Proceedings of the 2023 Winter Simulation Conference
C. G. Corlu, S. R. Hunter, H. Lam, B. S. Onggo, J. Shortle, and B. Biller, eds.

BOOTSTRAPPING AND BATCHING FOR OUTPUT ANALYSIS

Raghu Pasupathy

Department of Statistics
Purdue University
West Lafayette, IN 47906, USA

ABSTRACT

We review *bootstrapping* and *batching* as devices for statistical inference in simulation output analysis. Bootstrapping, discovered in the late 1970s and developed over the ensuing three decades, is widely held as being among the important scientific discoveries of the previous century due primarily to its facility for general statistical inference. By contrast, batching was introduced in the 1960s but was developed within the simulation community (in the 1980s) for the narrower contexts of variance parameter estimation and confidence interval construction. In recent years, however, there has been increasing realization that batching, much like bootstrapping, can be used also for general statistical inference, and that batching often compares favorably with bootstrapping in dependent data contexts. Bootstrapping and batching have tremendous applicability for uncertainty quantification in simulation, and are prime candidates for adoption in simulation software. We describe the general principles underlying bootstrapping and batching, outline guarantees, and discuss implementation.

1 INTRODUCTION

Suppose we have a "dataset" $(X_1, X_2, \ldots, X_n)$ of identically distributed $\mathscr{X}$-valued random variables obtained somehow, e.g., using a simulation, in the service of estimating an unknown quantity θ. We stipulate only that $(\mathscr{X}, \mathscr{A})$ is some measurable space, and that the $\mathscr{X}$-valued random variables $(X_1, X_2, \ldots, X_n)$ form the initial segment of a time-series in steady-state. For our purposes, the desired unknown "parameter" θ is general — it can reside in the d-dimensional Euclidean space $\mathbb{R}^d, d \geq 1$, or can also be function-valued although the technical parts of this tutorial do not treat the latter case. It is important that the X_js may not be independent, and can exhibit heavy serial correlation as is often the case in simulation settings. Since $(X_1, X_2, \ldots, X_n)$ come from a steady-state distribution, we can assume each of $X_1, X_2, \ldots, X_n$ is distributed according to P (unknown), and that θ_n, constructed using the dataset $(X_1, X_2, \ldots, X_n)$, estimates the unknown parameter θ. The error in the estimator θ_n is thus $\varepsilon_n := \theta_n - \theta$.

A substantial portion of simulation output analysis, and all of statistical inference, is about understanding the nature of F_{ε_n}, the sampling distribution of ε_n. For instance, when $\theta_n, \theta \in \mathbb{R}$, statistical inference on ε_n means estimating such objects as the standard error

$$\mathrm{se}(\varepsilon_n) = \sqrt{\mathrm{Var}(\varepsilon_n)} := \sqrt{\int_{-\infty}^{\infty} (x - \mu_{\varepsilon_n})^2 \, \mathrm{d}F_{\varepsilon_n}(x)}; \quad \mu_{\varepsilon_n} := \int_{-\infty}^{\infty} x \, \mathrm{d}F_{\varepsilon_n}(x),$$

the bias

$$\mathrm{bias}(\theta_n, \theta) = \mathbb{E}[\theta_n] - \theta,$$

the γ-quantile

$$Q_\gamma(\varepsilon_n) := \min\{x : F_{\varepsilon_n}(x) \leq x) \geq \gamma\}, \quad \gamma \in [0, 1],$$

979-8-3503-6966-3/23/$31.00 ©2023 IEEE

or the $(1-\alpha)$-confidence interval on θ, that is, an interval I_n constructed from data $(X_1, X_2, \ldots, X_n)$ such that $\lim_{n\to\infty} P(\theta \in I_n) = 1-\alpha$. All such effort to understand F_{ε_n} is in the important service of providing a simulation practitioner with some measure of uncertainty on the estimator θ_n.

1.1 Bootstrapping and Batching in a Nutshell

In this tutorial, we detail two "omnibus" methods for estimating aspects of the distribution F_{ε_n}. The first of these, called *bootstrapping* (Efron and Tibshirani 1994; Efron 1979; van der Vaart and Wellner 1996; Hall 1992), broadly works as follows. Since F_{ε_n} is unknown, bootstrapping identifies another random variable ε_n^* whose distribution $F_{\varepsilon_n^*}$ approximates F_{ε_n} as $n \to \infty$ (in a sense that will be made precise), and from which observations can be generated easily. The facility to generate from $F_{\varepsilon_n^*}$, called resampling, is important because observations generated from $F_{\varepsilon_n^*}$ can then be used to estimate virtually any aspect of $F_{\varepsilon_n^*}$. Thus, for instance, when seeking a $(1-\alpha)$ confidence interval on $\theta \in \mathbb{R}$, the logic of bootstrapping suggests approximating the exact (but unknown) $(1-\alpha)$ confidence interval $\left(\theta_n - F_{\varepsilon_n}^{-1}(\frac{\alpha}{2}),\ \theta_n + F_{\varepsilon_n}^{-1}(1-\frac{\alpha}{2})\right)$ with the interval

$$\left(\theta_n - F_{B,\varepsilon_n^*}^{-1}(\frac{\alpha}{2}),\ \theta_n + F_{B,\varepsilon_n^*}^{-1}(1-\frac{\alpha}{2})\right),$$

where

$$F_{B,\varepsilon_n^*}(x) = \frac{1}{B}\sum_{j=1}^{B} \mathbb{I}(Y_j \le x), \quad Y_j \overset{\text{iid}}{\sim} F_{\varepsilon_n^*}$$

and the notation $\mathbb{I}(A) = 1$ if A is true and 0 otherwise.

The second method we discuss in this tutorial, called *batching* (Su et al. 2023; Pasupathy et al. 2023; Calvin and Nakayama 2013; Alexopoulos et al. 2007), works as follows. Assume there exists a "variance constant" σ so that $\sqrt{n}\,\varepsilon_n/\sigma$ stabilizes for large n (in a sense to be made clear). Then batching constructs an estimator $\hat{\sigma}_n$ of σ, and identifies a limiting random variable T_{OB} so that

$$\sqrt{n}\frac{\varepsilon_n}{\hat{\sigma}_n} \overset{\text{d}}{\to} T_{\text{OB}} \text{ as } n \to \infty, \tag{1}$$

where "$\overset{\text{d}}{\to}$" refers to convergence in distribution. A crucial point in batching is that the limit T_{OB} is "distribution free" in that it does not depend on any unknown parameters. Thus, when seeking a $(1-\alpha)$ confidence interval on $\theta \in \mathbb{R}$, batching suggests approximating the $(1-\alpha)$ confidence interval $\left(\theta_n - F_{\varepsilon_n}^{-1}(\frac{\alpha}{2}),\ \theta_n + F_{\varepsilon_n}^{-1}(1-\frac{\alpha}{2})\right)$ with

$$\left(\theta_n - t_{\alpha/2,\text{OB}}\frac{\hat{\sigma}_n}{\sqrt{n}},\ \theta_n + t_{1-\alpha/2,\text{OB}}\frac{\hat{\sigma}_n}{\sqrt{n}}\right), \tag{2}$$

where $t_{\gamma,\text{OB}}$ denotes the γ-quantile value of T_{OB} — $t_{\gamma,\text{OB}}$ is known because T_{OB} is free from unknown parameters, and "percentile tables" for T_{OB} can be (and have been) computed.

When one seeks an object other than a $(1-\alpha)$ confidence interval on θ, e.g., $\text{se}(\varepsilon_n)$, $\text{bias}(\theta_n, \theta)$, or $Q_\gamma(\varepsilon_n)$, similar ideas apply, as we briefly outline in Section 3 and Section 4.

1.2 Paper Organization

The remaining portion of the tutorial is organized as follows. The ensuing section, in an attempt to further the reader's intuition, details three settings where a simulationist might naturally want to perform statistical inference. This is followed by Section 3 and then by Section 4 which describe bootstrapping and batching, respectively.

2 EXAMPLE SETTINGS

To provide the reader a sense of the diversity of contexts that come under the purview of the methods described in this paper, we present three example settings. In each case, we clarify the unknown parameter θ, the estimator θ_n, and the dataset $(X_1, X_2, \ldots, X_n)$.

Example I (Wait Time in a $G/G/1$ Queue)

Consider the $G/G/1$ queue where a single server that serves customers arriving according to an arrival process with independent and identically distributed (iid) inter-arrival times having distribution G_1. Customers are served in the order in which they arrive after joining a queue having infinite capacity. Service times for customers are iid according to a distribution G_2. Suppose G_1, G_2 and the initial conditions are such that the system is at steady-state, that is, $W_n \stackrel{\mathrm{d}}{=} W \; \forall n \geq 1$, where W is a well-defined random variable having distribution F_W. Let $\theta = \min\{w : F_W(w) \geq 0.90\}$ denote the 0.9-quantile of W, and suppose that $(X_1, X_2, \ldots, X_n)$ are the observed waiting times of the first n customers in the system, so that

$$\theta_n := \min\{w : F_{n,W}(w) \geq 0.90\}; \quad F_{n,W}(w) := \frac{1}{n}\sum_{j=1}^{n} \mathbb{I}(W \leq w).$$

And, as described in the introduction, a simulationist interested in statistical inference on θ_n is essentially attempting to understand the sampling distribution of $\varepsilon_n = \theta_n - \theta$.

Example II (Time-Dependent Inventory Levels in a Supply Chain)

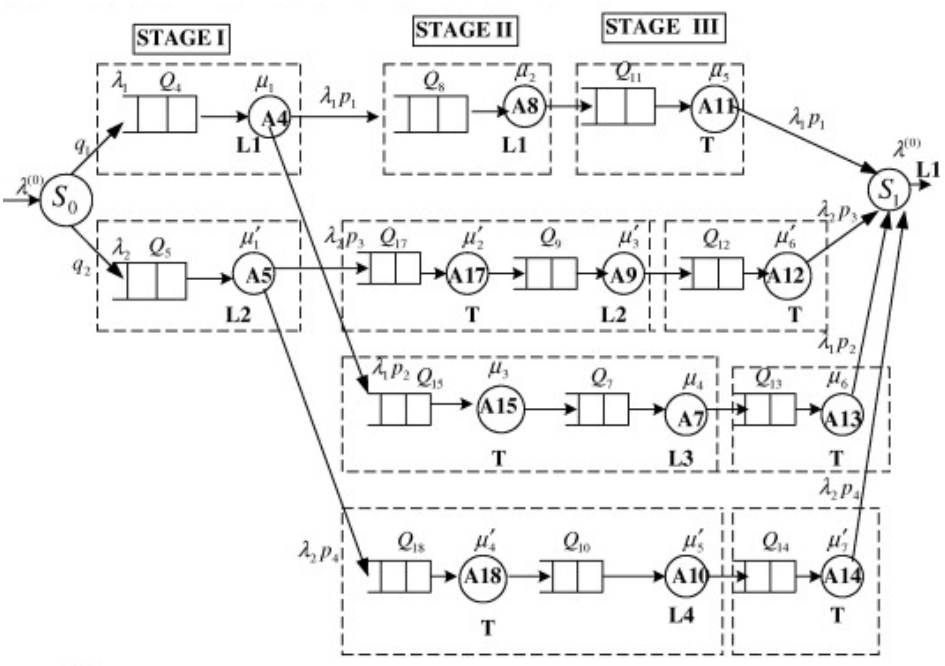

Figure 1: A queueing network model of a supply chain.

As a more elaborate example, consider the global supply chain introduced in the tutorial by Ingalls (2014), where the simulationist wishes to analyze the delivery of computing servers produced in Europe to the Asia-Pacific region, with the specific intention of evaluating whether it may be wise to move production to Singapore. Due to the complexity and scale of such a supply chain, it is easy to see why a simulation model would be helpful in answering many narrow questions, e.g., effect on inventory, effect on on-time

delivery, effect on costs and revenue, which together will be pertinent to the broader question of whether a move to Singapore is warranted.

Consider one such narrow question, that of *time-dependent inventory level*, that is, inventory as a function of time, at a specified location and observed over a horizon $[0,T]$ of interest. The simulationist executes n runs of the simulation, producing time-dependent inventory level $X_j(t), t \in [0,T]$ during the j-th run. Importantly, notice that the j-th "observation" denoted $X_j := X_j(t), t \in [0,T]$ is a function of time, or a *random function*. Suppose the simulationist is especially interested in analyzing low inventory levels and so chooses the parameter θ to be the 20-th percentile inventory level as a function of time, that is, $\theta := \theta(t), t \in [0,T]$, where $\theta(t)$ is the 20-th percentile inventory at time t. Recognize again that the parameter θ is a fixed, unknown function in time. Recalling the "dataset" $(X_1, X_2, \ldots, X_n)$ generated by n runs of the simulation, an estimator $\theta_n := \theta_n(t), t \in [0,T]$ of θ can then be constructed as:

$$\theta_n(t) := \min \left\{ y : \frac{1}{n} \sum_{j=1}^{n} \mathbb{I}(X_j(t) \leq y) \geq 0.2) \right\}, \quad t \in [0,T]. \tag{3}$$

The simulationist may have chosen a different parameter of interest, e.g., the mean vector of inventory levels at d specific locations $\ell_1, \ell_2, \ldots, \ell_d$ in the supply chain, at the fixed time instant T. In this case, denoting $\pi_{j,T}, j = 1,2,\ldots,d$ as the inventory level distribution at time T in location j, and denoting $X_{i,j}(T), j = 1,2,\ldots,d; i = 1,2,\ldots,n$ as the i-th observed inventory level in location j at time T, we can write:

$$\theta := \left(\int_\ell \ell \, \pi_{1,T}(\mathrm{d}\ell), \int_\ell \ell \, \pi_{2,T}(\mathrm{d}\ell), \ldots, \int_\ell \ell \, \pi_{d,T}(\mathrm{d}\ell) \right).$$

In such a case, the estimator $\theta_n := \left(\frac{1}{n}\sum_{i=1}^n X_{i,1}(T), \frac{1}{n}\sum_{i=1}^n X_{i,2}(T), \ldots, \frac{1}{n}\sum_{i=1}^n X_{i,d}(T) \right)$.

Example III (Nonlinear System of Equations)

Variable toll pricing has become a popular method to manage traffic on highways, by shifting purely discretionary traffic to off-peak hours or other roadways. Accordingly, a question of immense interest involves identifying the relationship between the toll price and the resulting congestion levels at steady state, toward better congestion pricing policies.

Let's introduce notation to make this question more precise. Suppose $p = (p_1, p_2, \ldots, p_d), p_i \in [0,M]$ represents the prevailing toll price for d vehicle classes, and $\theta = \{(\theta_1(p), \theta_2(p), \ldots, \theta_d(p)), p \in [0,M]^d\}$ the corresponding expected steady state waiting time at the tolls for each of the d classes. Given the complicated relationship between the expected wait time and the toll price, a simulation (whose mechanics are not relevant for our purposes) is used to estimate the parameter θ. Suppose the simulation yields the output $(X_1, X_2, \ldots, X_n)$, where $X_i = (X_{i1}(p), X_{i2}(p), \ldots, X_{id}(p)), p \in [0,M]^d$ represents the i-th realization of the wait time vector, that is, the vector wait times corresponding to the i-th vechicle in each of the d classes, with p held fixed. It is important to observe that each output observation X_i in this example is a *random function or surface* of the toll price. A useful thought experiment that clarifies the nature of X_i is as follows. Fix and hold all "random elements" of the simulation while varying the toll price p to form a time series of observations, each of which is a function of the price p.

Suppose the simulationist is interested in setting the tolls $p = (p_1, p_2, \ldots, p_d)$ so that the expected wait times for the d classes matches target wait times $\gamma_1, \gamma_2, \ldots, \gamma_d$, respectively. Then the parameter θ is the solution (in p) to the following nonlinear system of equations:

$$\int_x x_j \pi_p(\mathrm{d}x) = \gamma_j, \quad j = 1,2,\ldots,d. \tag{4}$$

Of course the solution θ to (4) is unknown, but can be estimated as θ_n by solving the corresponding system constructed using the data generated by simulation, that is, by solving the system:

$$\frac{1}{n}\sum_{i=1}^{n} X_{ij}(p) = \gamma_j, \quad j = 1,2,\ldots,d. \tag{5}$$

(There are existence and uniqueness issues pertaining to the solution of (5) but we omit discussion about such details here.) And, as in Example I and Example II, the inference question here is whether anything can be inferred about the nature of the error $\theta_n - \theta$.

The type of inference considered within each of the examples described is *conditional* on a given simulation model and pertains to quantifying the output uncertainty $\theta_n - \theta$. This is in important contrast to another modern popular topic called simulation *input uncertainty* (Henderson 2003; Chick 2001; Cheng and Holland 1997; Barton 2012; Lam 2016), which quantifies the effect of errors in the input distributions that form the primitives to the simulation. In effect, the type of inference treated in this paper provides a sense of how decision-making might be affected due to performing too few simulation runs, whereas input uncertainty deals with the corresponding effects due to a lack of adequate real-world data used when estimating the distributional input to the simulation.

Both output uncertainty and input uncertainty in simulation are subsumed by the recently phrased "umbrella" topic *uncertainty quantification* (Abdar et al. 2021; Najm 2009; Soize 2017) which should be understood loosely as the effort to quantify the effect of all sources of error, e.g., input parameters, structure, logic, and solution, within models that include, but not limited to, simulation. Some examples of models other than simulation are stochastic differential equations (Hoel et al. 1986), neural networks (Bottou et al. 2018), and regression (Wasserman 2004).

3 BOOTSTRAPPING

Recall the "observed dataset" $(X_1, X_2, \ldots, X_n)$ in $(\mathscr{X}, \mathscr{A})$ and the empirical measure

$$P_n(A) := n^{-1}\sum_{j=1}^{n} \delta_{X_j}(A), \quad A \in \mathscr{A}$$

constructed from the observed dataset. Also recall the notation $\theta \equiv \theta(P)$ and $\theta_n \equiv \theta(P_n)$ for the unknown parameter of interest and its estimator, respectively, and $\varepsilon_n := \theta(P_n) - \theta(P)$. (Writing θ and θ_n as $\theta(P)$ and $\theta(P_n)$ allows treating these objects as functions of the probability measures P and P_n, respectively; so, $\theta(\cdot)$ can be viewed as a statistical functional.) We wish to (a) estimate $\psi(F_{\varepsilon_n})$, where $\psi : \mathscr{W} \to \mathbb{R}$ is a statistical functional that subsumes such objects as $\mathrm{se}(\varepsilon_n)$, $\mathrm{bias}(\theta_n, \theta)$, or $Q_\gamma(\varepsilon_n)$; or (b) construct an asymptotically valid $(1-\alpha)$ confidence interval on θ. For simplicity of exposition, let's suppose that $\theta, \theta_n \in \mathbb{R}$.

In the simplest and most pervasive flavor of the bootstrap, a "bootstrap dataset" $(X_1^*, X_2^*, \ldots, X_n^*)$ is defined through uniform iid sampling with replacement from the observed dataset $(X_1, X_2, \ldots, X_n)$, that is, each $X_j^* \overset{\text{iid}}{\sim} P_n, \quad j = 1,2,\ldots,n$. Let $P_n^*(A) := n^{-1}\sum_{j=1}^{n} \delta_{X_j^*}(A), A \in \mathscr{A}$ denote the empirical measure constructed from the bootstrap dataset $(X_1^*, X_2^*, \ldots, X_n^*)$. Then, the following two loosely stated observations underlie the bootstrapping principle and naturally lead to a method to perform inference on $\theta_n - \theta$.

(a) Under arguably weak conditions, the "conditional distribution" of $\sqrt{n}(P_n^* - P_n)$ converges almost surely, as $n \to \infty$, to the weak limit of $\sqrt{n}(P_n - P)$. (When we refer to the distribution of $\sqrt{n}(P_n^* - P_n)$, we are referring to its distribution *conditional* on the "dataset" $(X_1, X_2, \ldots, X_n)$, and so converging "almost surely" means given almost all sequences $\{X_n, n \geq 1\}$.)

(b) Under (a), and if the functional $\theta(\cdot)$ is well-behaved at P in the sense of being Hadamard differentiable (van der Vaart and Wellner 1996, page 373), the distribution of $\sqrt{n}(\theta_n(P) - \theta(P))$ stabilizes to a P-Brownian bridge process (defined in Section 3.1) and is consistently approximated by the distribution of $\sqrt{n}(\theta(P_n^*) - \theta(P))$.

The observations in (a) and (b) lead to a basic bootstrapping algorithm to estimate $\psi(F_{\varepsilon_n})$ and to construct an asymptotically valid $(1-\alpha)$ confidence interval on θ. To see how, notice from (a) and (b) that the distribution of $\varepsilon_n^* := \theta(P_n^*) - \theta(P_n)$ is a "good approximation" to the distribution of $\varepsilon_n := \theta(P_n) - \theta(P)$. Furthermore, it is in principle easy to generate iid observations from $F_{\varepsilon_n^*}$ through the generation of multiple bootstrap datasets, thus allowing to readily estimate $\psi(F_{\varepsilon_n^*})$.

Algorithm 1: The Basic Bootstrap

Compute $\theta_n := \theta(P_n)$, where $P_n = \frac{1}{n}\sum_{j=1}^n \delta_{X_j}$.
for $j = 1, 2, ...B$ **do**
  $(X_{j,1}^*, X_{j,2}^*, \ldots, X_{j,n}^*) \overset{\text{iid}}{\sim} P_n$;
  Compute $\theta_{j,n}^* := \theta(P_{j,n}^*)$ for $j = 1, 2, \ldots, B$ where $P_{j,n}^* := \frac{1}{n}\sum_{j=1}^n \delta_{X_{j,n}^*}$;
  Compute $\varepsilon_{j,n}^* := \theta_{j,n}^* - \theta_n$;
end
Compute $\hat{\psi}(F_{\varepsilon_n^*}) := \psi(P_{B,\varepsilon_n^*}^*)$, where $P_{B,\varepsilon_n^*}^* := \frac{1}{B}\sum_{j=1}^B \delta_{\varepsilon_{j,n}^*}$.

So, as an example, if $\theta, \theta_n \in \mathbb{R}$ and the standard-error of the error distribution is the target of inference, Algorithm 1 suggests estimating $\text{se}(\varepsilon_n)$ as

$$\hat{\text{se}}\,(\varepsilon_n^*) := \sqrt{\frac{1}{B}\sum_{j=1}^B (\theta_{j,n}^* - \theta_n)^2}.$$

Similarly, if the γ-quantile $Q_\gamma(\varepsilon_n)$ of the error is the target of inference, then Algorithm 1 implies estimating $Q_\gamma(\varepsilon_n)$ as

$$\hat{Q}_\gamma(\varepsilon_n^*) := \min\left\{x : \sum_{j=1}^B \mathbb{I}(\theta_{j,n}^* - \theta_n^* \le x) \ge \gamma\right\}.$$

And, as Section 1 notes, a $(1-\alpha)$ confidence interval on θ as suggested by Algorithm 1 becomes

$$\left(\theta_n - F_{B,\varepsilon_n^*}^{-1}(\frac{\alpha}{2}),\ \theta_n + F_{B,\varepsilon_n^*}^{-1}(1-\frac{\alpha}{2})\right), \tag{6}$$

where $F_{B,\varepsilon_n^*}(x) = \frac{1}{B}\sum_{j=1}^B \mathbb{I}(\varepsilon_{j,n}^* \le x)$.

3.1 Bootstrap Guarantee

The bootstrap procedure in Algorithm 1 is a natural outgrowth of the two ideas outlined in (a) and (b). In what follows, we make the ideas in (a) and (b) rigorous through two theorems stated without proof. We ignore all measurability issues when stating these theorems; see (van der Vaart and Wellner 1996) for a complete treatment.

Suppose $\mathscr{F}$ is a collection of measurable functions ("random variables") from $\mathscr{X} \to \mathbb{R}$. This automatically defines the $\mathscr{F}$-indexed *empirical process* $\mathbb{G}_n$ given by

$$f \mapsto \mathbb{G}_n f = \sqrt{n}(P_n - P)f,$$

where we have used the notation $Qf = \int f\,\mathrm{d}Q$ and $(P_n - P)f = \int f\,\mathrm{d}P_n - \int f\mathrm{d}P$. For a given $f \in \mathscr{F}$, if Pf exists we have the classical "law of large numbers" $P_n f \overset{\text{wp1}}{\to} Pf$ as $n \to \infty$; and if $Pf^2 < \infty$, we have the classical central limit theorem $\mathbb{G}_n \overset{\text{d}}{\to} N(0, P(f-Pf)^2)$ as $n \to \infty$. (As a matter of terminology, each $\mathbb{G}_n$ is a "process" indexed by $f \in \mathscr{F}$, that is, each $\mathbb{G}_n$ is a collection of random variables labeled by $f \in \mathscr{F}$.

Less confusion ensues if we do not use the word *process* but simply refer to $\mathbb{G}_n$ as a "random variable" or "random object" and, correspondingly, to $\{\mathbb{G}_n, n \geq 1\}$ as a sequence of random variables.)

Define the *envelope* F associated with $\mathscr{F}$ as $F(x) := \sup_{f\in\mathscr{F}} |f(x) - Pf|, \quad x \in \mathscr{X}$, and notice that if $F(x) < \infty$ for each x, then $\mathbb{G}_n f \in \ell^\infty(\mathscr{F})$ and so the process $\{\mathbb{G}_n f, f \in \mathscr{F}\}$ can be viewed as a map into $\ell^\infty(\mathscr{F})$, where $\ell^\infty(\mathscr{F})$ is the space of uniformly bounded real-valued functions on $\mathscr{F}$, that is, the set of $g : \mathscr{F} \to \mathbb{R}$ such that $\sup_{f\in\mathscr{F}} |g(f)| < \infty$.

The class $\mathscr{F}$ is said to be *P-Donsker* if the sequence $\{\mathbb{G}_n, n \geq 1\}$ converges weakly to a tight Borel-measurable element in $\ell^\infty(\mathscr{F})$: $\mathbb{G}_n \xrightarrow{\mathrm{d}} \mathbb{G}$ in $\ell^\infty(\mathscr{F})$. In such a case, the limit process $\mathbb{G}$ is a mean-zero Gaussian process called the *P-Brownian bridge* having covariance

$$\begin{aligned}\mathbb{E}[\mathbb{G}f_1 f_2] := \int \mathbb{G}f_1 \mathbb{G}f_2 &= P(f_1 - Pf_1)(f_2 - Pf_2)\\ &= Pf_1 f_2 - Pf_1 Pf_2.\end{aligned}$$

Corresponding to the empirical process $\mathbb{G}_n$, let's also define the *bootstrap empirical process*

$$\mathbb{G}_n^* := \sqrt{n}(P_n^* - P_n) = \frac{1}{n}\sum_{j=1}^{n}(M_{n,i} - 1)\delta_{X_i},$$

where $M_{n,i}$ is the number of times the observation X_i was chosen from the "original dataset" $(X_1, X_2, \ldots, X_n)$ during the iid resampling process. Accordingly, the vector $M := (M_{n,1}, M_{n,2}, \ldots, M_{n,n})$ is a multinomial random vector having parameters n and $(1/n, 1/n, \ldots, 1/n)$, independent of $(X_1, X_2, \ldots, X_n)$. We are now ready to state a theorem that rigorizes the statement in (a) made earlier; BL_1 in Theorem 1 refers to the bounded Lipschitz metric (van der Vaart and Wellner 1996, page 73).

Theorem 1 (Theorem 3.6.2, (van der Vaart and Wellner 1996)) Let $\mathscr{F}$ be a class of $\mathscr{X} \to \mathbb{R}$ measurable functions with a finite envelope. Then, the following statements are equivalent.

(i) $\mathscr{F}$ is P-Donsker and $\int \sup_{f\in\mathscr{F}}(f - Pf)^2 < \infty$.

(ii) $\sup_{h\in\mathrm{BL}_1} |\mathbb{E}_M h(\mathbb{G}_n^*) - \mathbb{E}[h(\mathbb{G})]| \xrightarrow{\mathrm{wp1}} 0$.

It is important to recognize that the assertion in (ii) of Theorem 1 is conditional on the dataset $(X_1, X_2, \ldots, X_n)$, that is, given almost all sequences $X_1, X_2, \ldots$. In what sense does Theorem 1 rigorize the statement in (a)? Notice from the arguments preceding Theorem 1 that if $\mathscr{F}$ is P-Donsker, then $\mathbb{G}_n$ converges weakly to the P-Brownian bridge process. With the added condition that $\int \sup_{f\in\mathscr{F}}(f - Pf)^2 < \infty$, Theorem 1 guarantees that the empirical bootstrap process $\mathbb{G}_n^*$ also converges weakly to the P-Brownian bridge process, rigorously establishing the idea loosely stated in (a).

Under Hadamard differentiability of the statistical functional $\theta(\cdot)$ along with the postulates of Theorem 1, the following theorem provides a rigorous statement of the principle stated in (b).

Theorem 2 (Theorem 3.9.11, (van der Vaart and Wellner 1996)) Let $\theta : \mathscr{W} \to \mathbb{R}$ be a statistical functional that is Hadamard differentiable on the normed space $\mathscr{W}$, and suppose the sequence $P_n \in \mathscr{W}$ for $n \geq 1$. Let $\mathscr{F}$ be a class of $\mathscr{X} \to \mathbb{R}$ measurable functions such that $\mathscr{F}$ is P-Donsker and $\int \sup_{f\in\mathscr{F}}(f - Pf)^2 < \infty$. Then

$$\sup_{h\in\mathrm{BL}_1} \left|\mathbb{E}_M\left[h\left(\sqrt{n}\left(\theta\left(P_n^*\right) - \theta(P_n)\right)\right)\right] - \mathbb{E}\left[h\left(\theta'(\mathbb{G})\right)\right]\right| \to 0,$$

where $\mathbb{G}$ is the P-Brownian bridge process.

The requirement for Hadamard differentiability is in general weak but can be further weakened. Also, Thoerem 2 demonstrates consistency (that is, convergence in probability) of the conditional law of $\sqrt{n}(\theta(P_n^*) - \theta(P_n))$ to the law of $\sqrt{n}(\theta(P_n) - \theta(P))$. Such convergence can be strengthened to almost sure convergence if a certain form of uniform Hadamard differentiability is assumed — see Theorem 3.9.13 in van der Vaart and Wellner (1996).

Recall that the broad idea in bootstrap is to approximate the unknown sampling distribution F_{ε_n} of ε_n with the "known" distribution $F_{\varepsilon_n^*}$. Theorem 1 in effect assures us that under certain regularity conditions, as $n \to \infty$,

$$\sup_{t \in \mathbb{R}^d} \left| P\left(\sqrt{n}\,\varepsilon_n^* \le t\right) - P\left(\sqrt{n}\,\varepsilon_n \le t\right) \right| \overset{\text{wp1}}{\to} 0, \tag{7}$$

where $\varepsilon_n^* = \theta(P_n^*) - \theta(P_n)$ and $\varepsilon_n = \theta(P_n) - \theta(P)$. In other words, $F_{\varepsilon_n^*}$ converges to F_{ε_n} almost surely on the $\sqrt{n}$ scale. One of the most celebrated aspects of the bootstrap is what is called *higher order accuracy*. Specifically, under certain conditions usually imposed on the higher order moments associated with $\theta(P_n)$, a guarantee such as what follows obtains.

$$\sqrt{n} \sup_{t \in \mathbb{R}^d} \left| P\left(\sqrt{n}\,\varepsilon_n^* \le t\right) - P\left(\sqrt{n}\,\varepsilon_n \le t\right) \right| \overset{\text{p}}{\to} 0, \tag{8}$$

Loosely, the guarantee in (8) states that the supremum norm deviation between $F_{\varepsilon_n^*}$ and F_{ε_n} (on the $\sqrt{n}$ scale) converges to zero in probability faster than $O(1/\sqrt{n})$. We do not go into further detail on the specific nature of a guarantee such as (8) but instead direct the reader to (Shao and Tu 2012).

3.2 Named Bootstrap Contexts

The bootstrap principle as reflected through (a), (b) and Algorithm 1 is remarkably general and has been applied for statistical inference in a wide variety of classical contexts such as estimating standard errors in curve fitting (Efron and Tibshirani 1994, page 70), regression (Efron and Tibshirani 1994, Chapter 9), and bias estimation (Efron and Tibshirani 1994, Chapter 10). As a reflection of such widespread use, particular contexts in which the bootstrap has been used has given rise to names that have become popular in the literature. For example, the *percentile bootstrap interval* (Efron and Tibshirani 1994, page 170) is used to describe contexts where a required $(1-\alpha)$ confidence interval on θ is constructed as in (6), using the quantiles associated with the empirical cdf F_{B,ε_n^*}. By contrast, the *bootstrap-t* interval constructs the $(1-\alpha)$ confidence interval for the same context as

$$\left(\theta_n - F_{B,t}^{-1}(1-\alpha)\hat{\text{se}}_{B,n}, \theta_n + F_{B,t}^{-1}(\alpha)\hat{\text{se}}_{B,n}\right),$$

where $F_{B,t}$ is the empirical cdf constructed from $Z_j = (\theta_{j,n}^* - \theta_n)/\hat{\text{se}}_{B,n}$, $j = 1,2,\ldots,B$. Similarly, the *residual bootstrap* (Efron and Tibshirani 1994, page 113) refers to bootstrapping residuals from a complicated model in service of estimating standard errors within the model; and *parametric bootstrap* (Cheng 2017; Efron and Tibshirani 1994) where an assumed parametric model for the underlying population drives resampling as opposed to the empirical measure used in the basic bootstrap algorithm we have described.

3.2.1 Moving Blocks Bootstrap

The basic bootstrap in Algorithm 1 and weighted variations such as the exchangeable bootstrap (van der Vaart and Wellner 1996, Section 3.6.2) assume that the data $(X_1, X_2, \ldots, X_n)$ in the original dataset are iid. This, of course, need not be the case. In fact, in simulation contexts such as those discussed in Section 2, it is routinely the case that the data $(X_1, X_2, \ldots, X_n)$ form the initial segment of a steady state time series that exhibits heavy autocorrelation. In such contexts, applying the bootstrap principle by iid resampling will destroy the underlying correlation structure that is present in the dataset and potentially lead to inconsistency.

The Moving Blocks Bootstrap (MBB) is a variation on the basic bootstrap procedure designed to preserve the dependence in the underlying time series. In a nutshell, MBB performs iid sampling of "blocks of data" as shown in Figure 2. Each block contains m_n contiguous observations from the original dataset. So, MBB and the basic bootstrap differ only in the way the B bootstrap datasets are generated. While the

j-th bootstrap dataset in the basic bootstrap contains observations $(X^*_{j,1}, X^*_{j,2}, \ldots, X^*_{j,n})$, where $X^*_{j,n} \overset{\text{iid}}{\sim} P_n$, the j-th MBB dataset contains the data

$$\left(X_{L_1}, X_{L_1+1}, \ldots, X_{L_1+m_n-1}, X_{L_2}, X_{L_2+1}, \ldots, X_{L_2+m_n-1}, \ldots, X_{L_{b_n}}, X_{L_{b_n}+1}, \ldots, X_{L_{b_n}+m_n-1}\right),$$

where $m_n b_n = n$ and the block starting locations $L_1, L_2, \ldots, L_{b_n}$ are obtained using uniform iid sampling from $\{1, 2, \ldots, n\}$.

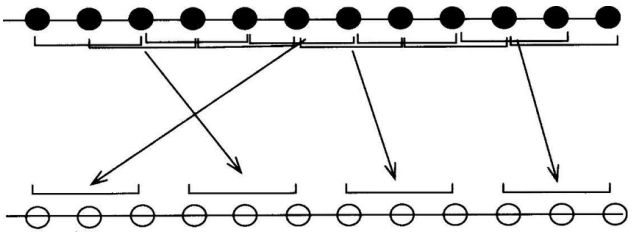

Figure 2: Illustration of Moving Blocks Bootstrap with block size $m_n = 3$, adapted from Efron and Tibshirani (1994).

It is easy to see that the block size m_n will play a crucial role in determining the performance of MBB. For instance, making m_n too small, e.g., $m_n = 1$ will make the resulting procedure resemble the basic bootstrap. At least as envisioned and stated in (Efron and Tibshirani 1994, page 102), the blocks in MBB should be large enough that observations more than m_n time units apart should be nearly independent. The choice of block size m_n, however, remains a question that is not yet fully resolved in the MBB literature.

4 BATCHING

Recall again the "observed dataset" $(X_1, X_2, \ldots, X_n)$ in $(\mathscr{X}, \mathscr{A})$, the empirical measure

$$P_n(A) := n^{-1} \sum_{j=1}^{n} \delta_{X_j}(A), \quad A \in \mathscr{A}$$

constructed from the observed dataset, the unknown parameter of interest $\theta \equiv \theta(P)$, and the estimator $\theta_n \equiv \theta(P_n)$. As before, we seek a $(1-\alpha)$ confidence interval on θ, or an estimate of $\psi(F_{\varepsilon_n})$, where $\psi : \mathscr{W} \to \mathbb{R}$ is a statistical functional subsuming such objects as $\text{se}(\varepsilon_n)$, $\text{bias}(\theta_n, \theta)$, or $Q_\gamma(\varepsilon_n)$.

Fundamental to batching, and analogous to the "moving block" in MBB (Section 3.2.1), is a *batch* of contiguous observations from the dataset $(X_1, X_2, \ldots, X_n)$. Let's introduce notation to make this idea precise. Partition $(X_1, X_2, \ldots, X_n)$ into b_n possibly overlapping batches each of size m_n as shown in Figure 3. The first of these batches consists of observations $X_1, X_2, \ldots, X_{m_n}$, the second consists of observations $X_{d_n+1}, X_{d_n+2}, \ldots, X_{d_n+m_n}$, and so on, and the last batch consists of observations $X_{(b_n-1)d_n+1}, X_{(b_n-1)d_n+2}, \ldots, X_n$. The quantity $d_n \geq 1$ represents the offset between batches, with the choice $d_n = 1$ corresponding to "fully-overlapping" batches and any choice $d_n \geq m_n$ corresponding to "non-overlapping" batches. Notice then that the offset d_n and the number of batches b_n are related as $d_n = \frac{n-m_n}{b_n-1}$. Now use the data in batches $1, 2, \ldots, b_n$ to construct the corresponding empirical measures:

$$P_{i,n}(A) := m_n^{-1} \sum_{j=1}^{m_n} \delta_{X_{(i-1)d_n+j}}(A), \quad A \in \mathscr{A};\ i = 1, 2, \ldots, b_n.$$

Analogous to the observations (a) and (b) that explain the bootstrap principle, the following two observations underlie the batching principle.

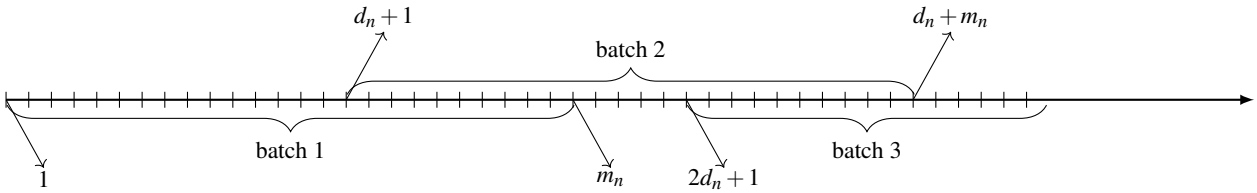

Figure 3: The figure depicts partially overlapping batches. Batch 1 consists of observations $X_j, j = 1,2,\ldots,m_n$; batch 2 consists of observations $X_j, j = d_n+1, d_n+2, \ldots, d_n+m_n$, and so on, with batch i consisting $X_j, j = (i-1)d_n+1, (i-1)d_n+2, \ldots, (i-1)d_n+m_n$. There are thus $b_n := d_n^{-1}(n-m_n)+1$ batches in total, where n is the size of the dataset.

(c) The *variance parameter* $\sigma^2 := \lim_{n\to\infty} n\mathbb{E}\left[(\theta(P_n)-\theta(P))^2\right]$, assumed to exist, can be estimated in one of various ways, e.g.,

$$S_{\text{OB-S}}(m_n,b_n) = \sqrt{\frac{m_n}{n-m_n}\times m_n \times \frac{1}{b_n}\sum_{j=1}^{b_n}(\theta(P_{j,m_n})-\theta(P_n))^2}. \tag{9}$$

(The first multiplier $m_n/(n-m_n)$ in (9) ensures asymptotic unbiasedness of $S_{\text{OB-S}}(m_n,b_n)$ with respect to σ^2, and the second multiplier m_n in (9) accounts for using batches of size m_n to estimate σ^2. The "OB" and the "-S" in the notation $S_{\text{OB-S}}(m_n,b_n)$ stand for "overlapping batch" and "sectioning," respectively.)

(d) Under certain regularity conditions, the following weak limit

$$T_{\text{OB-S}}(m_n,b_n) := \sqrt{n}\left(\frac{\theta(P_n)-\theta(P)}{S_{\text{OB-S}}(m_n,b_n)}\right) \xrightarrow{\text{d}} T_{\text{OB-S}} \text{ as } n\to\infty \tag{10}$$

exists, and the random variable $T_{\text{OB-S}}$ can be characterized. Importantly, $T_{\text{OB-S}}$ is "distribution-free," that is, it does not depend on unknown parameters.

The limit in (10) immediately suggests the symmetric $(1-\alpha)$ confidence interval

$$\left(\theta_n - F^{-1}_{T_{\text{OB-S}}}(\frac{\alpha}{2})\frac{S_{\text{OB-S}}(m_n,b_n)}{\sqrt{n}},\ \theta_n + F^{-1}_{T_{\text{OB-S}}}(1-\frac{\alpha}{2})\frac{S_{\text{OB-S}}(m_n,b_n)}{\sqrt{n}}\right), \tag{11}$$

where $F_{T_{\text{OB-S}}}$ denotes the cdf of $T_{\text{OB-S}}$. And, if estimating $\psi(F_{\varepsilon_n})$ is the object of statistical inference, then, assuming $S_{\text{OB-S}}(m_n,b_n) \xrightarrow{\text{p}} \sigma^2$ as $n\to\infty$, the limit in (10) suggests the estimator $\psi(\frac{S_{\text{OB-S}}(m_n,b_n)}{\sqrt{n}}T_{\text{OB-S}})$.

Why should the confidence interval suggested by batching "work well"? And, in particular, why should it work better than what is suggested through bootstrap? The straightforward answer to this question is that batching assumes more structure (as encoded through Assumption 1 to be stated in Section 4.1) and exploits it for efficiency. To be more precise, batching uses overlapping sets of data when estimating the variance parameter σ. Overlapping batches reduce loss of information, but also introduce dependence across batches estimates. The crucial point is that Assumption 1 allows to capture such dependence through the characterized random variable $T_{\text{OB-S}}$. This also means that batching can be expected to not perform as well when Assumption 1 is violated, which is indeed something we observe in numerical experiments.

At least two measures are important when considering the quality of a reported confidence interval: (i) *coverage probability*, that is, does the probability of a reported confidence interval such as (11) containing θ tend to the nominal probability $(1-\alpha)$ as $n\to\infty$, and if so, how rapidly? and (ii) *expected halfwidth*, that is, what is the expected half-width of the reported confidence interval? Better confidence interval

procedures exhibit rapid convergence (as $n \to \infty$) of the coverage probability to $1-\alpha$, along with low expected half-widths.

Extensive numerical experimentation (Pasupathy et al. 2023; Su et al. 2023) reveals that batch sizes m_n have a dominant effect on coverage probability, with large batch sizes ensuring more rapid convergence to the nominal probability $1-\alpha$. Numerical experimentation also reveals that the effect of batch sizes m_n on the expected half-width is more muted, with the number of batches b_n playing a more dominant role. Increasing b_n tends to rapidly decrease the expected half-width of the confidence intervals, especially for small values of b_n. These two insights from experimentation, (i) large m_n generally leads to better coverage, and (ii) large b_n generally leads to smaller expected half-widths, together suggest the use of fully-overlapping batches when constructing confidence intervals such as (11). (We recognize that "large m_n" in (i) is not a precise statement, but choosing m_n so that $m_n/n \approx 0.15$ exhibits good performance across diverse experimental settings; and, fully overlapping batches means $d_n = 1$ and $b_n = n - m_n + 1$ but this choice needs to be traded-off against the resulting increased need for computation.)

4.1 Theoretical Guarantee

In this section, we present Theorem 3 as a rigorous statement of the weak limit stated as a key principle in (d) of Section 4. Theorem 3 relies crucially on a certain type of regularity assumption called *strong approximation* (Glynn 1998; Csörgö and Révész 1981; Su et al. 2023) on the sequence $\{\theta(P_n), n \geq 1\}$.

Assumption 1 (Strong Invariance) The sequence $\{\theta(P_n), n \geq 1\}$ of estimators satisfies the following strong invariance principle. On a rich enough probability space, there exists a standard Wiener process $\{W(t), t \geq 0\}$ and a stationary stochastic process $\{\tilde{X}_n, n \geq 1\} \stackrel{d}{=} \{X_n, n \geq 1\}$ such that as $n \to \infty$,

$$\sup_{0 \leq t \leq n} \left| \sigma^{-1} \left(\theta(P_{\lfloor t \rfloor}) - \theta(P) \right) - t^{-1} W(t) \right| \leq \Gamma n^{-1/2-\delta} \sqrt{\log n} \quad a.s., \tag{12}$$

where the constant $\delta > 0$ and the real-valued random variable Γ satisfies $\mathbb{E}[\Gamma] < \infty$.

Assumption 1 is a statement about $\{\theta(P_n), n \geq 1\}$ "looking like" a Wiener process on a certain scaling. There is evidence that Assumption 1 holds in diverse contexts (Su et al. 2023), although a proof of Theorem 3 also suggests that Assumption 1 can be relaxed to a functional central limit theorem (Serfling 1980) without losing the strength of the assertions in Theorem 3.

Theorem 3 (Su, Pasupathy,Yeh, and Glynn (2023)) Suppose that Assumptions 1 holds, and that $\beta = \lim_{n\to\infty} m_n/n \in (0,1)$. Assume also that $b_n \to b \in \{2, 3, \ldots, \infty\}$ as $n \to \infty$. Define

$$\chi^2_{\text{OB-S}}(\beta, b) := \begin{cases} \dfrac{1}{\kappa_1(\beta,b)} \dfrac{1}{\beta(1-\beta)} \displaystyle\int_0^{1-\beta} \left(W(u+\beta) - W(u) - \beta W(1)\right)^2 du & b = \infty; \\ \dfrac{1}{\kappa_1(\beta,b)} \dfrac{1}{\beta b} \displaystyle\sum_{j=1}^{b} \left(W(c_j+\beta) - W(c_j) - \beta W(1)\right)^2 & b \in \mathbb{N} \setminus \{1\}, \end{cases} \tag{13}$$

where $\kappa_1(\beta, b) = 1-\beta$ and $c_j := (j-1)\frac{1-\beta}{b-1}$. Then, as $n \to \infty$,

$$S^2_{\text{OB-S}}(m_n, b_n) \stackrel{\text{d}}{\to} \sigma^2 \chi^2_{\text{OB-S}}(\beta, b); \quad \text{and} \quad T_{\text{OB-S}}(m_n, b_n) \stackrel{\text{d}}{\to} \frac{W(1)}{\sqrt{\chi^2_{\text{OB-S}}(\beta, b)}} =: T_{\text{OB-S}}. \tag{14}$$

The real-valued random variable $T_{\text{OB-S}} := \frac{W(1)}{\sqrt{\chi^2_{\text{OB-S}}(\beta,b)}}$ has been tabulated (Pasupathy et al. 2023; Su et al. 2023) thereby allowing one to construct the confidence interval in (11).

4.2 Batching Variants

Batching variants arise as a result of using estimators other than $S_{\text{OB-S}}(m_n, b_n)$ (in (9)), or by replacing the estimator $\theta(P_n)$ in (9) and (10) with the alternate estimator

$$\bar{\theta}_n := \frac{1}{b_n} \sum_{j=1}^{b_n} \theta(P_{j,m_n}). \tag{15}$$

For example, suppose $\bar{\theta}_n$ is used in place of $\theta_n = \theta(P_n)$ in (9) to get the alternate estimator

$$S_{\text{OB-B}}(m_n, b_n) = \sqrt{\frac{1}{\kappa_2} \times m_n \times \frac{1}{b_n} \sum_{j=1}^{b_n} \left(\theta(P_{j,m_n}) - \bar{\theta}_n\right)^2}, \tag{16}$$

where κ_2 is a bias correction factor (see Theorem 4). Then, the weak limit analogous to (10) becomes

$$T_{\text{OB-B}}(m_n, b_n) := \sqrt{n}\left(\frac{\bar{\theta}_n - \theta(P)}{S_{\text{OB-B}}(m_n, b_n)}\right) \xrightarrow{\text{d}} T_{\text{OB-B}} \text{ as } n \to \infty, \tag{17}$$

where $T_{\text{OB-B}}$ exists and is given through Theorem 4. Importantly, and like $T_{\text{OB-S}}$, $T_{\text{OB-B}}$ does not depend on unknown parameters. The resulting $(1-\alpha)$ confidence interval becomes

$$\left(\bar{\theta}_n - F^{-1}_{T_{\text{OB-B}}}\left(\frac{\alpha}{2}\right)\frac{S_{\text{OB-B}}(m_n, b_n)}{\sqrt{n}},\ \bar{\theta}_n + F^{-1}_{T_{\text{OB-B}}}\left(1-\frac{\alpha}{2}\right)\frac{S_{\text{OB-B}}(m_n, b_n)}{\sqrt{n}}\right), \tag{18}$$

where $F_{T_{\text{OB-B}}}$ denotes the cdf of $T_{\text{OB-B}}$. The following theorem characterizes $T_{\text{OB-B}}$.

Theorem 4 (Su et al. (2023)) Suppose that Assumption 1 holds, and that $\beta := \lim_{n\to\infty} m_n/n > 0$. Assume also that $b_n \to b \in \{2, 3, \ldots, \infty\}$ as $n \to \infty$. Define

$$\chi^2_{\text{OB-B}}(\beta, b) := \begin{cases} \dfrac{1}{\kappa_2(\beta,\infty)} \dfrac{\beta^{-1}}{1-\beta} \displaystyle\int_0^{1-\beta} \left(\tilde{W}_u(\beta) - \frac{1}{1-\beta}\int_0^{1-\beta} \tilde{W}_s(\beta)\, ds\right)^2 du & b = \infty; \\ \dfrac{1}{\kappa_2(\beta,b)} \dfrac{1}{\beta} \dfrac{1}{b} \displaystyle\sum_{j=1}^{b} \left(\tilde{W}_{c_j}(\beta) - \frac{1}{b}\sum_{i=1}^{b} \tilde{W}_{c_i}(\beta)\right)^2 & b \in \mathbb{N} \setminus 1, \end{cases} \tag{19}$$

where $\tilde{W}_x(\beta) := W(x+\beta) - W(x), x \in [0, 1-\beta]$, $\{W(t), t \in [0,1]\}$ is the standard Brownian motion, $c_i := (i-1)\frac{1-\beta}{b-1}, i = 1, 2, \ldots, b$, and $\kappa_2(\beta, b)$ is the "bias-correction" factor given by

$$\kappa_2(\beta, b) := \begin{cases} 1 & \beta = 0; \\ 1 - 2\left(\dfrac{\beta}{1-\beta} \wedge 1\right) + \dfrac{1}{\beta}\left(\dfrac{\beta}{1-\beta} \wedge 1\right)^2 - \dfrac{2}{3}\dfrac{1-\beta}{\beta}\left(\dfrac{\beta}{1-\beta} \wedge 1\right)^3 & \beta > 0, b = \infty; \\ 1 - \dfrac{1}{b} - \dfrac{2}{b}\displaystyle\sum_{h=1}^{b} \left(1 - \frac{h}{b-1}\frac{1-\beta}{\beta}\right)^+ (1 - h/b) & \beta > 0, b \in \mathbb{N} \setminus 1. \end{cases} \tag{20}$$

Then, as $n \to \infty$,

$$S^2_{\text{OB-B}}(m_n, b_n) \xrightarrow{\text{d}} \sigma^2 \chi^2_{\text{OB-B}}(\beta, b); \tag{21}$$

and

$$T_{\text{OB-B}}(m_n,b_n) \xrightarrow{\text{d}} \begin{cases} \dfrac{1}{\sqrt{\chi^2_{\text{OB-B}}(\beta,b)}} \dfrac{1}{\beta} \dfrac{1}{(1-\beta)} \displaystyle\int_0^{1-\beta} (W(s+\beta)-W(s))\, ds & b=\infty; \\ \dfrac{1}{\sqrt{\chi^2_{\text{OB-B}}(\beta,b)}} \dfrac{1}{\beta} \dfrac{1}{b} \displaystyle\sum_{i=1}^{b} W(c_i+\beta)-W(c_i) & b \in \mathbb{N} \setminus 1, \end{cases} \tag{22}$$

where $c_i := (i-1)\frac{1-\beta}{b-1}, i = 1,2,\ldots,b$.

Another prominent batching variant arises due to using the *weighted area estimator* (Schruben 1983; Alexopoulos et al. 2007; Goldsman and Schruben 1990; Goldsman et al. 1990) in place of $S^2_{\text{OB-S}}(m_n,b_n)$ or $S^2_{\text{OB-B}}(m_n,b_n)$ to estimate the variance constant σ^2. See Su et al. (2023) for the corresponding weak limit and also for variants that result from using small batch sizes, that is, m_n such that $m_n/n \to 0$.

5 NOTES FOR FURTHER DISCUSSION AND STUDY

The following notes are salient and will be discussed during the oral presentation of this paper.

(a) Bootstrapping and batching are resampling devices that appear to have wide applicability for statistical inference in simulation settings. They are both easy to implement and remarkably effective in diverse settings so that their incorporation into general simulation software seems appropriate.

(b) Our presentation has assumed that all aspects of $F_{\varepsilon_n^*}$ can be easily obtained. This is generally not true in practice, especially when n is large. In such cases, implementers typically resort to what is called *bootstrap Monte Carlo* where aspects of $F_{\varepsilon_n^*}$ are estimated by drawing B observations from $F_{\varepsilon_n^*}$. Incorporating the error due to such sampling is generally a very challenging problem.

(c) For batching, we only treated the construction of confidence intervals. While batching methods for estimating ψ follow in a straightforward fashion, various aspects such as strong convergence and higher order accuracy have not been studied yet. Some other narrow questions like the theoretical characterization of the effect of batch sizes on coverage error and expected half-width in batching are also open.

(d) A frequent question among simulation practitioners is whether resampling is needed if "additional simulation runs can be performed" easily. This question becomes moot if efficiency (in the sense of teasing out more information from a given amount of data) is of interest. Batching and bootstrapping are methods that allow for efficient inference.

(e) The parameter θ is routinely not real-valued, that is, they can be vector-valued or function valued as in the examples we described. In such cases, the interpretation of simulation output $X_j, j = 1,2,\ldots,n$, and the ensuing inference, needs to be performed carefully even though the fundamental insights do not change.

(f) Our treatment assumes that the simulation output data $(X_1,X_2,\ldots,X_n)$ are in steady state. This is usually not the case in practice, leading to what has been called the *initial transient problem*. See Pasupathy and Schmeiser (2010) for an annotated bibliography on this problem. For appropriate inference, ideas from removing the initial transient need to be used in concert with batching, constituting what is an interesting research question.

(g) A consistent estimator of the variance constant σ^2 is neither needed nor preferred when constructing a confidence interval using batching. Interestingly, however, in the sequential context where the data $X_1,X_2,\ldots$ are revealed one by one, a risk-optimal estimator of θ might entail consistently estimating σ^2. See Pasupathy and Yeh (2020) for more.

(h) Virtually all discussion in this paper applies to estimators constructed in the context of *digital twins* (Biller et al. 2022).

(i) Parametric batching, analogous to parametric bootstrap (Cheng 2017), has not been sufficiently explored and should form a topic of future research.
(j) Bias estimation tends to be tricky and delicate, and should be performed with care. This issue is not specific to batching and similar caution has been issued even in the context of the bootstrap and the jackknife (Efron and Tibshirani 1994).
(k) There is a deep and interesting connection between variance estimation and certain types of input model uncertainty, as explained through semi-parametric estimation (Kosorok 2008).

ACKNOWLEDGMENTS

Raghu Pasupathy gratefully acknowledges the Office of Naval Research for support provided by the grants N000141712295 and 13000991. He also thanks Prof. Susan Hunter at Purdue Industrial Engineering for insightful discussions.

REFERENCES

Abdar, M. et al. 2021. "A Review Of Uncertainty Quantification In Deep Learning: Techniques, Applications And Challenges". *Information Fusion* 76:243–297.

Alexopoulos, C., N. T. Argon, D. Goldsman, G. Tokol, and J. R. Wilson. 2007. "Overlapping Variance Estimators for Simulation". *Operations Research* 55(6):1090–1103.

Barton, R. R. 2012. "Tutorial: Input Uncertainty in Output Analysis". In *Proceedings of the 2012 Winter Simulation Conference*, edited by R. P. O. R. C. Laroque, J. Himmelspach, and A. Uhrmacher, 1–12. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Biller, B., J. Xi, J. Yi, and P. Venditti. 2022. "Simulation: The Critical Technology in Digital Twin Development". In *Proceedings of the 2022 Winter Simulation Conference*, edited by B. Feng, G. Pedrielli, Y. Peng, S. Shashaani, E. Song, C. G. Corlu, L. H. Lee, E. P. Chew, T. Roeder, and P. Lendermann, 1340–1355. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Bottou, L., F. Curtis, and J. Nocedal. 2018. "Optimization Methods for Large-Scale Machine Learning". *SIAM Review* 60(2).

Calvin, J. M., and M. K. Nakayama. 2013. "Confidence Intervals for Quantiles with Standardized Time Series". In *Proceedings of the 2013 Winter Simulations Conference*, edited by R. Pasupathy, S.-H. Kim, A. Tolk, R. Hill, and M. E. Kuhl., 601–612. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Cheng, R. 2017. *Non-Standard Parametric Statistical Inference*. Oxford University Press.

Cheng, R. C., and W. Holland. 1997. "Sensitivity of Computer Simulation Experiments to Errors in Input Data". *Journal of Statistical Computation and Simulation* 1-4(57):219–241.

Chick, S. E. 2001. "Input Distribution Selection for Simulation Experiments: Accounting for Input Uncertainty". *Operations Research* 49(5):744–758.

Csörgö, M., and Révész. 1981. *Strong Approximations in Probability and Statistics*. Probability and mathematical statistics. New York: Academic Press.

Efron, B. 1979. "Bootstrap Methods: Another Look at the Jackknife". In *Breakthroughs in Statistics*, Springer Series in Statistics, 569–593. New York, NY: Springer New York.

Efron, B., and R. Tibshirani. 1994. *An Introduction to the Bootstrap*. Monographs on statistics and applied probability. New York: Chapman & Hall.

Glynn, P. W. 1998. "Strong Approximations in Queueing Theory". In *Asymptotic methods in probability and statistics*, 135–150. Elsevier.

Goldsman, D., M. Meketon, and L. W. Schruben. 1990. "Properties of Standardized Time Series Weighted Area Variance Estimators". *Management Science* 36(5):602–612.

Goldsman, D., and L. Schruben. 1990. "Note–New Confidence Interval Estimators Using Standardized Time Series". 36(3):393–397.

Hall, P. 1992. *Principles of Edgeworth Expansion*, 39–81. New York, NY: Springer New York.

Henderson, S. G. 2003. "Input Model Uncertainty: Why Do We Care And What Should We Do About It?". In *Proceedings of the 2003 Winter Simulation Conference*, edited by S. Chick, P. J. Sánchez, D. Ferrin, and D. J. Morrice, 1–12. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Hoel, P. G., S. C. Port, and C. J. Stone. 1986. *Introduction to Stochastic Processes*. Waveland Press.

Ingalls, R. 2014. "Introduction to Supply Chain Simulation". In *Proceedings of the 2014 Winter Simulation Conference*, 36–50: ACM.

Kosorok, M. R. 2008. *Introduction to Empirical Processes and Semiparametric Inference.* Springer.

Lam, H. 2016. "Advanced Tutorial: Input Uncertainty and Robust Analysis in Stochastic Simulation". In *Proceedings of the 2016 Winter Simulation Conference*, edited by T. M. K. Roeder, P. I. Frazier, R. Szechtman, E. Zhou, T. Huschka, and S. E. Chick, 1–12. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Najm, H. N. 2009. "Uncertainty Quantification and Polynomial Chaos Techniques in Computational Fluid Dynamics". *Annual review of fluid mechanics* 41:35–52.

Pasupathy, R., and B. W. Schmeiser. 2010. "DARTS — Dynamic Adaptive Random Target Shooting". In *Proceedings of the 2010 Winter Simulation Conference*, edited by B. Johansson, S. Jain, J. Montoya-Torres, J. Hugan, and E. Yücesan: Institute of Electrical and Electronics Engineers: Piscataway, New Jersey.

Pasupathy, R., D. Singham, and Y. Yeh. 2023. "Overlapping Batch Confidence Regions on the Steady-State Quantile Field". *Operations Research*.

Pasupathy, R., and Y. Yeh. 2020. "Risk-Efficient Sequential Simulation Estimators". In *Proceedings of the 2020 Winter Simulation Conference*, edited by K.-H. Bae, B. Feng, S. Kim, S. Lazarova-Molnar, Z. Zheng, T. Roeder, and R. Theising, 289–300. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Schruben, L. W. 1983. "Confidence Interval Estimation Using Standardized Time Series". *Operations Research* 31(6):1090–1108.

Serfling, R. J. 1980. *Approximation Theorems of Mathematical Statistics*. New York, New York: John Wiley & Sons, Inc.

Shao, J., and D. Tu. 2012. *The jackknife and bootstrap*. Springer Science & Business Media.

Soize, C. 2017. *Uncertainty Quantification*. Springer.

Su, Z., R. Pasupathy, Y. Yeh, and P. Glynn. 2023. "Overlapping Batch Confidence Intervals on Statistical Functionals Constructed from Time Series: Application to Quantiles, Optimization, and Estimation". *ACM Transactions on Modeling and Computer Simulation (TOMACS)*.

van der Vaart, A. W., and J. A. Wellner. 1996. *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer Series in Statistics. New York, NY: Springer New York.

Wasserman, L. 2004. *All of Statistics: A Concise Course in Statistical Inference*, Volume 26. Springer.

AUTHOR BIOGRAPHIES

RAGHU PASUPATHY is Professor of Statistics at Purdue University. His current research interests lie broadly in stochastic optimization, statistical inference, and Monte Carlo. He has been actively involved with the Winter Simulation Conference for the past 15 years. Raghu Pasupathy's email address is pasupath@purdue.edu, and his web page https://web.ics.purdue.edu/~pasupath contains links to papers, software codes, and other material.