - 1}{\epsilon} \eta_0^{1-1/\epsilon}\right) \gamma'_{1,n}.$$

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

S3.4 Proof of Lemma S2 (Concentration of conditional risk)

Lemma S2 (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,*

$$\sup_{M \in \mathbb{N}, k \in \mathcal{K}_n} |R_{M,k} - \mathfrak{R}_{M,k}| = \mathcal{O}_p\left(C_1(|\mathcal{K}_n|)\gamma_{1,n} + C_2(|\mathcal{K}_n|)\gamma_{1,n}\right), \tag{24}$$

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 S2. By Proposition 1, we have for $\{I_{k,\ell}\}_{\ell=1}^M \overset{\text{SRS}}{\sim} \mathcal{I}_k$

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

Then, it follows that

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

$$\leq \sup_{M\geq 1}\left|\frac{1}{M}\sum_{\ell\in[M]}\sup_{k\in\mathcal{K}_n}|R(\widetilde{f}_{1,k};\,\mathcal{D}_n,\{I_{k,\ell}\})-\mathfrak{R}_{1,k}|\right| \\ + 2\sup_{M\geq 2}\left|\frac{1}{M(M-1)}\sum_{i,j\in[M],i\neq j}\sup_{k\in\mathcal{K}_n}|R(\widetilde{f}_{2,k};\,\mathcal{D}_n,\{I_{k,i},I_{k,j}\})-\mathfrak{R}_{2,k}|\right|. \tag{25}$$

We start by observing that the two terms

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

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 S1, 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 (25), it follows that

$$\begin{aligned}
&\sup_{M\in\mathbb{N},k\in\mathcal{K}_n} |R(\widetilde{f}_M;\mathcal{D}_n,\{I_{k,\ell}\}_{\ell=1}^M) - \mathfrak{R}_{M,k}| \\
&= \sup_{M\geq 1} |U_M| + 2\sup_{M\geq 2} |U'_M| \\
&= \mathcal{O}_p(n^\epsilon(\gamma_{1,n}+\gamma_{2,n})). \qquad (26)
\end{aligned}$$

This completes the proof. □

S4 Proofs of results in Section 4

S4.1 Proof of Theorem 2 (δ-optimality of ECV)

Proof of Theorem 2. 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 1, we have

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

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

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

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

Step 1: Bounding the difference between $\widehat{R}^{\text{ECV}}_\infty(\widetilde{f}_{M_0,\widehat{k}})$ 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 (4) 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

$$\widehat{R}^{\mathrm{ECV}}_{\infty,\widehat{k}} = \inf_{k\in\mathcal{K}_n} R_{\infty,k} + \mathcal{O}_p(\zeta_n) = \inf_{M\in\mathbb{N},k\in\mathcal{K}_n} R_{M,k} + \mathcal{O}_p(\zeta_n) \tag{27}$$

where the first equality is due to Theorem 1.

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

$$\begin{aligned}
&|R_{\widehat{M},\widehat{k}} - \inf_{M\in\mathbb{N},k\in\mathcal{K}_n} R_{M,k}| \\
&= |R_{\widehat{M},\widehat{k}} - \widehat{R}^{\mathrm{ECV}}_{\widehat{M},\widehat{k}} + \widehat{R}^{\mathrm{ECV}}_{\widehat{M},\widehat{k}} - \widehat{R}^{\mathrm{ECV}}_{\infty,\widehat{k}} + \widehat{R}^{\mathrm{ECV}}_{\infty,\widehat{k}} - R_{\infty,k^*}| \\
&\le |R_{\widehat{M},\widehat{k}} - \widehat{R}^{\mathrm{ECV}}_{\widehat{M},\widehat{k}}| + |\widehat{R}^{\mathrm{ECV}}_{\widehat{M},\widehat{k}} - \widehat{R}^{\mathrm{ECV}}_{\infty,\widehat{k}}| + |\widehat{R}^{\mathrm{ECV}}_{\infty,\widehat{k}} - R_{\infty,k^*}|.
\end{aligned} \tag{28}$$

From Part (1) we know that the first term in (28) can be bounded as $|R_{\widehat{M},\widehat{k}} - \widehat{R}^{\mathrm{ECV}}_{\widehat{M},\widehat{k}}| = \mathcal{O}_p(\zeta_n)$. From (27) we know that the last term in (28) 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 (28) . By the definition of $\widehat{R}^{\mathrm{ECV}}_{M,k}$ in (7), we have that

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

Since $\widehat{M} = \lceil 2/\delta \cdot \widehat{R}^{\mathrm{ECV}}_{1,\widehat{k}} - \widehat{R}^{\mathrm{ECV}}_{2,\widehat{k}} \rceil \ge 2/\delta \cdot \widehat{R}^{\mathrm{ECV}}_{1,\widehat{k}} - \widehat{R}^{\mathrm{ECV}}_{2,\widehat{k}}$, the second term in (28) is bounded by

$$|\widehat{R}^{\mathrm{ECV}}_{\widehat{M},\widehat{k}} - \widehat{R}^{\mathrm{ECV}}_{\infty,\widehat{k}}| = \frac{2}{\widehat{M}} \left|\widehat{R}^{\mathrm{ECV}}_{1,\widehat{k}} - \widehat{R}^{\mathrm{ECV}}_{2,\widehat{k}}\right| \le \delta. \tag{29}$$

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

S4.2 Proof of multiplicative optimality

Proposition S5 (Multiplicative optimality)**.** *Under the same conditions in Theorem 2, 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*

$$\widehat{R}^{\mathrm{ECV}}_{\widehat{M},\widehat{k}} \leq (1+\delta) \inf_{M \in \mathbb{N}, k \in \mathcal{K}_n} \widehat{R}^{\mathrm{ECV}}_{M,k}, \tag{30}$$

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 S5. By definition of (30), we have

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

where the inequality is from Theorem 2 and the last equality is from the assumption that the risks are lower bounded. Further, since from Theorem 1, $\sup_{M \in \mathbb{N}, k \in \mathcal{K}_n} |R_{M,k} - \widehat{R}^{\mathrm{ECV}}_{M,k}| = \mathcal{O}_p(\zeta_n)$, we have

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

This finishes the proof. □

S4.3 Concrete example: bagged ridge predictors

Application of Theorem 2 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 $\{I_\ell\}_{\ell=1}^M \overset{\text{SRS}}{\sim} \mathcal{I}_k$.

We consider Assumptions S3-S4 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 S3 (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 $k \geq 2$. 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) $r_1 \geq r_2 \geq \cdots \geq r_p \geq 0$, and $\boldsymbol{W} \in \mathbb{R}^{p \times p}$ is an orthonormal matrix containing the associated eigenvectors $\boldsymbol{w}_1, \boldsymbol{w}_2, \ldots, \boldsymbol{w}_p \in \mathbb{R}^p$. Let H_p denote the empirical spectral distribution of $\boldsymbol{\Sigma}$ (supposed on $\mathbb{R}_{>0}$) whose value at any $r \in \mathbb{R}$ is given by*

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

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

Assumption S4 (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 S3-S4 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 S6, Assumptions S3-S4 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 S3-S4, 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 S3 and ϵ_i in Assumption S3 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 S6.

Proposition S6 (ECV for ridge predictors)**.** *Suppose Assumptions S3-S4 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 2 hold.*

We remark that Proposition S6 verifies Theorem 2 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 S3-S4 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 2. The generality of Theorem 2 to general predictors is illustrated empirically in the next section.

Proof of Proposition S6. 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 $M = 1$:

$$
\begin{aligned}
&R(\widetilde{f}_{1,k}; \mathcal{D}_I) \\
&= \mathbb{E}\{(y_0 - \boldsymbol{x}_0^\top \boldsymbol{\beta}_0)^2\} + \|\widehat{\boldsymbol{\beta}}_1(\mathcal{D}_I) - \boldsymbol{\beta}_0\|_{\boldsymbol{\Sigma}^{1/2}}^2 \\
&= \mathbb{E}\{(y_0 - \boldsymbol{x}_0^\top \boldsymbol{\beta}_0)^2\} + \boldsymbol{\beta}_0^\top \widehat{\boldsymbol{\Sigma}}_I (\widehat{\boldsymbol{\Sigma}}_I + \lambda \boldsymbol{I}_p)^{-2} \widehat{\boldsymbol{\Sigma}}_I \boldsymbol{\beta}_0 + \frac{1}{k^2} \boldsymbol{\varepsilon}^\top \boldsymbol{L}_I \boldsymbol{X} (\widehat{\boldsymbol{\Sigma}}_I + \lambda \boldsymbol{I}_p)^{-2} \boldsymbol{X}^\top \boldsymbol{L}_I \boldsymbol{\varepsilon}. \quad (31)
\end{aligned}
$$

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

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

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

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

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

which we denote as $C_{0,q}$. From the bounded moment assumption S4, 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 S3-S4, 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 $\{I_1, I_2\} \overset{\text{SRS}}{\sim} \mathcal{I}_k$,

$$R(\widetilde{f}_{1,k}; \mathcal{D}_I) - \mathfrak{R}_1(\phi, \phi_s) \to 0, \qquad R(\widetilde{f}_{2,k}; \mathcal{D}_n, \{I_1, I_2\}) - \mathfrak{R}_2(\phi, \phi_s) \to 0, \tag{32}$$

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

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

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

$$\limsup_{n\to\infty} \sup_{\eta>0} \eta^q \mathbb{P}(Z_n \geq \eta) \leq C_1^q + (C_{0,2q}^q + \mathfrak{R}_1(\phi, \phi_s)^{\frac{q}{2}}) 2^{q-1} C^{\frac{1}{2}}, \tag{33}$$

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

$$0 \leq R(\widetilde{f}_{2,k}; \mathcal{D}_n, \{I_1, I_2\})^{1/\epsilon} \leq 2^{-1/\epsilon}(R(\widetilde{f}_{1,k}; \mathcal{D}_n, \{I_1\}) + R(\widetilde{f}_{1,k}; \mathcal{D}_n, \{I_2\}))^{1/\epsilon}. \tag{34}$$

We next apply Pratt's lemma to show $L^{1/\epsilon}$-convergence for $M = 2$. Note that the tail bound condition (33) 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 $\gamma_{2,n} = \gamma_{2,n}^{\prime\epsilon}$ and $\gamma_{2,n} = o(n^{-\epsilon})$. Thus, Assumption 2 is satisfied under proportional asymp-

totics.

Part (4) Bounded variance proxy for CV estimates. From results by Patil et al. (2022, Remark 2.19), Assumption S3 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 2. 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 1 (Hypergeometric random variable)**.** *A random variable X follows the hypergeometric distribution $X \sim \text{Hypergeometric}(n, K, N)$ 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

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

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 S3 (Asymptotic proportions of shared observations, adapted from Patil et al. (2023))**.** *For $n \in \mathbb{N}$, define $\mathcal{I}_k = \{\{i_1, i_2, \ldots, i_k\} : 1 \leq i_1 < i_2 < \ldots < i_k \leq n\}$. Let $I_1, I_2 \overset{\text{SRSWR}}{\sim} \mathcal{I}_k$, define the random variable $i_0^{\text{SRSWR}} = |I_1 \cap I_2|$ to be the number of shared samples, and define i_0^{SRSWOR} accordingly. Then $i_0^{\text{SRSWR}} \sim \text{Binomial}(k, k/n)$ and $i_0^{\text{SRSWOR}} \sim \text{Hypergeometric}(k, k, n)$. 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, $n_m^{\nu} \leq k_m \leq n_m$ for some constant $\nu \in (0, 1)$. Then, $i_0^{\text{SRSWR}}/k_m - k_m/n_m \to 0$, and $i_0^{\text{SRSWOR}}/k_m - k_m/n_m \to 0$ 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 (7) 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	Fig. S3	Fig. S2	Fig. S5	Fig. S4	Fig. S6	Fig. 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	Fig. S8	Fig. 3	Fig. S9	Fig. S10	Fig. S11	Fig. S12

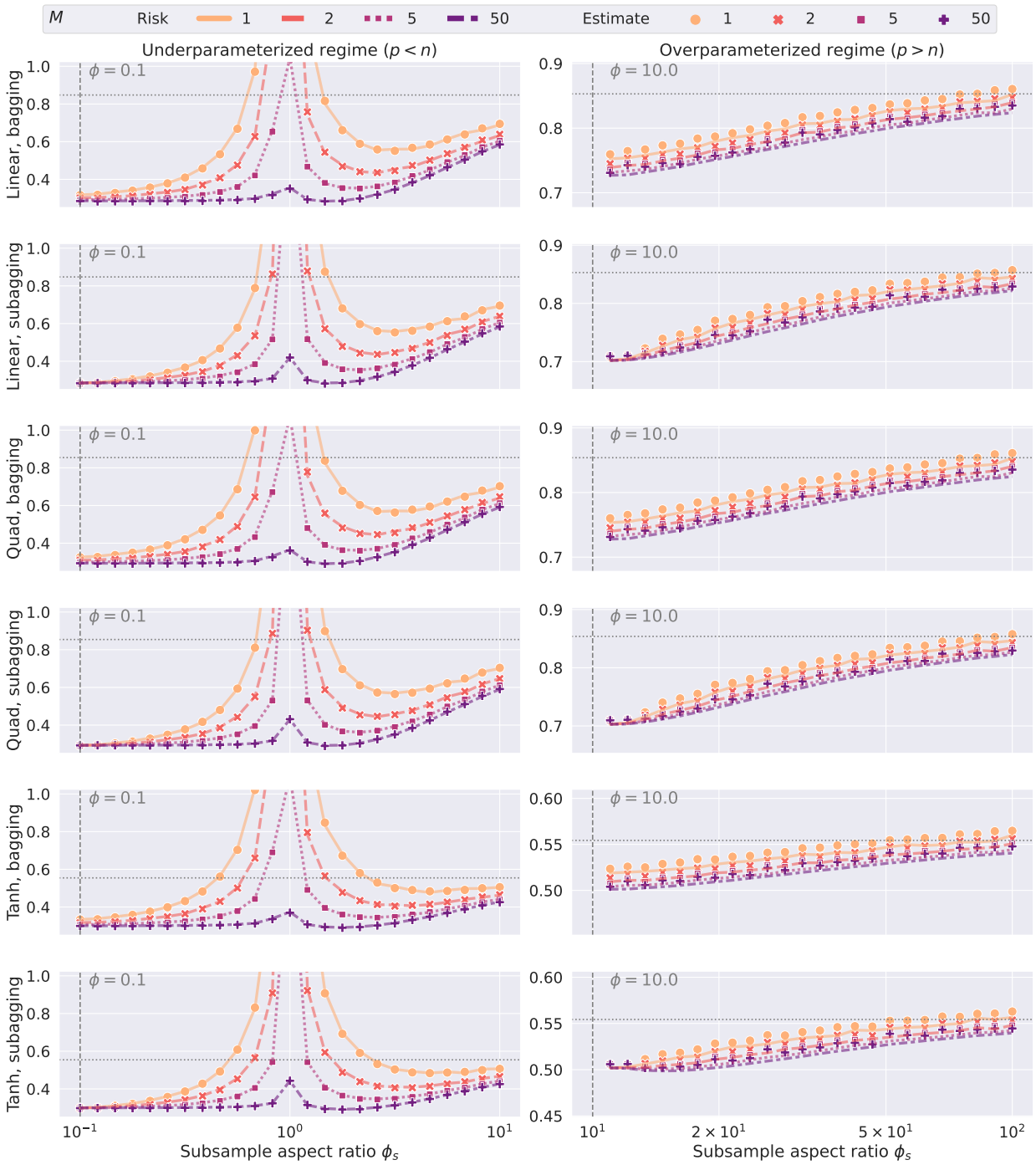

Figure S2: Finite-sample ECV estimate (points) and prediction risk (lines) for ridge predictors ($\lambda = 0.1$) using bagging and subagging, under nonlinear quadratic and tanh models (Section 5.1) for varying subsample size $k = \lfloor p/\phi_s \rfloor$, and the ensemble size M. For each value of M, the points denote the finite-sample ECV cross-validation estimates (7), and the lines denote the out-of-sample prediction error. See Appendix S6.1 for more details.

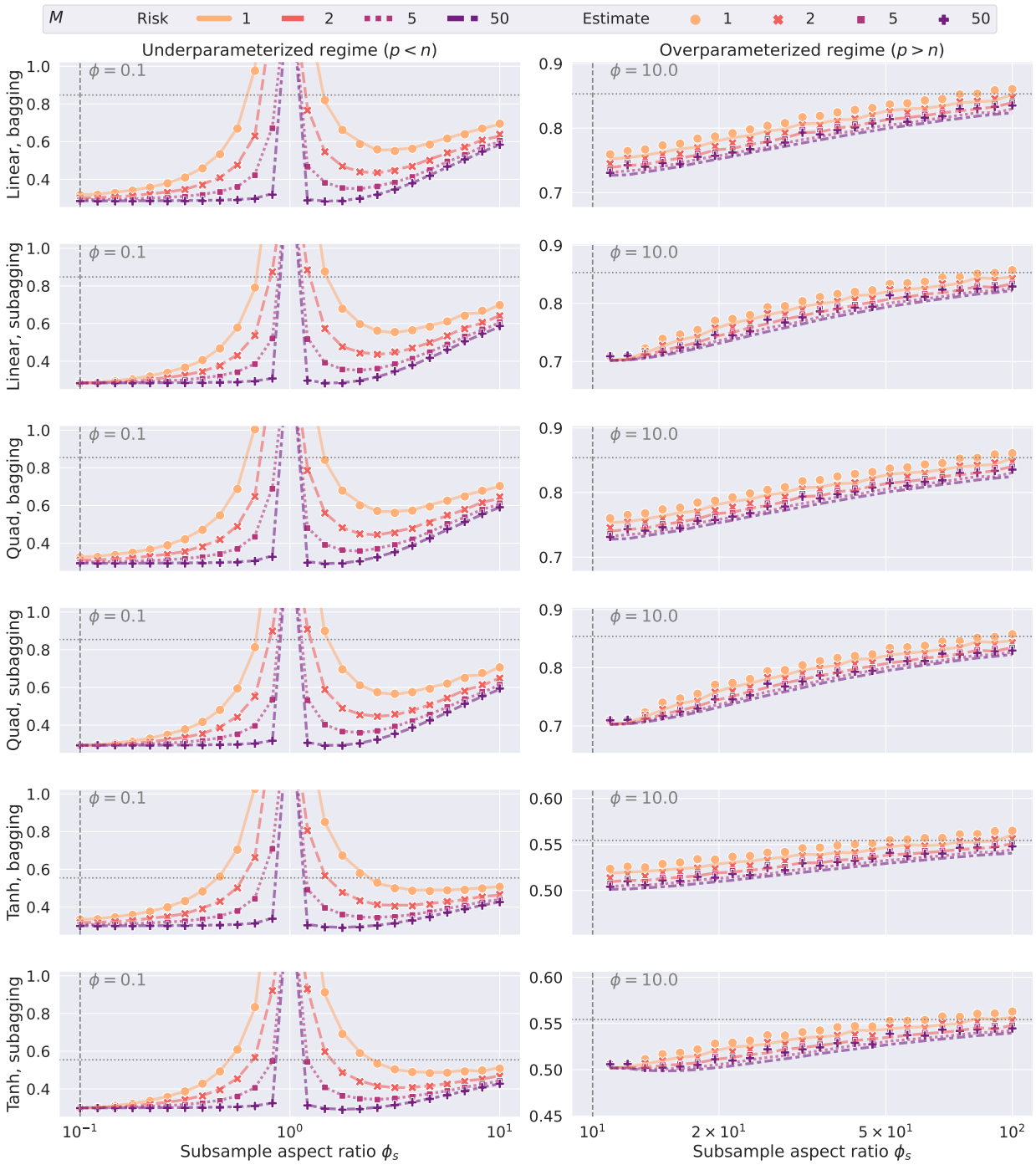

Figure S3: Finite-sample ECV estimate (points) and prediction risk (lines) for ridgeless predictors using bagging and subagging, under nonlinear quadratic and tanh models (Section 5.1) for varying subsample size $k = \lfloor p/\phi_s \rfloor$, and the ensemble size M. For each value of M, the points denote the finite-sample ECV cross-validation estimates (7), and the lines denote the out-of-sample prediction error. See Appendix S6.1 for more details.

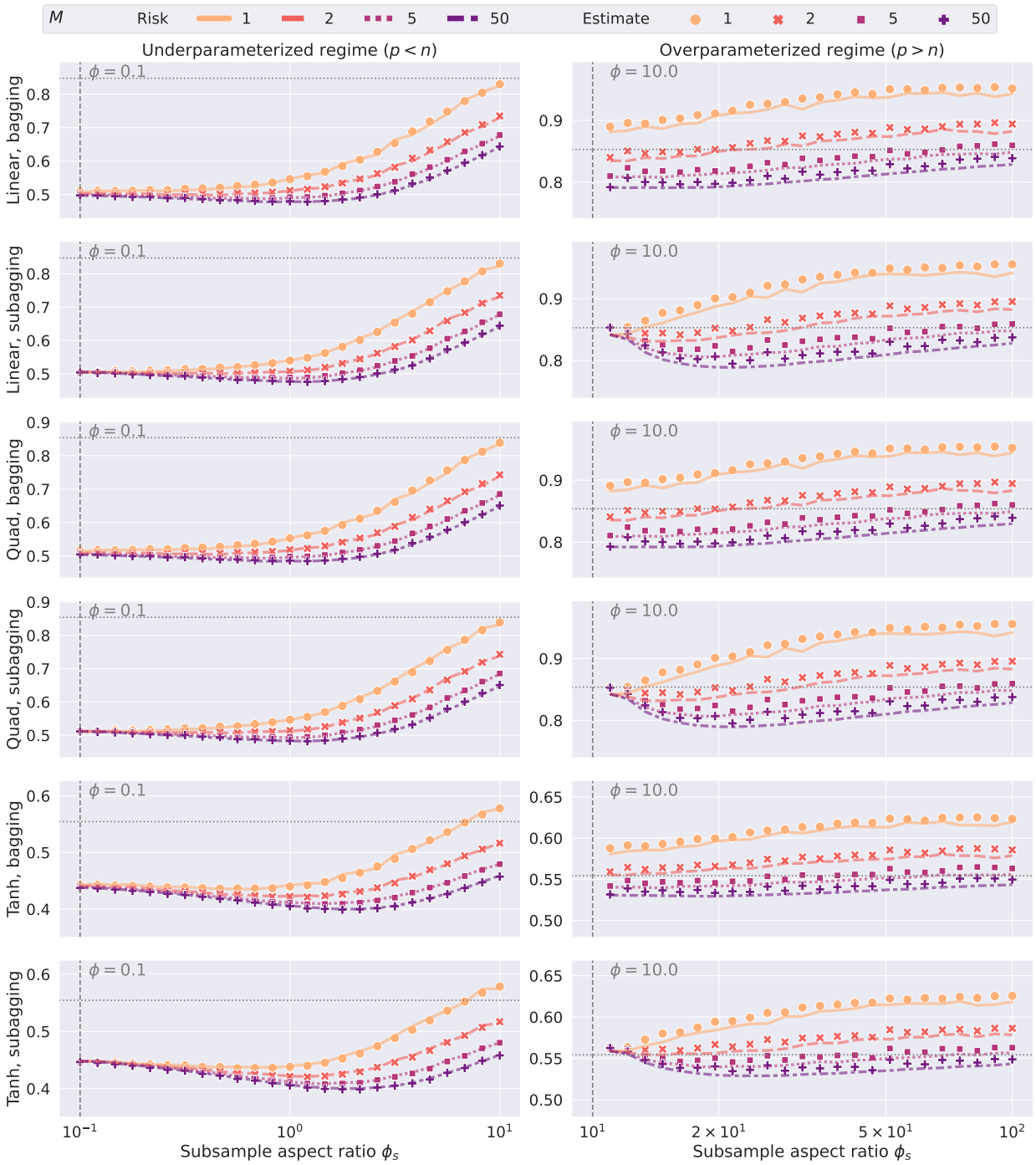

Figure S4: Finite-sample ECV estimate (points) and prediction risk (lines) for lasso predictors ($\lambda = 0.1$) using bagging and subagging, under nonlinear quadratic and tanh models (Section 5.1) for varying subsample size $k = \lfloor p/\phi_s \rfloor$, and the ensemble size M. For each value of M, the points denote the finite-sample ECV cross-validation estimates (7), and the lines denote the out-of-sample prediction error. See Appendix S6.1 for more details.

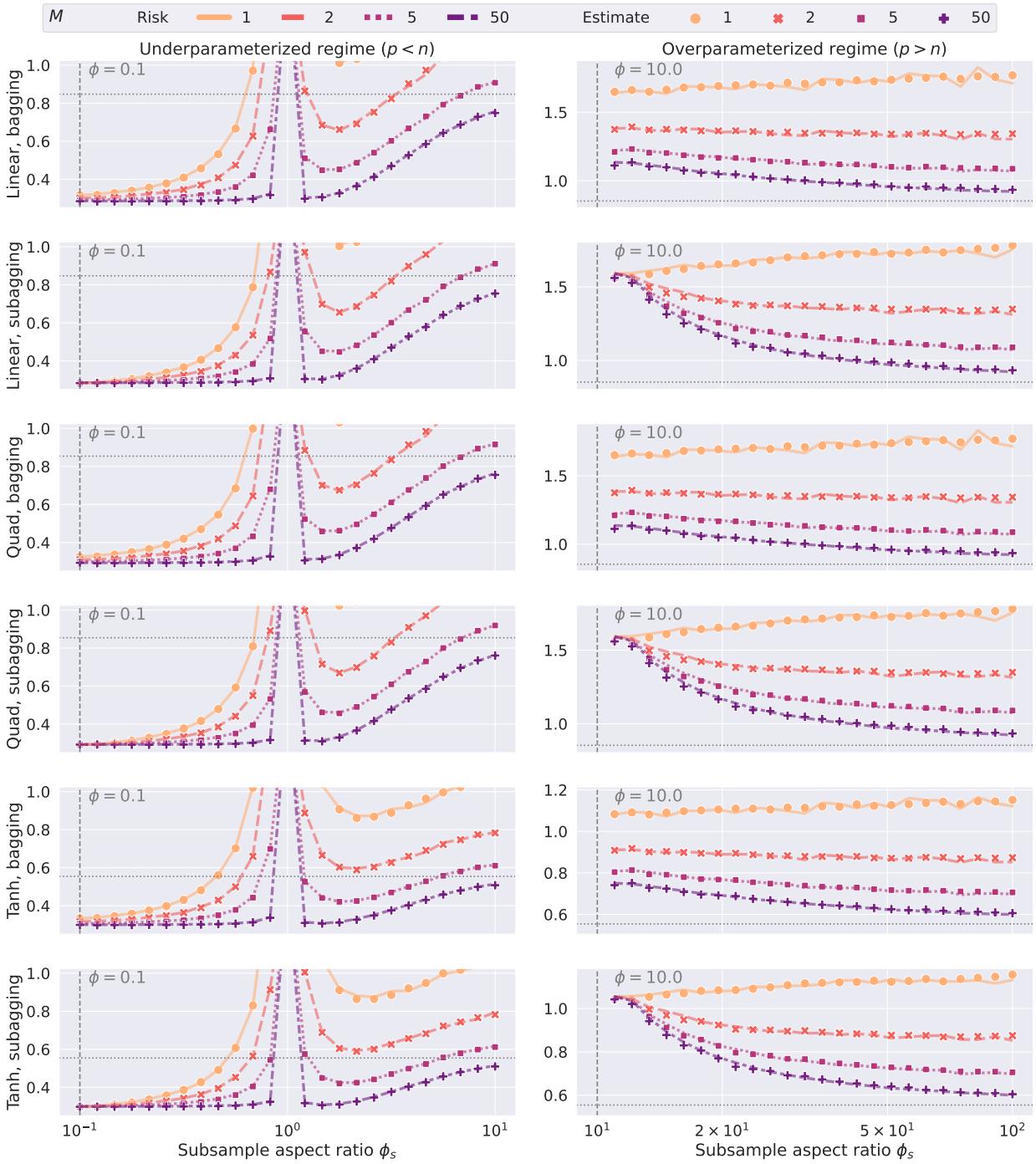

Figure S5: Finite-sample ECV estimate (points) and prediction risk (lines) for lassoless predictors using bagging and subagging, under nonlinear quadratic and tanh models (Section 5.1) for varying subsample size $k = \lfloor p/\phi_s \rfloor$, and the ensemble size M. For each value of M, the points denote the finite-sample ECV cross-validation estimates (7) 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.

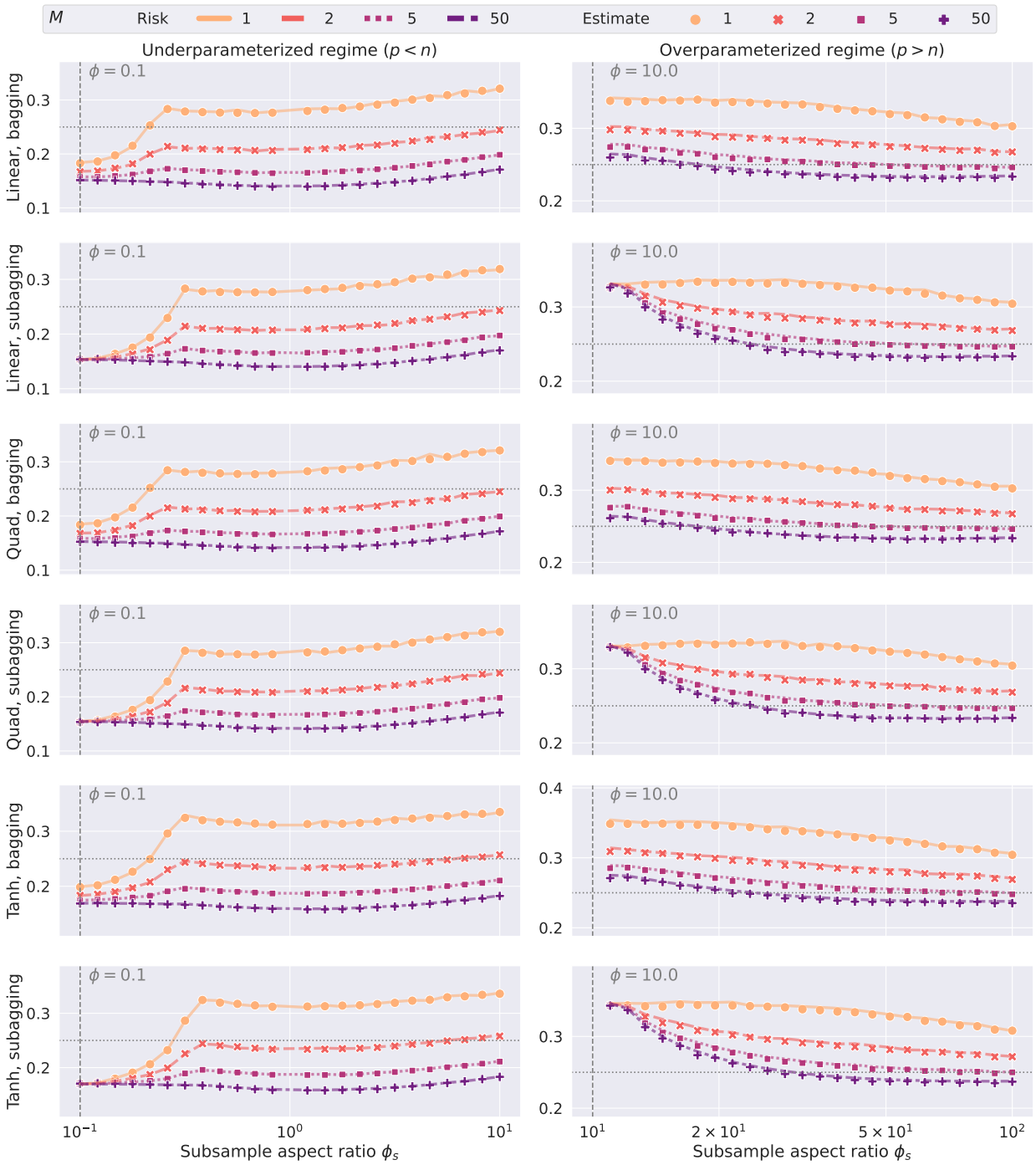

Figure S6: Finite-sample ECV estimate (points) and prediction risk (lines) for logistic predictors using bagging and subagging, under nonlinear quadratic and tanh models (Section 5.1) for varying subsample size $k = \lfloor p/\phi_s \rfloor$, and the ensemble size M. For each value of M, the points denote the finite-sample ECV cross-validation estimates (7), and the lines denote the out-of-sample prediction error. See Appendix S6.1 for more details.

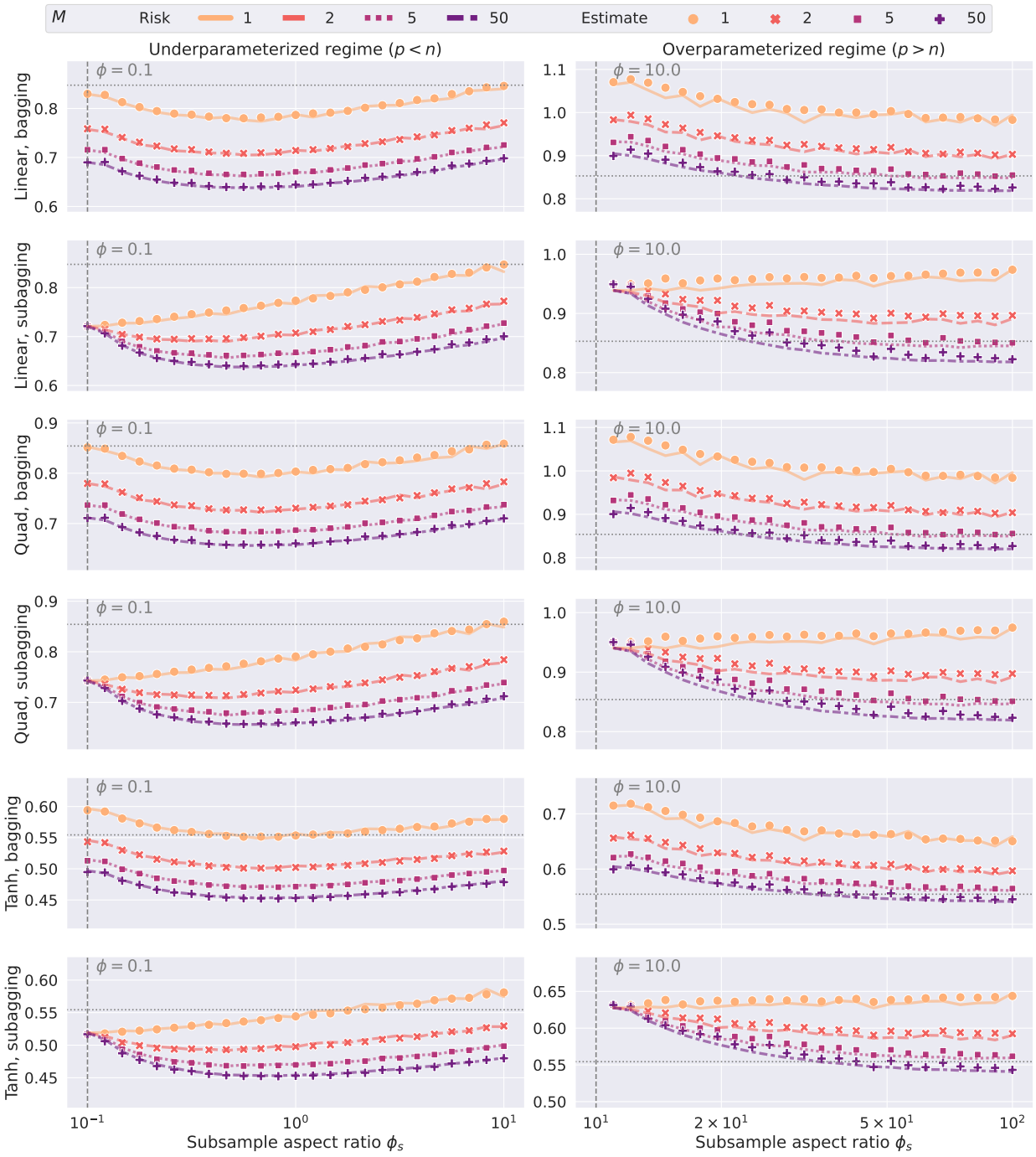

Figure S7: Finite-sample ECV estimate (points) and prediction risk (lines) for kNN predictors using bagging and subagging, under nonlinear quadratic and tanh models (Section 5.1) for varying subsample size $k = \lfloor p/\phi_s \rfloor$, and the ensemble size M. For each value of M, the points denote the finite-sample ECV cross-validation estimates (7), and the lines denote the out-of-sample prediction error. See Appendix S6.1 for more details.

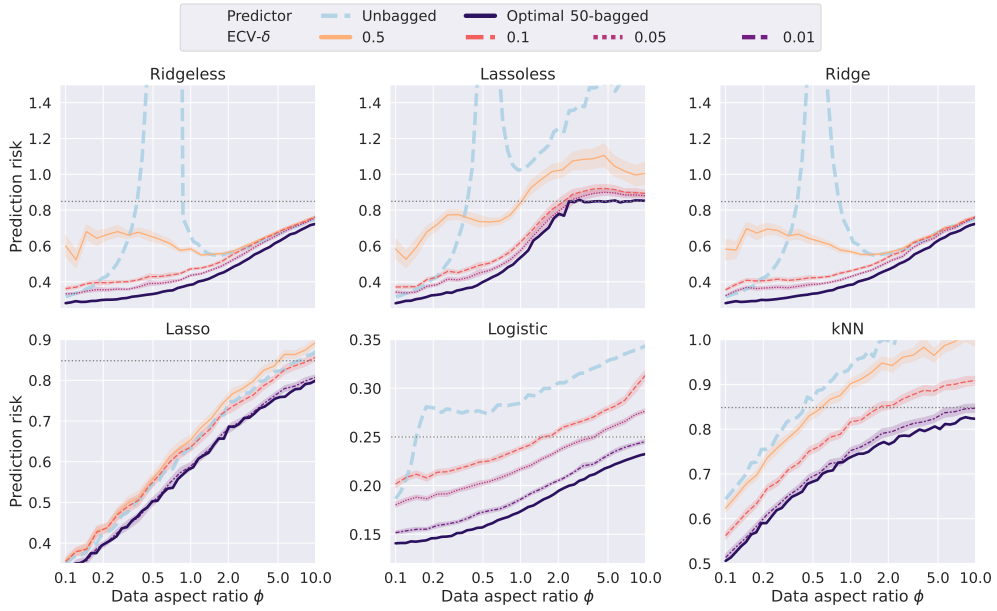

Figure S8: Prediction risk for different bagged predictors by ECV, under model (M1), for varying ϕ and tolerance threshold δ. An ensemble is fitted when (11) is satisfied with $\zeta = 5$. The null risks and the risks for the non-ensemble predictors are marked as gray dotted lines and blue dashed lines, 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.

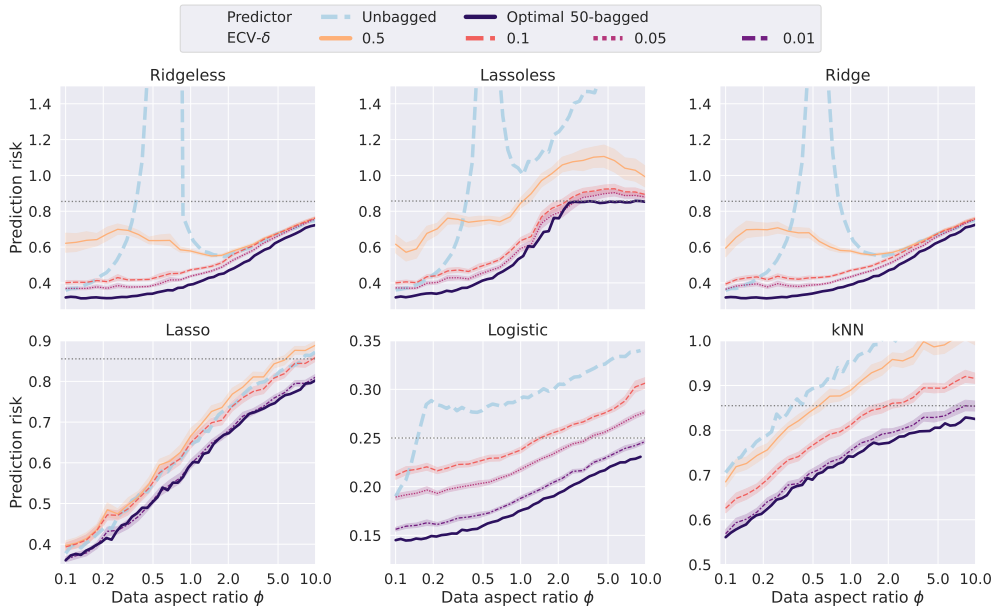

Figure S9: Prediction risk for different bagged predictors by ECV, under model (M3), for varying ϕ and tolerance threshold δ. An ensemble is fitted when (11) is satisfied with $\zeta = 5$. The null risks and the risks for the non-ensemble predictors are marked as gray dotted lines and blue dashed lines, 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.

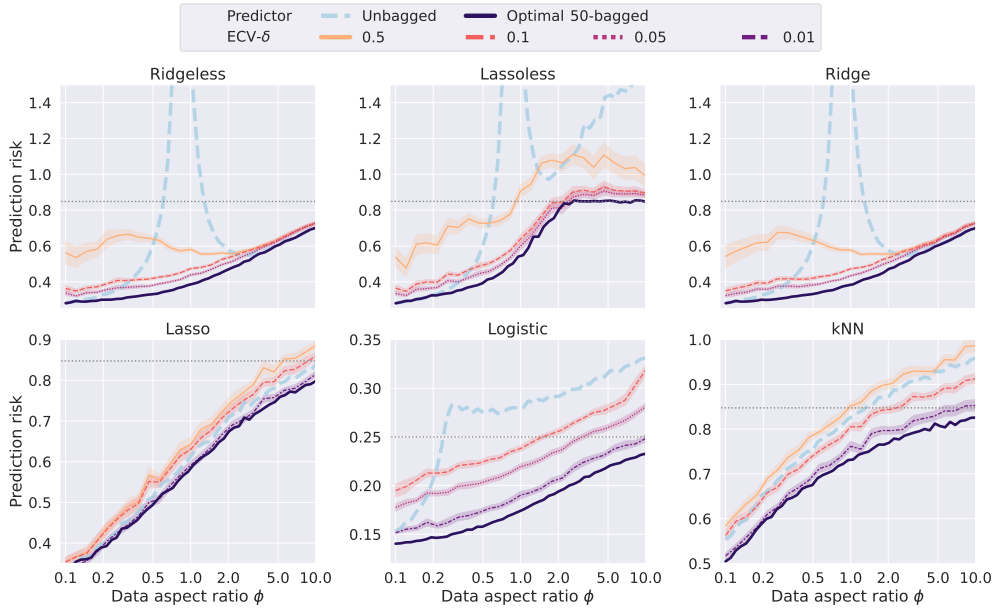

Figure S10: Prediction risk for different subagged predictors by ECV, under model (M1), for varying ϕ and tolerance threshold δ. An ensemble is fitted when (11) is satisfied with $\zeta = 5$. The null risks and the risks for the non-ensemble predictors are marked as gray dotted lines and blue dashed 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.

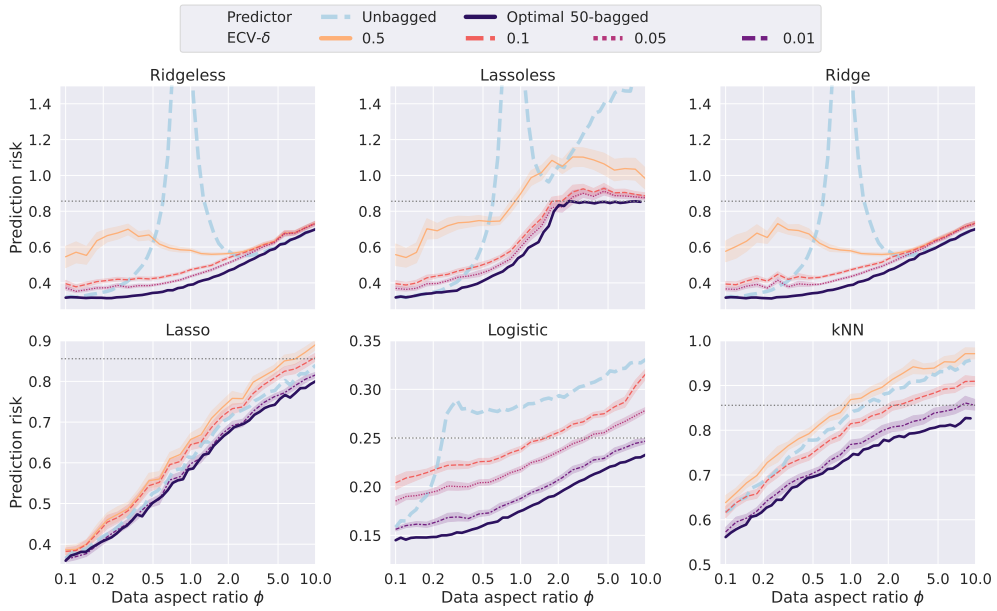

Figure S11: Prediction risk for different subagged predictors by ECV, under model (M2), for varying ϕ and tolerance threshold δ. An ensemble is fitted when (11) is satisfied with $\zeta = 5$. The null risks and the risks for the non-ensemble predictors are marked as gray dotted lines and blue dashed 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.

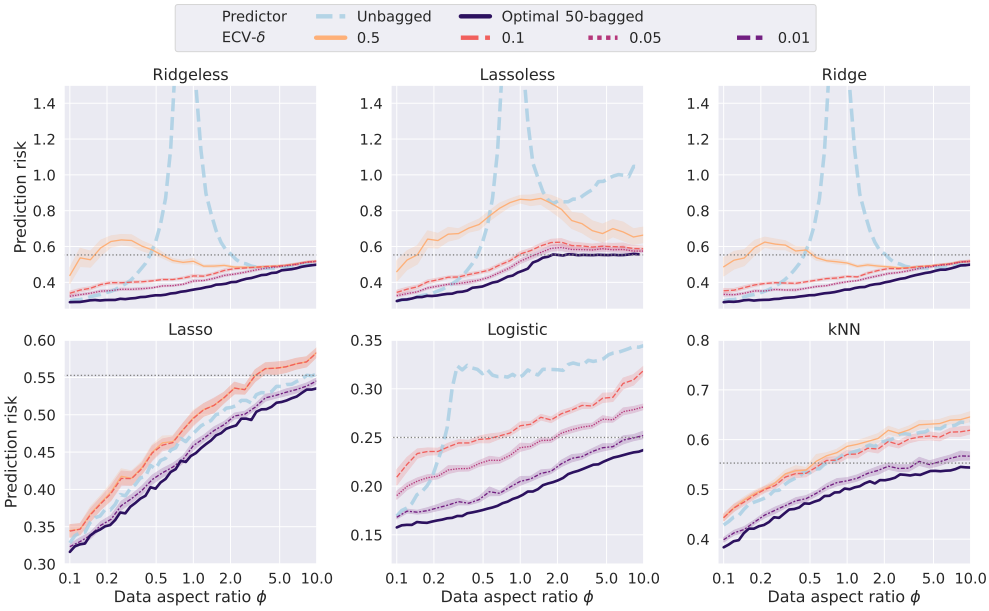

Figure S12: Prediction risk for different subagged predictors by ECV, under model (M3), for varying ϕ and tolerance threshold δ. An ensemble is fitted when (11) is satisfied with $\zeta = 5$. The null risks and the risks for the non-ensemble predictors are marked as gray dotted lines and blue dashed 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 Fig. 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 Fig. 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.

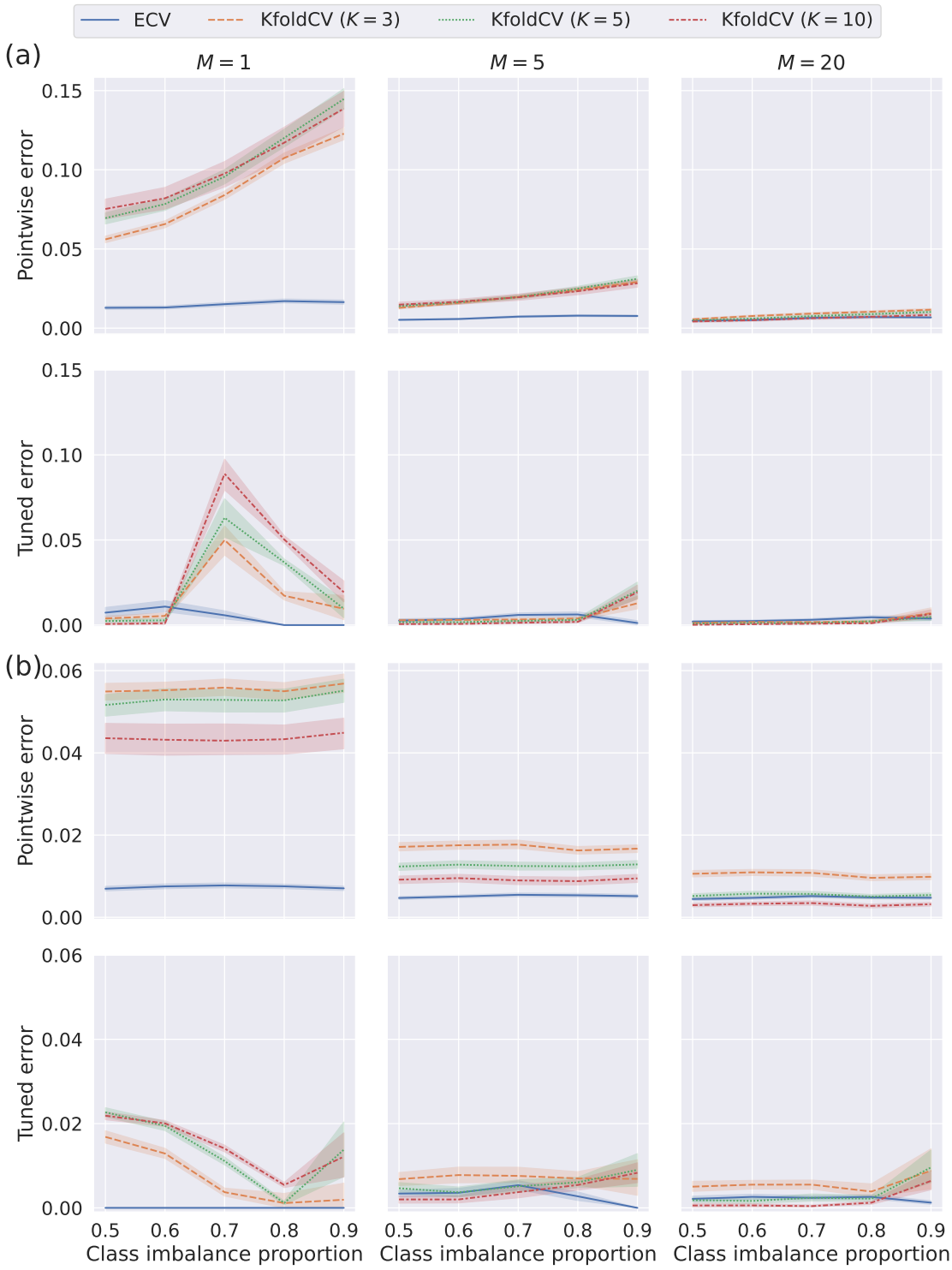

Figure S13: The pointwise and tuned errors for Logistic predictors with varying negative class proportions, in (a) the underparameterized regime ($\phi = 0.1$) and (b) the overparameterized ($\phi = 10$) regimes, with $n = 1,000$ and $p = \lfloor n\phi \rfloor$ across 50 repetitions. The setup is under 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 prespecified threshold, as in Lopes (2019); Lopes et al. (2020).

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

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.

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 Fig. 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 (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_`

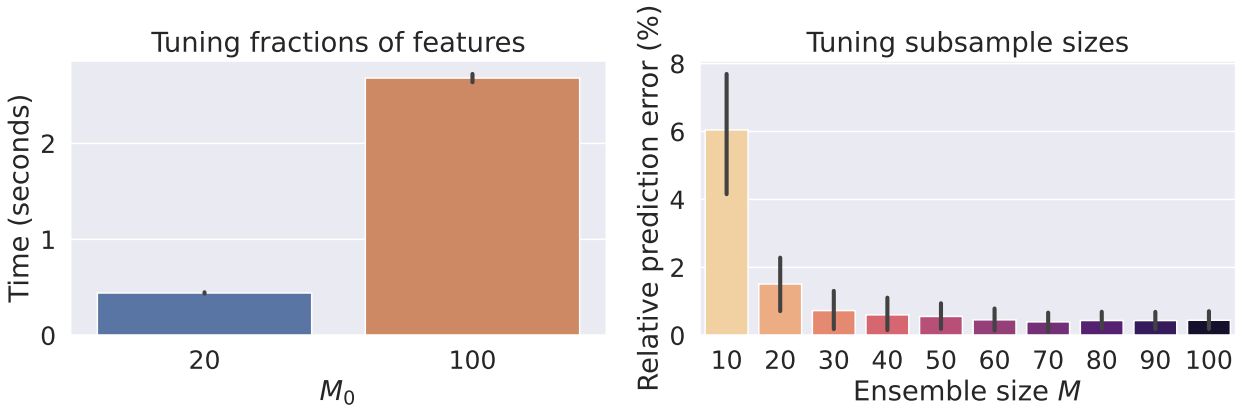

Figure S14: Tuning factions of features, subsample size by ECV. The left panel shows that computational time with extrapolation ($M_0 = 20$) and without extrapolation ($M_0 = 100$). The right panel shows the relative prediction error (the difference between the prediction errors without and with subsample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

`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.

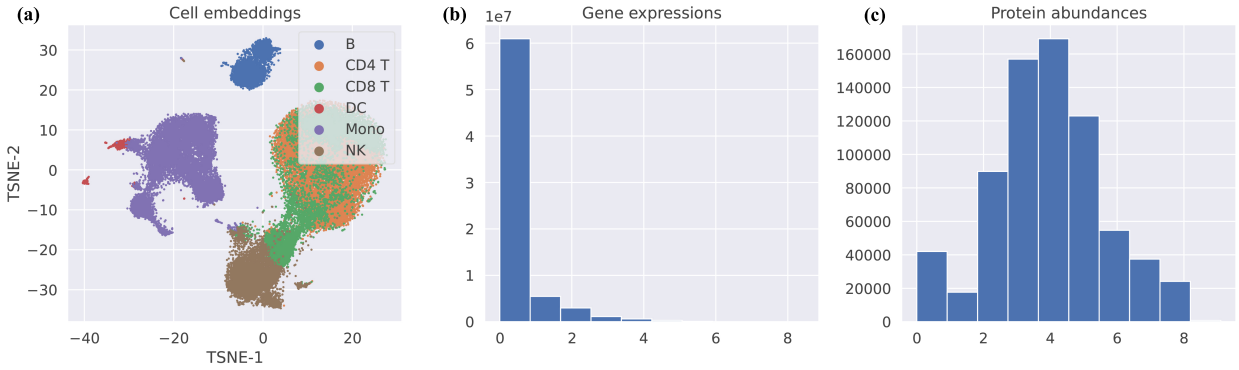

Figure S15: Overview of the single-cell CITE-seq dataset from Hao et al. (2021). (a) The 2-dimensional tSNE cell embeddings. (b-c) The histogram of overall log-normalized gene expressions and protein abundances in the Mono cell type.

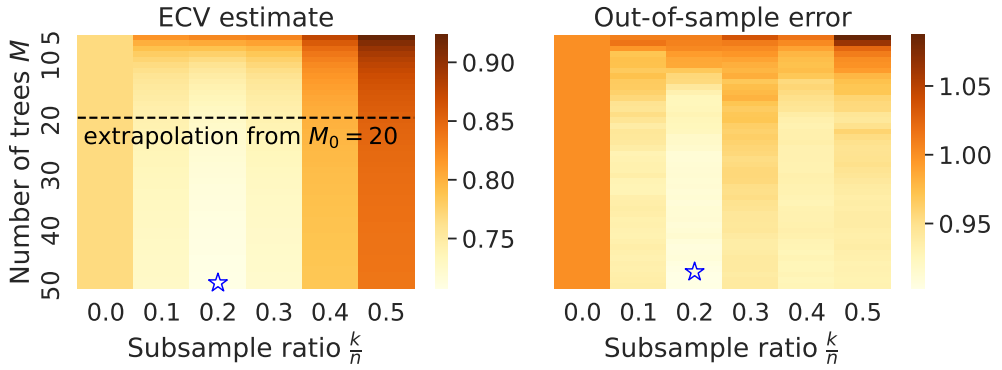

Figure S16: Heatmap of ECV performance on predicting the abundances of surface protein CD103 in the DC cell type by random forests (subagging). The left and right panels show the NMSEs of OOB risk estimates and out-of-sample prediction risk, respectively. The values are normalized by the empirical variance 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.

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

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 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.

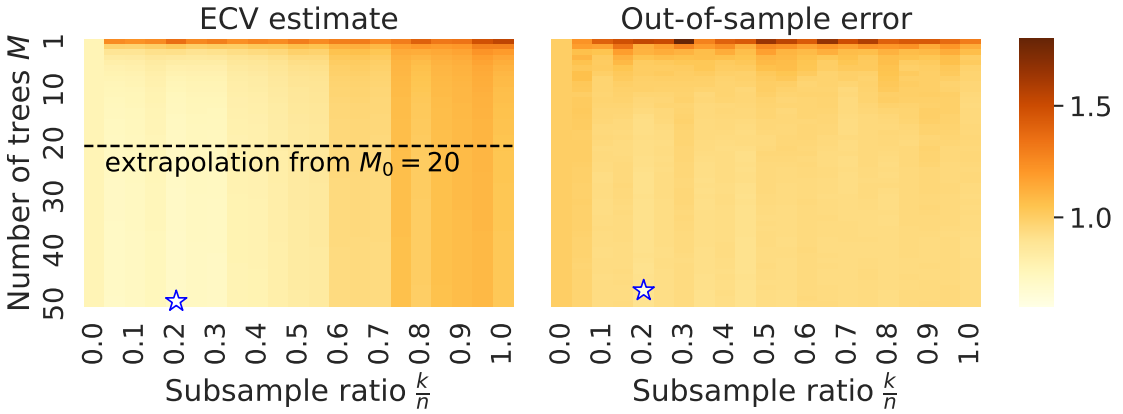

Figure S17: Heatmap of ECV performance on predicting the abundances of surface protein CD103 in the DC cell type by random forests (subagging) as in Fig. S16 with $M \in \{1, \ldots, 50\}$ and k/n ranging from 0 to 1 displayed.

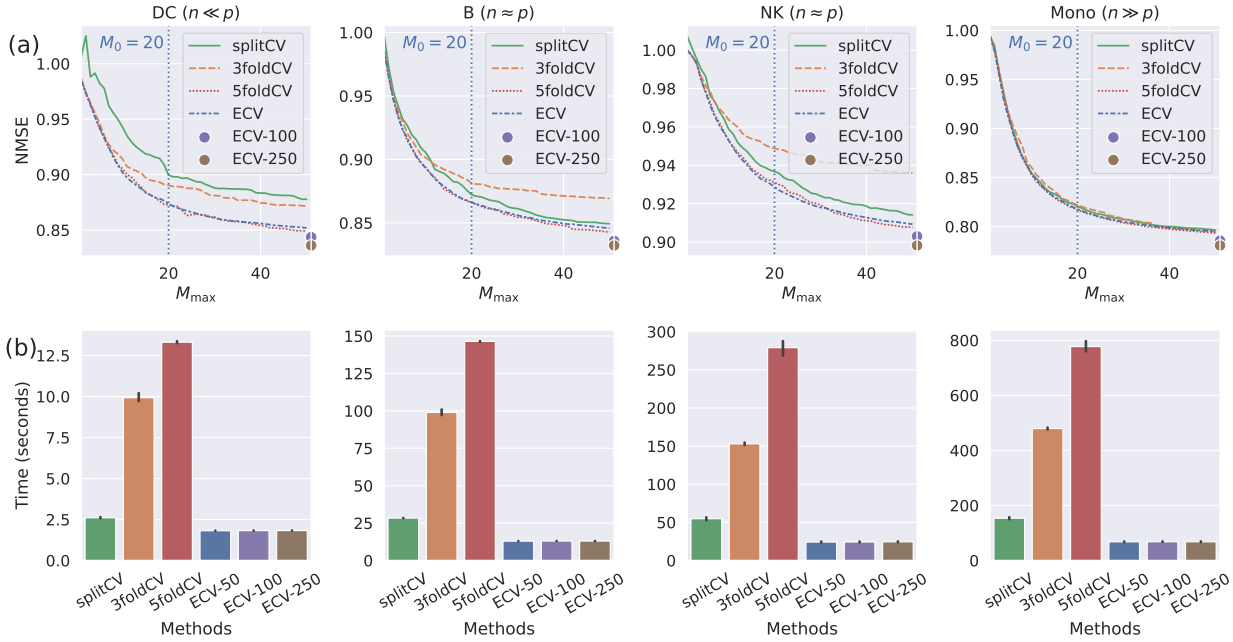

Figure S18: Performance of cross-validation methods on predicting the protein abundances in different cell types. (a) The average normalized mean squared error (NMSE) of the cross-validated predictors for different methods. ECV (bagging) uses $M_0 = 20$ trees to extrapolate risk estimates, and the dashed 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.