Scalable subsampling: computation, aggregation and inference

By Dimitris N. Politis
Department of Mathematics
and Halicioglu Data Science Institute
University of California, San Diego
La Jolla, CA 92093-0112, USA
dpolitis@ucsd.edu

Summary

Subsampling has seen a resurgence in the *Big Data* era where the standard, full-resample size bootstrap can be infeasible to compute. Nevertheless, even choosing a single random subsample of size b can be computationally challenging with both b and the sample size n being very large. The paper at hand shows how a set of appropriately chosen, non-random subsamples can be used to conduct effective—and computationally feasible—subsampling distribution estimation. Furthermore, the same set of subsamples can be used to yield a procedure for subsampling aggregation—also known as subagging—that is scalable with big data. Interestingly, the scalable subagging estimator can be tuned to have the same (or better) rate of convergence as compared to $\hat{\theta}_n$. Statistical inference could then be based on the scalable subagging estimator instead of the original $\hat{\theta}_n$.

Some key words: Bagging, Big Data, bootstrap, distributed inference, subagging.

1 Introduction

Assume data $X_1, \ldots, X_n$ that are independent, identically distributed (i.i.d.) taking values in an arbitrary space. Often, this space will be $\mathbf{R}^d$ but other choices exist, e.g., it can be a function space, a space of networks, etc. A statistic $\hat{\theta}_n = T_n(X_1, \ldots, X_n)$ is employed to estimate a parameter θ associated with the common distribution of the data. Assume that θ takes values on a normed linear space $\mathbf{\Theta}$ with norm denoted by $||\cdot||$.

Let $J_n(x) = P\{\tau_n g(\hat{\theta}_n - \theta) \leq x\}$ where the rate of convergence τ_n diverges to infinity as n increases. We will generally assume that the real-valued function $g(\cdot)$ has the properties of a norm on $\mathbf{\Theta}$ with one interesting exception: if $\mathbf{\Theta} = \mathbf{R}$, then $g(\cdot)$ could be taken to be the identity function, leading to one-sided inference. In the general case where $g(\cdot)$ is indeed a norm, note that it can be a different norm than $||\cdot||$. For example, if $\mathbf{\Theta} = \mathbf{R}^p$ equipped with Euclidean norm, it is often useful to let $g(\cdot)$ be the sup-norm on $\mathbf{R}^p$; estimating the quantiles of J_n would then lead to simultaneous confidence intervals and/or simultaneous hypothesis tests for all p coordinates of θ. We will also assume:

Assumption A. *There exists a non-degenerate probability distribution J, such that $J_n(x) \to J(x)$ as $n \to \infty$ for all x points at which $J(x)$ is continuous.*

Subsampling is a general statistical method developed in the 1990s aimed at estimating the sampling distribution J_n in order to conduct nonparametric inference such as the construction of confidence intervals and hypothesis tests; see Politis, Romano and Wolf (1999) and the references therein. To describe it, let the subsample size b be a positive integer less than n, and consider all the subsets of size b of the sample $X_1, \ldots, X_n$. There are $Q = \binom{n}{b}$ such subsets that can be ordered in an arbitrary fashion and denoted by $\mathcal{B}_j$ for $j = 1, \ldots, Q$.

Compute the subsample statistics $\hat{\theta}_{b,j} = T_b(\mathcal{B}_j)$ for $j = 1, \ldots, Q$, and subsampling distribution $L_{n,b}(x) = Q^{-1} \sum_{i=1}^{Q} 1\{\tau_b g(\hat{\theta}_{b,i} - \hat{\theta}_n) \leq x\}$. Under Assumption A and the additional conditions

$$n \to \infty \ \text{ and } \ b \to \infty \ \text{ but with } \ b/n \to 0 \tag{1}$$

$$\text{and } \ \tau_b/\tau_n \to 0 \tag{2}$$

it was shown that $L_{n,b}(x) \xrightarrow{P} J(x)$ for all x points of continuity of J, where $\xrightarrow{P}$ denotes convergence in probability; see Theorems 3.1 and 3.2 of Politis and Romano (1994). Note that $L_{n,b}(x) \xrightarrow{P} J(x)$ implies that $L_{n,b}(x) - J_n(x) \xrightarrow{P} 0$, i.e., $L_{n,b}(x)$ can be used as a proxy for the unknown $J_n(x)$ so as to conduct statistical inference based on $\hat{\theta}_n$. Also note that if the rate of convergence satisfies

$$\tau_n = n^{\alpha} \mathcal{L}(n) \ \text{ for some } \ \alpha > 0. \tag{3}$$

for some slowly varying function $\mathcal{L}(\cdot)$ such that $\lim_{n\to\infty} \frac{\mathcal{L}(sn)}{\mathcal{L}(n)} = 1$ for any $s > 0$, then eq. (2) follows.

It was recognized early on that if n is large, it is not realistic to compute $\hat{\theta}_{b,j}$ for $j = 1, \ldots, Q$ since Q can be astronomically large. For that reason, Corollary 2.1 of Politis and Romano (1994) showed that a stochastic approximation to $L_{n,b}(x)$ can be used instead. The stochastic approximation relies on B randomly chosen subsamples from the set $\{\mathcal{B}_j,\ j = 1, \ldots, Q\}$ with B tending to infinity; in practice, B is taken large enough so that the error of the stochastic approximation is negligible.

Subsampling has seen a resurgence in the *Big Data* era of the 21st century where the standard, full-resample size bootstrap can be infeasible to compute; see e.g., Jordan (2013), Kleiner et al. (2014), and Sengupta et al. (2016). Nevertheless, even choosing a single random subsample of size b can be computationally challenging. As pointed out in Ting (2021), drawing a random sample of size b from n items using the Sparse Fisher-Yates Sampler takes $O(b)$ time and space which corresponds to optimal time and space complexity for this problem. To perform subsampling inference, we need to generate B such subsamples each of size b. Therefore, the computational cost of just drawing the random subsamples is of order $O(bB)$ where both b and B are meant to tend to infinity. Approximate solutions such as Poisson sampling are often used instead; see Bertail et al. (2017).

In the next section we show how a set of appropriately chosen, non-random subsamples can be used to conduct effective—and computationally feasible—distribution estimation via subsampling. In Section 4 we show how the same set of subsamples can be used to yield a procedure for subsampling aggregation—also known as subagging—that is scalable in an attempt to remedy computability issues discussed in Section 3. Interestingly, the scalable subagging estimator can be tuned to have the same (or better) rate of convergence as compared to $\hat{\theta}_n$. The paper is concluded by providing details on how to conduct inference, e.g. confidence intervals, based on the scalable subagging estimator instead of the original $\hat{\theta}_n$. Some numerical illustrations are presented in the online Supplement.

2 Scalable subsampling distribution estimation

Recall the set of all size b subsamples $\{\mathcal{B}_j,\ j = 1, \ldots, Q\}$, and re-arrange it so that the first subsamples are obtained as blocks of consecutive data points, i.e., $\mathcal{B}_j = (X_{(j-1)h+1}, X_{(j-1)h+2}, \ldots, X_{(j-1)h+b})$. Recall that the block size b is an integer in $[1, n]$, and so is h; in particular, h controls the amount of overlap (or separation) between $\mathcal{B}_j$ and $\mathcal{B}_{j+1}$. If $h = 1$, then the overlap is the maximum possible; if $h \sim 0.2\ b$, then there is an approximate 80% overlap between $\mathcal{B}_j$ and $\mathcal{B}_{j+1}$; if $h = b$, then there is *no* overlap between $\mathcal{B}_j$ and $\mathcal{B}_{j+1}$ but these two blocks are adjacent; finally, if $h \sim 1.2\ b$, then there is a block of about $0.2\ b$ data points from the data sequence $X_1, \ldots, X_n$ that separate the blocks $\mathcal{B}_j$ and $\mathcal{B}_{j+1}$. In general, b and h are functions of n, but these dependences will not be explicitly denoted.

The collection of all available block–subsamples of size b, is then $\{\mathcal{B}_j,\ j = 1, \ldots, q\}$ where $q = \lfloor (n-b)/h \rfloor + 1$; here, $\lfloor \cdot \rfloor$ and $\lceil \cdot \rceil$ denote the floor and ceiling functions respectively. We claim that this non-random collection is sufficient for effective and computationally feasible subsampling distribution estimation. To see why, note that subsampling using the aforementioned block–subsamples has been found to be consistent in the setting where the data sequence $X_1, \ldots, X_n$ is a finite stretch of a strong-mixing, stationary time series; see e.g. Politis and Romano (1994, Section 3.2). Since the i.i.d. case can be considered as a special case of a stationary time series, the claim follows.

To elaborate, we define the subsample statistics $\hat{\theta}_{b,j} = T_b(\mathcal{B}_j)$ for $j = 1, \ldots, q$ as before, and construct a new subsampling distribution as $L_{n,b,h}(x) = q^{-1} \sum_{i=1}^{q} 1\{\tau_b g(\hat{\theta}_{b,i} - \hat{\theta}_n) \leq x\}$.

Proposition 2.1 *Assume Assumption A, and conditions (1) and (3). Also assume that either $h = 1$, or that h satisfies*

$$h \sim c_1\ b \quad \textit{for some constant} \quad c_1 > 0. \tag{4}$$

Then, $L_{n,b,h}(x) \xrightarrow{P} J(x)$ for all x points of continuity of J.

The Proposition follows from Corollary 3.2 of Politis and Romano (1994) who worked under the assumption that $1 \leq h \leq b$; the case where $h > b$ —but still with $h = O(b)$— can be proven in a similar way. The essence of the argument is that

$$EL_{n,b,h}(x) \approx J_b(x) \to J(x) \quad \text{as} \quad b \to \infty \tag{5}$$

where x a point of continuity of J. In addition,

$$Var(L_{n,b,h}(x)) = O(b/n). \tag{6}$$

Eq. (5) and (6) together with Chebyshev's inequality imply $L_{n,b,h}(x) \xrightarrow{P} J(x)$.

Note that the bound (6) holds true regardless as to whether $h = 1$ or h satisfies condition (4); it is just the proportionality constant in $O(b/n)$ that becomes smaller (but is bounded below) as h decreases. Therefore, for reasons of parsimony and computational tractability, we will not propose using full-overlap block-subsamples, i.e., the case $h = 1$, in what follows. Instead we will work under condition (4), in which case $q = O(n/b)$. Hence, assuming $\hat{\theta}_n$ can be computed in $O(n^\zeta)$ operations (for some constant $\zeta > 0$), the construction of $L_{n,b,h}$ and its quantiles has computational complexity $O(n^\zeta) + O(nb^{\zeta-1}) = O(n^\zeta)$ which is the same as computing the statistic $\hat{\theta}_n$ itself; therefore, we may call the construction of $L_{n,b,h}$ under condition (4) as being *scalable*.

Implementation problems might ensue when $\zeta > 1$ and n is large, but in this case even the computation of the original statistic $\hat{\theta}_n$ may be problematic; we will address this issue next.

3 Computability issues

As discussed in the last section, the statistic $\hat{\theta}_n$ will generally be computable in $O(n^\zeta)$ operations. If ζ is small, then no issues incur. Unfortunately, examples abound with $\zeta > 1$, making the computability of $\hat{\theta}_n$ questionable in the Big Data era. Some examples are as follows:

1. The X_i are univariate, and $\hat{\theta}_n$ is the sample mean (or median) of $X_1, \ldots, X_n$. Then, $\zeta = 1$.

2. The X_i take values in $\mathbf{R}^d$, and $\hat{\theta}_n$ is the sample mean of $X_1, \ldots, X_n$. Then, $\hat{\theta}_n$ is computable in $O(dn)$ operations. If d is a constant, then $\zeta = 1$ as above. However, it may be that d grows with n; if d grows linearly with n, then $\zeta = 2$.

3. Suppose that $X_i = (Y_i, W_i)$ where Y_i is the univariate response associated with regressor W_i that takes values in $\mathbf{R}^d$; this is the standard regression situation. If d is large, then LASSO regression can be employed; see Tibshirani (1996). A popular method to compute the LASSO has computational complexity $O(d^3 + d^2 n)$ as long as $d < n$; see Efron et al. (2004). If d grows linearly with n, e.g., when $d \sim n/2$, then $\zeta = 3$.

Remark 3.1 There is a growing body of work dealing with the possibility that the sample size n is so large that it may not be feasible to compute $\hat{\theta}_n$. One branch of this literature is devoted to 'optimal subsampling' whose meaning is different than the subsampling-based inference discussed so far. In a nutshell, if $\hat{\theta}_n$ is not computable, one can use just one of the subsample statistics $\hat{\theta}_{b,i}$ to estimate θ. The question 'which one to use' is tantamount to 'optimal subsampling'; see Yao and Wang (2021) for a review. The problem with this approach is that the practitioner is effectively throwing out most of the data. A Divide-and-Conquer alternative is proposed in the next section.

4 Scalable subsampling aggregation

4.1 Computation

Subsample aggregation, also known as *subagging*, was proposed by Bühlmann and Yu (2002). In the context of the present paper, the subagging estimator can be written as $\bar{\theta}_{b,SA} = Q^{-1} \sum_{i=1}^{Q} \hat{\theta}_{b,i}$ for an appropriate choice of b; here, and for the remainder of the paper, we will assume a univariate θ, i.e., $\mathbf{\Theta} = \mathbf{R}$, and $g(x) = x$. Under regularity conditions, Bühlmann and Yu (2002) showed that

$$E\bar{\theta}_{b,SA} = E\hat{\theta}_{b,1}, \quad \text{and} \quad Var(\bar{\theta}_{b,SA}) \leq (b/n) Var(\hat{\theta}_{b,1}).$$

Hence, if the Bias of $\hat{\theta}_{b,1}$ is tolerable, subagging yields a welcome variance reduction.

With Big Data it is, of course, infeasible to compute (and average) all $Q = \binom{n}{b}$ subsample statistics. By analogy to the stochastic approximation to $L_{n,b}(x)$, Zou et al. (2021) proposed the use of randomly chosen subsamples to compute $\bar{\theta}_{b,SA}$, and provided two algorithms for implementation under constraints in computer memory. However, as already mentioned, choosing B random subsamples of size b presents computational difficulties when b and n are large.

Observe that although $\bar{\theta}_{b,SA}$ is an average of Q values, the variance is reduced by dividing by n/b not Q. The reason is that the Q subsamples have typically high overlap; hence their associated subsample statistics are highly dependent. We can achieve a similar variance reduction effect by using just the first q subsamples in the ordering described in Section 2, i.e., $\mathcal{B}_j = (X_{(j-1)h+1}, X_{(j-1)h+2}, \ldots, X_{(j-1)h+b})$ for $j = 1, \ldots, q$. We therefore define the *scalable subagging* estimator as $\bar{\theta}_{b,n,SS} = q^{-1} \sum_{i=1}^{q} \hat{\theta}_{b,i}$.

Proposition 4.1 *Assume condition (4), and that $E\hat{\theta}_n^2 < \infty$ for all n. Then,*

$$E\bar{\theta}_{b,n,SS} = E\hat{\theta}_{b,1}, \quad \textit{and} \quad Var(\bar{\theta}_{b,n,SS}) \leq \frac{2m-1}{q} Var(\hat{\theta}_{b,1}) \tag{7}$$

where $m = \lceil b/h \rceil$ and $q = \lfloor (n-b)/h \rfloor + 1$. If $h \geq b$, then $Var(\bar{\theta}_{b,n,SS}) = q^{-1} Var(\hat{\theta}_{b,1})$.

Proof. Note that $E\hat{\theta}_{b,j} = E\hat{\theta}_{b,1}$ and $Var(\hat{\theta}_{b,j}) = Var(\hat{\theta}_{b,1})$ for all j. The Cauchy-Schwarz inequality yields $|Cov(\hat{\theta}_{b,i}, \hat{\theta}_{b,j})| \leq Var(\hat{\theta}_{b,1})$ but $\hat{\theta}_{b,i}$ is independent of $\hat{\theta}_{b,j}$ when $|i-j| \geq m$ in which case $Cov(\hat{\theta}_{b,i}, \hat{\theta}_{b,j}) = 0$. Plugging these estimates into the expression $Var(\bar{\theta}_{b,n,SS}) = q^{-2} \sum_{i=1}^{q} \sum_{i=j}^{q} Cov(\hat{\theta}_{b,i}, \hat{\theta}_{b,j})$ shows (7); $\hat{\theta}_{b,1}, \ldots, \hat{\theta}_{b,q}$ are i.i.d. when $h \geq b$, and the result is sharper.$\diamond$

Remark 4.1 Since $q = O(n/b)$, we can compute $\bar{\theta}_{b,n,SS}$ with $O(qb^{\zeta}) = O(nb^{\zeta-1})$ operations, a significant saving over the $O(n^{\zeta})$ needed for $\hat{\theta}_n$. Recall the discussion of Remark 3.1 on (i) identifying the 'optimal subsample' denoted by $\mathcal{B}_{j^*}$, and then (ii) using as your final estimator the subsample statistic $\hat{\theta}_{b,j^*}$. The computational cost of part (ii) is of course $O(b^{\zeta})$; if you add to this the nontrivial computational cost of part (i) —that may require numerical optimization— the overall complexity may well exceed the $O(nb^{\zeta-1})$ needed to compute the scalable subagging estimator $\bar{\theta}_{b,n,SS}$.

4.2 Rate of convergence and choice of b for scalable subagging

If the computational complexity of $\bar{\theta}_{b,n,SS}$ is $O(nb^{\zeta-1})$, what is to stop us from taking b very small, even $b = 1$, to make it $O(n)$? The answer is the generally nonegligible bias of $\hat{\theta}_{b,1}$ that is inherited by $\bar{\theta}_{b,n,SS}$ as Proposition 4.1 showed. Consider the following three *Bias Conditions* (BC):

(I) Estimator $\hat{\theta}_n$ is exactly unbiased, as is the case with linear statistics; see Example 1.4.1 of Politis et al. (1999). Then, b can be taken to equal one but, of course, scalable subagging is not needed here as the computational complexity of $\hat{\theta}_n$ is $O(n)$ already.

(II) Estimator $\hat{\theta}_n$ is asymptotically unbiased, and its bias is asymptotically negligible even after multiplication by τ_n; in other words, the limit law J of Assumption A is centered at zero.

(III) Estimator $\hat{\theta}_n$ is asymptotically unbiased but its bias does not vanish after multiplication by τ_n; in other words, the limit law J of Assumption A is centered at a nonzero value.

Although the subsampling distribution estimator $L_{n,b,h}$ can work under all above eventualities — including BC case (III)—, to investigate the rate of convergence of scalable subagging we will work under the assumption of BC case (II). Note that an estimator falling under BC case (III) could be analytically *debiased* —by subtracting from it a consistent estimate of its bias—, allowing the debiased estimator to be handled under BC case (II). Hence, we formulate the following assumption:

Assumption B. *Assume that* $E\hat{\theta}_n^2 < \infty$ *for all* n, *and that* $\tau_n = n^{\alpha}$ *for some constants* $\gamma > \alpha > 0$, $C \in \mathbf{R} - \{0\}$, *and* $\sigma^2 > 0$ *such that* $n^{\gamma}(E\hat{\theta}_n - \theta) \to C$, *and* $Var(\tau_n\hat{\theta}_n) \to \sigma^2$ *as* $n \to \infty$.

In the above, we have simplified eq. (3) by omitting the slowly varying function $\mathcal{L}(n)$. Note that $\gamma > \alpha$ implies that the bias of $\hat{\theta}_n$ is negligible even after multiplication by τ_n as in BC case (II). Politis (2021) considered the possibility that $h/b \to \infty$ but reasons of efficiency of $\bar{\theta}_{b,n,SS}$ point towards adopting condition (4) as in Section 2. To this end, we will assume:

$$b \sim c_2 n^{\beta} \text{ and } h \sim c_3 n^{\beta} \text{ as } n \to \infty \text{ for positive constants } c_2, c_3, \text{ and constant } 0 < \beta < 1. \quad (8)$$

The following lemma shows that $\bar{\theta}_{b,n,SS}$ can be tuned to have the same (or better) rate of convergence as compared to $\hat{\theta}_n$. We will use the notation $a_n = \Theta(d_n)$ to denote 'exact order', i.e., that there exist constants $\underline{c}, \bar{c}$ satisfying $\underline{c} \cdot \bar{c} > 0$, and such that $\underline{c} d_n \le a_n \le \bar{c} d_n$.

Lemma 4.1 *Assume Assumption B, eq. (8), and* $\alpha \le 1/2$. *Choose a value of* β *satisfying*

$$\frac{1}{1+2(\gamma-\alpha)} \le \beta < \frac{1}{2(\gamma-\alpha)}. \quad (9)$$

(i) $MSE(\bar{\theta}_{b,n,SS}) = \Theta(n^{\beta-1-2\alpha\beta}) = O(\tau_n^{-2})$ *where MSE is short for Mean Squared Error.*
(ii) If $\beta > \frac{1}{1+2(\gamma-\alpha)}$, *then* $[Bias(\bar{\theta}_{b,n,SS})]^2 = o\big(Var(\bar{\theta}_{b,n,SS})\big)$, *i.e.,* $\bar{\theta}_{b,n,SS}$ *falls under BC case (II).*

(iii) If $\beta = \frac{1}{1+2(\gamma-\alpha)}$*, then* $[Bias(\bar{\theta}_{b,n,SS})]^2 = \Theta\left(Var(\bar{\theta}_{b,n,SS})\right)$*, i.e.,* $\bar{\theta}_{b,n,SS}$ *falls under BC case (III). Furthermore, the choice* $\beta = \frac{1}{1+2(\gamma-\alpha)}$ *minimizes the* $MSE(\bar{\theta}_{b,n,SS})$*, yielding*

$$MSE(\bar{\theta}_{b,n,SS}) = \Theta(n^{-2\gamma/[1+2(\gamma-\alpha)]}) \tag{10}$$

in which case the optimized rate of convergence of $\bar{\theta}_{b,n,SS}$ *is* $n^{\gamma/[1+2(\gamma-\alpha)]}$*.*

The proof is omitted as straightforward; some details are given in Proposition 4.2 of Politis (2021) under the simplifying condition $h \geq b$. Note that letting β equal the lower bound $\frac{1}{1+2(\gamma-\alpha)}$ kills two birds with one stone: (a) optimizes the rate of convergence of $\bar{\theta}_{b,n,SS}$, and (b) minimizes the computational complexity in computing $\bar{\theta}_{b,n,SS}$ making it $O(nb^{\zeta-1}) = O(n^{1+\beta(\zeta-1)}) = O\left(n^{1+\frac{(\zeta-1)}{1+2(\gamma-\alpha)}}\right)$.

We now discuss some examples:

1. **Linear statistics.** Consider a linear statistic $\hat{\theta}_n$ that is represented as $\hat{\theta}_n = n^{-1}\sum_{i=1}^n G(X_i)$ for some appropriate function G. Note that $\hat{\theta}_n$ estimates $\theta = EG(X_1)$, and is exactly unbiased for that. As mentioned under the description of BC case (I), subagging is not needed here because $\hat{\theta}_n$ is easily computed. Furthermore, this case can not really fit under the premises of Lemma 4.1 since Assumption B does not hold; recall that Assumption B implies that $\hat{\theta}_n$ has nonzero bias. We can intuit what would happen here by pretending that Assumption B holds (approximately) with a huge value of γ. Letting $\gamma \to \infty$ implies that the optimal β tends to zero. Hence, one would take $b = h = 1$, reducing $\bar{\theta}_{b,n,SS}$ to the original statistic $\hat{\theta}_n$.

2. **Approximately linear statistics.** A statistic $\hat{\theta}_n$ can be called approximately linear if it can be represented as $\hat{\theta}_n = n^{-1}\sum_{i=1}^n G(X_i) + o_P(n^{-1/2})$. Then, $\hat{\theta}_n$ is $\sqrt{n}$–consistent for $\theta = EG(X_1)$, i.e., $\alpha = 1/2$. Examples include the sample median and other sample quantiles, trimmed means, M-estimators, etc. In many such examples, it may often be verified that the bias of $\hat{\theta}_n$ is of order $1/n$, i.e., $\gamma = 1$. Part (iii) of Lemma 4.1 suggests that the choice $\beta = 1/2$ minimizes the $MSE(\bar{\theta}_{b,n,SS})$; with this choice, $\bar{\theta}_{b,n,SS}$ is $\sqrt{n}$–consistent as well.

3. **Nonparametric function estimators.** Consider the case where θ represents the value of function f at a point of interest; the function f can be a probability density, spectral density, or other function that should be estimated in a nonparametric setting. Let $\hat{\theta}_n$ denote a kernel-smoothed estimator of θ, and suppose that a nonnegative kernel is used. In this case, the MSE–optimal bandwidth is $\Theta(n^{-1/5})$. However, this bandwidth choice brings $\hat{\theta}_n$ under the realm of BC case (III), and the premises of Lemma 4.1 do not apply. As an experiment, consider a degree of *undersmoothing* in constructing $\hat{\theta}_n$. To fix ideas, suppose the bandwidth is chosen to be $\Theta(n^{-1/4})$ instead, yielding $Bias(\hat{\theta}_n) = O(n^{-1/2})$ and $Var(\hat{\theta}_n) = \Theta(n^{-3/4})$; in this case, $\alpha = 3/8$ and $\gamma = 1/2$. According to Lemma 4.1, β should be optimally chosen to equal 0.8; hence, the rate of convergence of $\bar{\theta}_{b,n,SS}$ becomes $n^{2/5}$. This rate is not only faster than the rate of $\hat{\theta}_n$ that used the suboptimal bandwidth $\Theta(n^{-1/4})$; it is actually the fastest rate achievable by *any* estimator that uses a nonnegative kernel with its associated MSE–optimal bandwidth. Nevertheless, $\bar{\theta}_{b,n,SS}$ can be computed faster than $\hat{\theta}_n$, and may thus be preferable.

The above example opens up the possibility that $\bar{\theta}_{b,n,SS}$ may be more efficient than $\hat{\theta}_n$ (on which it is based). Such "super-efficiency" was first pointed out by Banerjee et al. (2019) in the setting of isotonic regression but it is a more general phenomenon, associated with harder estimation problems. Some numerical illustrations to that effect are presented in the online Supplement.

4.3 Inference beyond point estimation

Having established that $\bar{\theta}_{b,n,SS}$ is a consistent estimator whose rate of convergence towards θ is fast (and sometimes faster than that of $\hat{\theta}_n$), the question now is how to conduct inference, e.g., confidence intervals, hypothesis tests, etc. based on $\bar{\theta}_{b,n,SS}$. If we take $h \geq b$, then $\hat{\theta}_{b,1}, \ldots, \hat{\theta}_{b,q}$ are i.i.d. but not a simple sequence of random variables; to see that, note that the value of $\hat{\theta}_{b,1}$ changes with b (which increases with n). Rather, $\hat{\theta}_{b,1}, \ldots, \hat{\theta}_{b,q}$ can be thought as the nth row of a triangular array with i.i.d. entries, and common distribution given (approximately, and after centering and standardizing) by J_b. Since $\bar{\theta}_{b,n,SS}$ is the sample mean of $\hat{\theta}_{b,1}, \ldots, \hat{\theta}_{b,q}$, a Central Limit Theorem (CLT) for triangular arrays —such as Theorem B.O.1 of Politis et al. (1999)—is helpful.

Corollary 4.1 *Assume the premises of Lemma 4.1. Also assume that there exist positive numbers ϵ and Δ such that $E|\hat{\theta}_n|^{2+\epsilon} \leq \Delta < \infty$ for any n. Then, for some $\sigma_0^2 > 0$, we have*

$$\kappa_n \left(\bar{\theta}_{b,n,SS} - \theta\right) \stackrel{\mathcal{L}}{\Longrightarrow} N(C_\beta, \sigma_0^2) \quad as \quad n \to \infty \tag{11}$$

where $\kappa_n = n^{\frac{-1+\beta-2\alpha\beta}{2}}$ and $\stackrel{\mathcal{L}}{\Longrightarrow}$ denotes convergence in law. Furthermore: (i) if $\beta > \frac{1}{1+2(\gamma-\alpha)}$, then $C_\beta = 0$; (ii) if $\beta = \frac{1}{1+2(\gamma-\alpha)}$, then $C_\beta = C$ as defined in Assumption B.

If $h < b$, then $\hat{\theta}_{b,1}, \ldots, \hat{\theta}_{b,q}$ are not independent; rather, they are m-dependent with $m = \lceil b/h \rceil$ as in Proposition 4.1 but Theorem B.O.1 of Politis et al. (1999) still applies. If $h \geq b$, then $\sigma_0^2 = \sigma^2$ as defined in Assumption B. Hence, if $\beta > \frac{1}{1+2(\gamma-\alpha)}$, then all that is needed is a consistent estimator of σ^2; this is easy to obtain as the sample variance of the nth row of the triangular array, i.e., letting $\hat{\sigma}^2 = b^{2\alpha} q^{-1} \sum_{i=1}^{q} \left(\hat{\theta}_{b,i} - \bar{\theta}_{b,n,SS}\right)^2$. Therefore, when $h \geq b$, an approximate 95% confidence interval for θ under case (i) would be $\bar{\theta}_{b,n,SS} \pm 1.96\ \hat{\sigma} \cdot n^{\frac{1-\beta+2\alpha\beta}{2}}$.

However, case (ii) of Corollary 4.1 is more interesting since it ensures the fastest rate of convergence of $\bar{\theta}_{b,n,SS}$. Here, the nontrivial asymptotic bias of the distribution of $\bar{\theta}_{b,n,SS}$ presents some difficulties at first inspection. Nevertheless, subsampling comes again to the rescue since eq. (11) shows that $\bar{\theta}_{b,n,SS}$ has a well-defined asymptotic distribution; the fact that the latter is not centered at zero is immaterial. In other words, $\bar{\theta}_{b,n,SS}$ satisfies Assumption A with $\bar{\theta}_{b,n,SS}$ in place of $\hat{\theta}_n$. Hence, the scalable subsampling construction of Section 2 can be applied to yield a consistent estimate of the distribution of estimator $\bar{\theta}_{b,n,SS}$; see Remark 4.3 for details on iterated subsampling.

Remark 4.2 (Connections with distributed inference.) The case $h = b$, i.e., splitting the sample into q non-overlapping parts, is closely related to the classical notion of q-fold cross-validation, as well as the more recent notion of Divide-and-Conquer (DaC) methods; see Jordan (2013). To elaborate, the scalable subagging estimator $\bar{\theta}_{b,n,SS}$ has been studied before in the following DaC distributed inference contexts (all with $h = b$): U-statistics by Lin and Xi (2010); generalized linear models by Chen and Xie (2014); M-estimators by Zhang, Duchi and Wainwright (2013); and a certain class of symmetric statistics (that includes L-statistics and smoothed functions of the sample mean) by Chen and Peng (2021). Note also that Bradic (2016) employed subagging using non-overlapping blocks of data, and applied it to variable selection in large-scale regression. The current section is meant to serve many purposes. One is to show that these ideas are universally applicable under minimal assumptions, such as Assumption B; for example, the asymptotic normality results of Chen and Peng (2021) actually follow from our Corollary 4.1 simply by checking its premises. Furthermore, it is important to note that the usefulness of scalable subagging and DaC distributed inference extends well beyond the realm of asymptotically linear, $\sqrt{n}$–consistent statistics that have been

considered so far; see our Section 4.2 including the example on nonparametric function estimation. Finally, the interplay of the tuning parameters h and b opens up interesting possibilities, e.g., the possibility that $\bar{\theta}_{b,n,SS}$ has a faster rate of convergence than $\hat{\theta}_n$ itself; see the aforementioned paper by Banerjee et al. (2019) who also proved a CLT like eq. (11) under a different set of assumptions.

4.4 Weakly dependent data

All subsampling constructions in this paper, including the scalable subsampling distribution $L_{n,b,h}(x)$ and scalable subagging estimator $\bar{\theta}_{b,n,SS}$, remain valid if there is (weak) dependence in the data, i.e., if $X_1, \ldots, X_n$ are a stretch of a strictly stationary, strong mixing time series. The reason is that the choice of block-subsamples described in Section 2 and used throughout the paper is actually the choice that is recommended in order to subsample time series; see e.g. Politis and Romano (1994). Hence, all results in Section 4.2 remain true as stated in the case where $X_1, \ldots, X_n$ are weakly dependent but some of the discussion in Section 4.3 may require a little tweak.

To see why, note that if $h = b$, then $\hat{\theta}_{b,1}, \ldots, \hat{\theta}_{b,q}$ will be independent only if the data $X_1, \ldots, X_n$ are independent. If $X_1, \ldots, X_n$ are stationary and strong mixing, we can still ensure that $\hat{\theta}_{b,1}, \ldots, \hat{\theta}_{b,q}$ are approximately independent if we require $h - b \to \infty$, e.g., $h = b + \lfloor \sqrt{b} \rfloor$; this would ensure that blocks $\mathcal{B}_j$ and $\mathcal{B}_{j+1}$ are separated by about $\sqrt{b}$ data points, rendering them approximately independent as b increases with n. However, Theorem B.O.1 of Politis et al. (1999)—on which Corollary 4.1 was based—holds true even when $\hat{\theta}_{b,1}, \ldots, \hat{\theta}_{b,q}$ are weakly dependent, e.g., when $h \sim c_1 b$ with $c_1 \leq 1$, and so does subsampling distribution estimation (based on block-subsamples).

Corollary 4.2 *Assume $X_1, \ldots, X_n$ is a stretch of a strictly stationary time series with exponentially decreasing strong mixing coefficients. Then, Lemma 4.1 and Corollary 4.1 remain true as stated.*

Note that if $h - b \to \infty$, then $\sigma_0^2 = \sigma^2$ as defined in Assumption B. If $h \sim c_1\, b$ with $c_1 \leq 1$, and also to work under case (ii) of Corollary 4.1, we need the aforementioned idea of *iterated subsampling*.

Remark 4.3 (Iterated subsampling) Let $b = \lfloor c_2 n^{\beta} \rfloor$ and apply scalable subsampling aggregation (SSA) to $X_1, \ldots, X_n$ to compute $\bar{\theta}_{b,n,SS}$. (a) Let $b' = \lfloor c_2 b^{\beta} \rfloor$ and apply SSA to the b elements of $\mathcal{B}_j$, i.e., as if $\mathcal{B}_j$ where the only data at hand, to produce the jth pseudo-SSA-statistic $\bar{\theta}^{(j)}_{b',b,SS}$. (b) Repeat part (a) for $j = 1, \ldots, q$ to yield the subsampling distribution $\tilde{L}_{b',b,SS}(x) = q^{-1} \sum_{j=1}^{q} 1\{\kappa_b \left(\bar{\theta}^{(j)}_{b',b,SS} - \bar{\theta}_{b,n,SS} \right) \leq x\}$. Then, under the premises of Corollary 4.2, we have

$$\sup_x |\tilde{L}_{b',b,SS}(x) - P\{\kappa_n \left(\bar{\theta}_{b,n,SS} - \theta \right) \leq x\}| \xrightarrow{P} 0 \tag{12}$$

allowing for the construction of confidence intervals for θ based on the quantiles of $\tilde{L}_{b',b,SS}(x)$. The latter is closely related to the notion of convolved subsampling; see Tewes et al. (2019).

Finally, note that an analog of Corollary 4.2 can also be formulated when the strong mixing coefficients only decay polynomially fast. In that case, however, the ϵ appearing in the moment assumption of Corollary 4.1 can not be any positive number, i.e., it can not be taken arbitrarily close to zero. Rather, the minimum value of ϵ allowed would be dictated by the polynomial rate of decay of strong mixing; see the assumptions of Theorem B.O.1 of Politis et al. (1999).

Acknowledgement. Many thanks are due to Ery Arias-Castro, Patrice Bertail, Jelena Bradic, Tucker McElroy and Yiren Wang for helpful suggestions, and to the Editor, Associate Editor and four referees for their constructive comments. Research supported by NSF grant DMS 19-14556.

References

[1] Banerjee, M., Durot, C., and Sen, B. (2019). Divide and conquer in nonstandard problems and the super-efficiency phenomenon. *Ann. Statist.*, 47, 720–757.

[2] Bertail, P., Chautru, E. and Clémencon, S. (2017). Empirical Processes in Survey Sampling with (Conditional) Poisson Designs, *Scandinavian Journal of Statistics*, 44: 97-111.

[3] Bradic, J. (2016). Randomized maximum-contrast selection: Subagging for large-scale regression. *Elect. J. Statist.*, 10: 121-170.

[4] Bühlmann, P. and Yu, B. (2002). Analyzing bagging. *Ann. Statist.*, 30(4): 927-961.

[5] Chen, S.X. and Peng, L. (2021). Distributed statistical inference for massive data. *Ann. Statist.* 49(5): 2851-2869.

[6] Chen, X. and Xie, M. (2014). A split-and-conquer approach for analysis of extraordinarily large data. *Statist. Sinica*, 24 1655–1684.

[7] Efron, B., Hastie, T., Johnstone, J. and Tibshirani, R. (2004). Least angle regression. *Ann. Statist.*, 32(2): 407-499.

[8] Jordan, M.J. (2013). On statistics, computation and scalability, *Bernoulli*, 19(4): 1378-1390.

[9] Kleiner, A., Talwalkar, A., Sarkar, P. and Jordan, M.J. (2014). A scalable bootstrap for massive data. *J. Roy. Soc., Ser. B*, 76(4): 795-816.

[10] Lin, N. and Xi, R. (2010). Fast surrogates of U-statistics. *Comp. Statist. Data Anal.*, 54:16-24.

[11] Politis, D.N. (2021). Scalable subsampling: computation, aggregation and inference, Preprint arXiv:2112.06434.

[12] Politis, D.N., and Romano, J.P. (1994), Large sample confidence regions based on subsamples under minimal assumptions, *Ann. Statist.*, 22: 2031-2050.

[13] Politis, D.N., Romano, J.P. and Wolf, M. (1999), *Subsampling,* Springer, New York.

[14] Sengupta, S., Volgushev, S. and Shao, X. (2016). A subsampled double bootstrap for massive data, *Journal of the American Statistical Association*, 111(515): 1222-1232.

[15] Tewes, J., Politis, D.N. and Nordman, D.J. (2019). Convolved subsampling estimation with applications to block bootstrap, *Annals of Statistics*, 47(1): 468-496.

[16] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B.*, 58: 267–288.

[17] Ting, D. (2021). Simple, optimal algorithms for random sampling without replacement preprint arXiv:2104.05091.

[18] Yao, Y. and Wang, H. (2021). A review on optimal subsampling methods for massive datasets. *Journal of Data Science*, 19(1): 1–22.

[19] Zhang, Y., Duchi, J. C. and Wainwright, M. J. (2013). Communication-efficient algorithms for statistical optimization. *J. Mach. Learn. Res.* 14 3321–3363.

[20] Zou, T., Li, X., Liang, X. and Wang, H. (2021). On the subbagging estimation for massive data, preprint arXiv:2103.00631.