Without Replacement Sampling for Particle Methods on Finite State Spaces

Rohan Shah Dirk P. Kroese

May 6, 2017

1 Introduction

Importance sampling is a widely used Monte Carlo technique that involves changing the probability distribution under which simulation is performed. Importance sampling algorithms have been applied to a variety of discrete estimation problems, such as estimating the locations of change-points in a time series (Fearnhead and Clifford, 2003), the permanent of a matrix (Kou and McCullagh, 2009), the $\mathcal{K}$-terminal network reliability (L'Ecuyer et al, 2011) and the number of binary contingency tables with given row and column sums (Chen et al, 2005).

Sequential importance resampling algorithms
(Doucet et al, 2001; Liu, 2001; Del Moral et al, 2006; Rubinstein and Kroese, 2017) combine importance sampling with some form of resampling. The aim of the resampling step is to remove samples that have an extremely low importance weight. In the case that the random variables of interest take on only finitely many values, forms of resampling that involve without-replacement sampling can be used (Fearnhead and Clifford, 2003).

The resulting algorithms are similar to particle-based algorithms with resampling, but the sampling and resampling steps are replaced by a single without-replacement sampling step. In the approach of Fearnhead and Clifford (2003), the authors use what we characterize as a *probability proportional to size* sampling design. These ideas have recently been incorporated into quasi Monte Carlo (Gerber and Chopin, 2015), as *sequential quasi Monte Carlo*. The *stochastic enumeration algorithm* of Vaisman and Kroese (2015) is another without-replacement sampling method, based on simple random sampling.

Use of without-replacement sampling has a number of advantages. This type of sampling tends to automatically compensate for deficiencies in the importance sampling density. If the importance sampling density wrongly assigns high probability to some values, then the consequence of this mistake is limited, as those values can still only be sampled once. This type of sampling can in principle reduce the effect of sample impoverishment (Gilks and Berzuini, 2001), as there is a lower limit to the number of distinct particles.

The first contribution of this paper is to highlight the links between the field of sampling theory and sequential Monte Carlo, in the discrete setting. In particular, we view the use of without-replacement sampling as an application of the famous *Horvitz–Thompson* estimator (Horvitz and Thompson, 1952), *unequal probability sampling designs* (Brewer and Hanif, 1983; Tillé, 2006) and multi-stage sampling. The links between these fields have received limited attention in the literature (Fearnhead, 1998; Carpenter et al, 1999; Douc et al, 2005), and the link with the Horvitz-Thompson estimator has not been made previously.

Our application of methods from sampling theory would likely be considered unusual by practitioners in that field. For example, in the Monte Carlo context, physical data collection is replaced by computation, so huge sample sizes become quite feasible. Also, it has traditionally been unusual to apply multi-stage methods with more than three stages of sampling, but in the Monte

Carlo context we apply such methods with thousands of stages.

The second contribution of this paper is to describe a new method of without-replacement sampling, using results from sampling theory. Specifically, we use the *Pareto design* (Rosén, 1997a,b) as a computationally efficient unequal probability sampling design. Our use of the Pareto design relies on results from Bondesson et al (2006).

The rest of this paper is organized as follows. Section 2 describes importance sampling and related particle algorithms. Section 3 gives an overview of sampling theory. Section 4 introduces the new sequential Monte Carlo method incorporating sampling without replacement, and lists some advantages and disadvantages of the proposed methodology. Section 5 gives some numerical examples of the effectiveness of without-replacement sampling. Section 6 summarizes our results and gives directions for further research.

2 Sequential Importance Resampling

2.1 Importance Sampling

Let $\mathbf{X}_d = (X_1, \ldots, X_d)$ be a random vector in $\mathbb{R}^d$, having density f with respect to a measure μ, e.g., the Lebesgue measure or a counting measure. Let $\mathbf{X}_t = (X_1, \ldots, X_t)$ be the first t components of $\mathbf{X}_d$. We wish to estimate the value of $\ell = \mathbb{E}_f \left[h \left(\mathbf{X}_d \right) \right]$, for some real-valued function h.

The crude Monte Carlo approach is to simulate n iid copies $\mathbf{X}_d^1, \ldots, \mathbf{X}_d^n$ according to f, and estimate ℓ by $n^{-1} \sum_{i=1}^n h \left(\mathbf{X}_d^i \right)$. However, there is no particular reason to use f as the sampling density. For any other density g such that $g(\mathbf{x}) = 0$ implies $h(\mathbf{x}) f(\mathbf{x}) = 0$,

$$\ell = \int h \left(\mathbf{x}_d \right) \frac{f \left(\mathbf{x}_d \right)}{g \left(\mathbf{x}_d \right)} g \left(\mathbf{x}_d \right) \mathrm{d}\mu \left(\mathbf{x}_d \right) = \int h \left(\mathbf{x}_d \right) w \left(\mathbf{x}_d \right) g \left(\mathbf{x}_d \right) \mathrm{d}\mu \left(\mathbf{x}_d \right),$$

where $w \left(\mathbf{x}_d \right) \overset{\text{def}}{=} \frac{f(\mathbf{x}_d)}{g(\mathbf{x}_d)}$ is the *importance weight*. If $\mathbf{X}_d^1, \ldots, \mathbf{X}_d^n$ are iid with density g, then the estimator

$$\widehat{\ell}_{\text{ub}} = n^{-1} \sum_{i=1}^n h \left(\mathbf{X}_d^i \right) w \left(\mathbf{X}_d^i \right) \tag{1}$$

is unbiased. This estimator is known as an *importance sampling* estimator (Marshall, 1956), with g being the *importance density*.

The quality of the importance sampling estimator depends on a good choice for the importance density. If h is a non-negative function, then the optimal choice is

$$g(\mathbf{x}) \propto h(\mathbf{x}) f(\mathbf{x}), \tag{2}$$

and the estimator has zero variance.

If the normalizing constant of f is unknown, then we can replace the weight function w with the unnormalized version $w_{\mathrm{r}}(\mathbf{x}) = \frac{cf(\mathbf{x}_d)}{g(\mathbf{x}_d)}$, where cf is a known function but c and f are unknown individually. In that case we use the asymptotically unbiased ratio estimator

$$\widehat{\ell}_{\mathrm{ratio}} = \frac{\sum_{i=1}^{n} h\left(\mathbf{X}_d^i\right) w_{\mathrm{r}}\left(\mathbf{X}_d^i\right)}{\sum_{i=1}^{n} w_{\mathrm{r}}\left(\mathbf{X}_d^i\right)}. \tag{3}$$

The central limit theorem (CLT) implies that if $\ell < \infty$ and $\mathrm{Var}_g\left(h\left(\mathbf{X}_d\right) w\left(\mathbf{X}_d\right)\right) < \infty$, then $\sqrt{n}\left(\widehat{\ell}_{\mathrm{ub}} - \ell\right)$ converges to a normal distribution as $n \to \infty$. By the strong law of large numbers, $\frac{1}{n}\sum_{i=1}^{n} w_{\mathrm{r}}\left(\mathbf{X}_d^i\right) \xrightarrow{a.s.} c$. By Slutsky's theorem and the asymptotic normality of $\widehat{\ell}_{\mathrm{ub}}$, $\sqrt{n}\left(\widehat{\ell}_{\mathrm{ratio}} - \ell\right)$ also converges to a normal distribution and is asymptotically unbiased.

Another context in which importance sampling can be applied is the estimation of the constant $c = \int cf(\mathbf{x})\,\mathrm{d}\mathbf{x}$. Importance sampling can still be applied if it is unclear how to simulate from f, and an unbiased estimator of c is

$$\widehat{c} = n^{-1}\sum_{i=1}^{n} w_{\mathrm{r}}\left(\mathbf{X}_d^i\right).$$

2.2 Sequential Importance Sampling

Let $\mathbf{x}_t = (x_1, \ldots, x_t)$. We adopt Bayesian notation, so that the interpretation of $f(\cdots)$ depends on its arguments, e.g., $f(x_3 \mid \mathbf{x}_2)$ is the density of X_3 conditional on $\mathbf{X}_2 = \mathbf{x}_2$. It can be difficult to directly specify an importance density on a high-dimensional space. The simplest method is often to build the distributions of the components sequentially. We first specify $g(x_1)$, then $g(x_2 \mid x_1), g(x_3 \mid \mathbf{x}_2)$, etc. If g is then used as an importance density, the importance weight is

$$w(\mathbf{x}) = \frac{f(x_1) f(x_2 \mid x_1) \cdots f(x_d \mid \mathbf{x}_{d-1})}{g(x_1) g(x_2 \mid x_1) \cdots g(x_d \mid \mathbf{x}_{d-1})}.$$

Early applications of this type of sequential build-up include Hammersley and Morton (1954) and Rosenbluth and Rosenbluth (1955). More recent uses include Kong et al (1994); Liu and Chen (1995). See Liu et al (2001) for further details.

It is often convenient to calculate the importance weights recursively as $u_1(x_1) = \frac{f(x_1)}{g(x_1)}$ and

$$u_t(\mathbf{x}_t) = u_{t-1}(\mathbf{x}_{t-1}) \frac{f(x_t \mid \mathbf{x}_{t-1})}{g(x_t \mid \mathbf{x}_{t-1})}, \qquad t = 2, \ldots, d. \tag{4}$$

It is clear that $u_d(\mathbf{x}_d) = w(\mathbf{x}_d)$. Note that computing u_t requires the factorization of $f(\mathbf{x}_t)$ in order to compute $f(x_t \mid \mathbf{x}_{t-1})$, which can be difficult. An alternative is to use a family $\{f_t(\mathbf{x}_t)\}_{t=1}^{d}$ of *auxiliary densities*, where it is

required that $f_d = f$. Using these densities we can compute the importance weights as $v_1 = \frac{f_1(x_1)}{g(x_1)}$ and

$$v_t(\mathbf{x}_t) = \frac{v_{t-1}(\mathbf{x}_{t-1}) f_t(\mathbf{x}_t)}{f_{t-1}(\mathbf{x}_{t-1}) g(x_t \mid \mathbf{x}_{t-1})}, \qquad t = 2, \ldots, d. \tag{5}$$

Note that $u_d(\mathbf{x}_d) = v_d(\mathbf{x}_d) = w(\mathbf{x}_d)$. We obtain u_t as a special case of v_t, where the auxiliary densities are the marginals of f. As v_t is more general, we use it to define our importance weights (unless otherwise stated). If the auxiliary densities are only known up to constant factors, then the unnormalized version of (5) involves setting $v_1(x_1) = \frac{c_1 f_1(x_1)}{g(x_1)}$ and

$$v_t(\mathbf{x}_t) = \frac{v_{t-1}(\mathbf{x}_{t-1}) c_t f_t(\mathbf{x}_t)}{c_{t-1} f_{t-1}(\mathbf{x}_{t-1}) g(x_t \mid \mathbf{x}_{t-1})}, \qquad t = 2, \ldots, d, \tag{6}$$

where the functions $\{c_t f_t(\mathbf{x}_t)\}$ are known, but the normalized functions $\{f_t(\mathbf{x}_t)\}$ may be unknown.

If $c_d = 1$ it is possible to evaluate f_d, and we can use the estimator $\widehat{\ell}_{\text{ub}}$ defined in (1), regardless of whether $c_t \neq 1$ for $t < d$. Otherwise, if f_d is known only up to a constant factor, we must use $\widehat{\ell}_{\text{ratio}}$. The variance of the corresponding importance sampling estimator is independent of the choice of auxiliary densities and of the constants $\{c_t\}$, but dependent on g. This will change in Section 2.3 with the introduction of resampling steps.

Sequential importance sampling can be performed by simulating all d components of $\mathbf{X}_d$ and repeating this process n times. Alternatively, we can simulate the first component of all n copies of $\mathbf{X}_d$. Then we simulate the second components conditional on the first, and so on. We adopt the second approach, as it leads naturally to *sequential importance resampling*.

2.3 Sequential Importance Resampling

It is often clear before all d components have been simulated that the final importance weight will be small. Samples with a small final importance weight will not contribute significantly to the final estimate. It makes sense to remove these samples before the full d components have been simulated. One way of achieving this is by resampling from the set of partially observed random vectors. In this context the partially observed vectors are known as *particles*.

Let $\{\mathbf{X}_t^i\}_{i=1}^n$ be the set of particles for a sequential importance sampling algorithm, and let $W_t^i = v_t(\mathbf{X}_t^i)$ be the importance weights in Section 2.2. Let $\{\mathbf{Y}_t^i\}_{i=1}^n$ be a sample of size n chosen with replacement from $\{\mathbf{X}_t^i\}_{i=1}^n$ with probabilities proportional to $\{W_t^i\}_{i=1}^n$, and let $\overline{W}_t = n^{-1} \sum_{i=1}^n W_t^i$. We can replace the variables $\{(\mathbf{X}_t^i, W_t^i)\}_{i=1}^n$ by $\{(\mathbf{Y}_t^i, \overline{W}_t)\}_{i=1}^n$ and continue the sequential importance sampling algorithm. This type of resampling is called *multinomial resampling*. The most famous use of multinomial resampling is in the *bootstrap filter* (Gordon et al, 1993). There are numerous other types of resampling, such as splitting or enrichment (Wall and Erpenbeck, 1959),

stratified resampling and residual resampling (Liu and Chen, 1995; Carpenter et al, 1999). See Liu et al (2001) for a recent overview.

3 Sampling Theory

Sampling theory aims to provide estimates about a finite population by examining a randomly chosen set of elements of the population, known as a *sample*. The population consists of N different objects known as *units*, denoted by the numbers $1, 2, \ldots, N$. We will assume that the size N of the population is known.

We assume that for each unit $i \in \{1, \ldots, N\}$ there is a fixed scalar value $y(i)$. These values are known only for the units selected in the sample. We wish to estimate some function $F(y(1), \ldots, y(N))$ of the values, most often the mean $\overline{y} = N^{-1} \sum_{i=1}^{N} y(i)$.

In its most abstract form, sampling theory is concerned with constructing random variables taking values in certain product sets. For example, a sample chosen *with replacement* corresponds to a random vector taking values in $\bigcup_{n=1}^{\infty} \{1, \ldots, N\}^n$. A sample of fixed size n chosen with replacement corresponds to a random variable taking values in $\{1, \ldots, N\}^n$. Define the *power set* $\mathcal{P}(X)$ as the set of all subsets of the set X. A sample *without replacement* corresponds to a random variable taking values in the power set $\mathcal{P}(\{1, \ldots, N\})$, and a sample without replacement of fixed size n corresponds to a random variable taking values in

$$\mathcal{S}_n = \{\mathbf{s} \in \mathcal{P}(\{1, \ldots, N\}) : |\mathbf{s}| = n\}.$$

These random variables have some distribution, and these types of distribution are known as *sampling designs*.

Units may be included in the sample with *equal probability* or *unequal probability*. Our focus in this section is on without-replacement sampling with a fixed sample size n and unequal probabilities. The probability of including unit i in the sample is called the *inclusion probability* of unit i, and denoted by $\pi(i)$. We assume that all the inclusion probabilities are strictly positive. The probability that both units i and j are included in the sample is denoted by $\pi(i, j)$. This is referred to as the *second-order inclusion probability*.

In order to apply unequal probability sampling designs, we assume that there are positive values $\{p(i)\}_{i=1}^{N}$ (known as *size variables*). For reasons specific to the application domain, these values are assumed to be positively correlated with the values in $\{y(i)\}_{i=1}^{N}$. In traditional sampling applications, $\{p(i)\}_{i=1}^{N}$ might correspond to (financially expensive) census of the population at a previous time, or estimates of the $\{y(i)\}_{i=1}^{N}$ which are easily obtainable but highly variable. In our setting the $\{p(i)\}_{i=1}^{N}$ play a similar role to the importance density in traditional importance sampling.

Unlike the $\{y(i)\}_{i=1}^{N}$, the $\{p(i)\}_{i=1}^{N}$ are known *before* sampling is performed. We aim to have $\{\pi(i)\}_{i=1}^{N}$ approximately proportional to $\{p(i)\}_{i=1}^{N}$, and therefore approximately proportional to the $\{y(i)\}_{i=1}^{N}$. For these reasons unequal

probability designs are also known as *probability proportional to size* (PPS) designs. Calculation of the inclusion probabilities for these designs is often difficult. See Tillé (2006) or Cochran (1977) for further details on general sampling theory.

3.1 The Horvitz–Thompson Estimator

Assume that we are using a without-replacement sampling design with fixed size n, and wish to estimate the total $N\overline{y}$ of the population values. If $\mathbf{s} \in \mathcal{S}_n$ is the chosen sample, then the *Horvitz–Thompson* estimator (Horvitz and Thompson, 1952) of the total is

$$\widehat{Y}_{\mathrm{HT}} = \sum_{i \in \mathbf{s}} y(i) \pi(i)^{-1}. \tag{7}$$

3.1.1 Systematic Sampling

Assume that $0 < p(i)$, and let $K = n^{-1} \sum_{i=1}^{N} p(i)$. We assume that all the $p(i)$ are smaller than K. Simulate U uniformly on $[0, K]$. The sample contains every unit j such that

$$\exists \text{ integer } l \geq 1, \text{ s. t. } \sum_{i=1}^{j-1} p(i) \leq U + lK \leq \sum_{i=1}^{j} p(i).$$

We have described *systematic sampling* (Madow and Madow, 1944) using a fixed ordering of units, in which case some pairwise inclusion probabilities are zero. Systematic sampling can also be performed using a random ordering, in which case every pairwise inclusion probability is positive.

The complexity of generating a systematic sample is $O(N)$ (Fearnhead and Clifford, 2003), which is asymptotically faster than generation of a Pareto sample.

3.1.2 Adjusting the Population

The existence of units with large size variables may preclude the existence of a sampling design with sample size n, for which $\pi(i) \propto p(i)$. As $\sum_{i=1}^{N} \pi(i) = n$, proportionality would require

$$\pi(i) = \frac{np(i)}{\sum_{i=1}^{N} p(i)}.$$

This may contradict $\pi(i) \leq 1$.

More generally, if a population does not satisfy the conditions for a particular design, units can be removed from the population and the sample size adjusted, until the conditions are satisfied. For example, consider the case where the Sampford design cannot be applied, because even though the $\{p(i)\}_{i=1}^{N}$ are positive, they cannot be rescaled to satisfy the conditions in Section **??**. We

iteratively remove the units with the largest size variable from the population, until the Sampford design can be applied with sample size $n-k$, where k is the number of units removed. The k removed units are deterministically included in the sample, and the Sampford design is applied to the remaining units, with sample size $n-k$.

4 Sequential Monte Carlo for Finite Problems

Our aim in this section is to develop a new sequential Monte Carlo technique that uses sampling without replacement. The algorithms we develop are based on the Horvitz–Thompson estimator and can be interpreted as an application of multistage sampling methods from the field of sampling theory.

We begin in Section 4.1 by describing our new sequential Monte Carlo technique without reference to any specific sampling design. In Section 4.2 we argue for the use of the Pareto design, with the inclusion probabilities being approximated by the inclusion probabilities of a related Sampford design. Section 4.5 gives some advantages and disadvantages of without-replacement sampling methods.

4.1 Sequential Monte Carlo Without Replacement

Assume that $\mathbf{X}_d = (X_1, \ldots, X_d)$ is a random vector in $\mathbb{R}^d$, taking values in the finite set $\mathscr{S}_d$ and having density f with respect to the counting measure on $\mathscr{S}_d$. We wish to estimate the value of

$$\ell = \mathbb{E}_f \left[h \left(\mathbf{X}_d \right) \right] = \sum_{\mathbf{x}_d \in \mathscr{S}_d} h \left(\mathbf{x}_d \right) f \left(\mathbf{x}_d \right).$$

Let $\mathbf{S}_i$ be a subset of the support of $\mathbf{X}_i = (X_1, \ldots, X_i)$. For $d \geq t > i \geq 1$, define $\mathscr{S}_t \left(\mathbf{S}_i \right)$ as

$$\mathscr{S}_t \left(\mathbf{S}_i \right) \stackrel{\text{def}}{=} \bigcup_{\mathbf{x}_i \in \mathbf{S}_i} \text{Support} \left(f \left(\mathbf{x}_t \mid \mathbf{x}_i \right) \right) = \text{Support} \left(\mathbf{X}_t \mid \mathbf{X}_i \in \mathbf{S}_i \right).$$

That is, $\mathscr{S}_t \left(\mathbf{S}_i \right)$ is the set of all extensions of a vector in $\mathbf{S}_i$ to a possible value for $\mathbf{X}_t$. For any value $\mathbf{x}_i$ of $\mathbf{X}_i$, let

$$\mathscr{S}_t \left(\mathbf{x}_i \right) = \text{Support} \left(\mathbf{X}_t \mid \mathbf{X}_i = \mathbf{x}_i \right).$$

It will simplify our algorithms to define

$$\mathscr{S}_1 \left(\emptyset \right) = \mathscr{S}_1 = \text{Support} \left(X_1 \right).$$

We begin by drawing a without-replacement sample from the set of all possible values of the first coordinate, X_1. That is, we select a sample $\mathbf{S}_1$ (of fixed or random size) from $\mathscr{S}_1$ according to a sampling design. For any $x_1 \in \mathscr{S}_1$ let $\pi^1 \left(x_1 \right)$ be the inclusion probability for element x_1 under this design. The specific choice of the sampling design is deferred to Section 4.2.

We now repeat this sampling process by drawing a without-replacement sample from the possible values of $\mathbf{X}_2$, conditional on the value of X_1 being contained in $\mathbf{S}_1$. That is, we select a without-replacement sample $\mathbf{S}_2$ from $\mathscr{S}_2(\mathbf{S}_1)$ according to a second sampling design. If $\mathbf{x}_2 \in \mathscr{S}_2(\mathbf{S}_1)$, let $\pi^2(\mathbf{x}_2)$ be the inclusion probability of element $\mathbf{x}_2$ under this second design, and so on.

In general, we draw a without-replacement sample $\mathbf{S}_t$ from $\mathscr{S}_t(\mathbf{S}_{t-1})$ according to a sampling design, and calculate the inclusion probabilities $\pi^t(\mathbf{x}_t)$. This process continues until a sample from $\mathscr{S}_d(\mathbf{S}_{d-1})$ is generated.

Algorithm 1: Sequential Monte Carlo without replacement

input : Density f, function h, sampling designs
output: Estimate of ℓ

$\mathbf{S}_0 \leftarrow \emptyset$
for $t = 1$ **to** d **do**
  $\mathbf{S}_t \leftarrow$ Sample from $\mathscr{S}_t(\mathbf{S}_{t-1})$ according to some design
  $\forall \mathbf{x}_t \in \mathbf{S}_t$ compute the inclusion probability $\pi^t(\mathbf{x}_t)$ of $\mathbf{x}_t$
return $\sum_{\mathbf{x}_d \in \mathbf{S}_d} h(\mathbf{x}_d) f(\mathbf{x}_d) \prod_{t=1}^{d} \pi^t(\mathbf{x}_d)^{-1}$

Abusing notation slightly, if $\mathbf{x}$ is a vector of dimension greater than t, then $\pi^t(\mathbf{x})$ will be interpreted as applying π^t to the first t coordinates. The only way for $(x_1, \ldots, x_d)$ to be selected as a member of $\mathbf{S}_d$ is if x_1 is contained in $\mathscr{S}_1$, (x_1, x_2) is contained in $\mathbf{S}_2$, (x_1, x_2, x_3) is contained in $\mathbf{S}_3$, etc. The final sample $\mathbf{S}_d$ is generated by a sampling design, for which the inclusion probability of $\mathbf{x}_d \in \mathbf{S}_d$ is $\prod_{t=1}^{d} \pi^t(\mathbf{x}_d)$. The Horvitz–Thompson estimator (See (7)) of ℓ is therefore

$$\widehat{\ell} = \sum_{\mathbf{x}_d \in \mathbf{S}_d} \underbrace{h(\mathbf{x}_d) f(\mathbf{x}_d)}_{y(i)} \underbrace{\left(\prod_{t=1}^{d} \pi^t(\mathbf{x}_d) \right)^{-1}}_{\pi(i)^{-1}}. \tag{8}$$

Computation of this estimator is described in Algorithm 1. The inclusion probabilities π^t depend on the sampling designs at the intermediate steps and the chosen samples. So the estimator is a function of the final set $\mathbf{S}_d$ and *implicitly* a function of $\mathbf{S}_1, \ldots, \mathbf{S}_{d-1}$. Appendix 7 shows that this estimator is unbiased. In practice, Algorithm 1 is implemented by maintaining a weight for each particle, and updating the particle weights by multiplying by $\frac{f(x_t \mid \mathbf{x}_{t-1})}{\pi^t(\mathbf{x}_t)}$ every time sampling is performed. That is,

$$\underbrace{\frac{f(\mathbf{x}_t)}{\prod_{i=1}^{t} \pi^i(\mathbf{x}_t)}}_{\text{new weight}} = \underbrace{\frac{f(\mathbf{x}_{t-1})}{\prod_{i=1}^{t-1} \pi^i(\mathbf{x}_t)}}_{\text{old weight}} \underbrace{\frac{f(x_t \mid \mathbf{x}_{t-1})}{\pi^t(\mathbf{x}_t)}}_{\text{new term}}. \tag{9}$$

Note the similarities between (9) and (4). The only difference is that the inclusion probabilities replace the importance density in the formula.

Example 1. To illustrate this methodology, assume that $d = 3$, that $\mathbf{X}_3$ is a random vector in $\{0,1,2\}^3$ with density f and that all our sampling designs select exactly two units. One possible realization of our proposed algorithm is shown in Figure 1. There are three possible values of X_1, and there are three possible samples of size 2. We select a sample $\mathbf{S}_1$ according to some sampling design. Assume that units 0 and 1 are chosen. So the initial sample $\mathbf{S}_1$ from $\mathscr{S}_1$ will be $\mathbf{S}_1 = \{0,1\}$. We compute the inclusion probabilities $\pi^1(0)$ and $\pi^1(1)$ of each of these units being contained in the sample $\mathbf{S}_1$.

Conditional on these values of X_1 there are six possible values of $\mathbf{X}_2$, which are

$$\mathscr{S}_2(\mathbf{S}_1) = \{(0,0),(0,1),(0,2),(1,0),(1,1),(1,2)\}.$$

The next step is to select a sample $\mathbf{S}_2$ of size 2 from these six units, according to some sampling design. Assume that the units $(0,1)$ and $(1,1)$ are chosen. We compute the inclusion probabilities $\pi^2(0,1)$ and $\pi^2(1,1)$ of each of these units being contained in the sample $\mathbf{S}_2$.

The final step is to sample $\mathbf{X}_3$ conditional on $\mathbf{X}_2$ being one of the values in $\mathbf{S}_2$. In this case $\mathscr{S}_3(\mathbf{S}_2)$ is

$$\{(0,1,0),(0,1,1),(0,1,2),(1,1,0),(1,1,1),(1,1,2)\}.$$

Assume that the sample of size 2 chosen is

$$\mathbf{S}_3 = \{(0,1,1),(1,1,1)\},$$

and compute the inclusion probabilities $\pi^3(0,1,1)$ and $\pi^3(1,1,1)$.

The overall inclusion probabilities of the two units in $\mathbf{S}_3$ are

$$\pi^1(0)\,\pi^2(0,1)\,\pi^3(0,1,1)$$

and

$$\pi^1(1)\,\pi^2(1,1)\,\pi^3(1,1,1).$$

In this case the Horzitz–Thompson estimator of ℓ is therefore

$$\begin{aligned} &h(0,1,1)\,f(0,1,1)\left(\pi^1(0)\,\pi^2(0,1)\,\pi^3(0,1,1)\right)^{-1} \\ &+ h(1,1,1)\,f(1,1,1)\left(\pi^1(1)\,\pi^2(1,1)\,\pi^3(1,1,1)\right)^{-1}. \end{aligned}$$ ■

We refer to the elements of the sets $\mathbf{S}_1,\ldots,\mathbf{S}_d$ as *particles*. A particle refers to an object that is chosen in a sampling step. We refer to elements of the sets $\mathscr{S}_1,\ldots,\mathscr{S}_d(\mathbf{S}_{d-1})$ as *units* to distinguish them from the particles. The term "unit" is traditional in survey sampling to refer to an element of a population, from which a sample is drawn.

If h is a non-negative function and

$$\prod_{t=1}^{d}\pi^t(\mathbf{x}_d) \propto h(\mathbf{x}_d)\,f(\mathbf{x}_d), \qquad \forall \mathbf{x}_d \in \mathscr{S}_d,$$

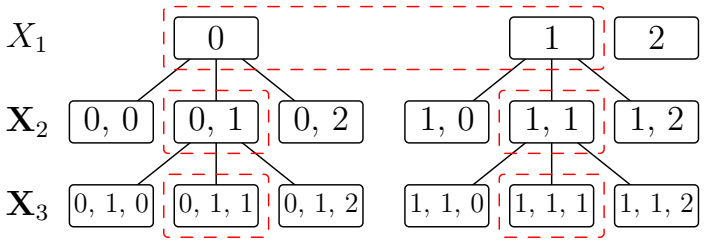


Figure 1: Illustration of the without-replacement sampling methodology, in the case that $d = 3$ and $\mathbf{X}_3$ is a random vector in $\{0, 1, 2\}^3$. The marked subsets of $X_1, \mathbf{X}_2$ and $\mathbf{X}_3$ are $\mathbf{S}_1$, $\mathbf{S}_2$ and $\mathbf{S}_3$.

we find that the estimator has zero variance. This formula is similar to the zero-variance importance sampling density given in (2). An alternative method of obtaining a zero–variance estimator is to choose the d sampling designs, such that at every sampling step, with probability 1 all the possible units are sampled. In this case the estimator corresponds to exhaustive enumeration of all the possible values of $\mathbf{X}_d$.

We can generalize to the case where $cf(\mathbf{x}_d)$ is known but the normalizing constant c is unknown, and the aim is to estimate c. The final estimator returned by Algorithm 1 should be changed to

$$\sum_{\mathbf{x}_d \in \mathbf{S}_d} cf(\mathbf{x}_d) \prod_{t=1}^{d} \pi^t(\mathbf{x}_d)^{-1}.$$

If the aim is to estimate $\mathbb{E}_f[h_d(\mathbf{X}_d)]$ but only $cf(\mathbf{x}_d)$ is known for some unknown constant c, then as in standard sequential Monte Carlo, we use the estimator

$$\left(\sum_{\mathbf{x}_d \in \mathbf{S}_d} h(\mathbf{x}_d) cf(\mathbf{x}_d) \prod_{t=1}^{d} \pi^t(\mathbf{x}_d)^{-1} \right) \left(\sum_{\mathbf{x}_d \in \mathbf{S}_d} cf(\mathbf{x}_d) \prod_{t=1}^{d} \pi^t(\mathbf{x}_d)^{-1} \right)^{-1}. \quad (10)$$

This estimator is no longer unbiased.

4.2 Choice of Sampling Design

So far we have not discussed the choice of the sampling design. Our preferred choice is to simulate from the Pareto design, due to the ease of simulation. The inclusion probabilities are difficult to calculate, but we use the connections to the Sampford design, for which the inclusion probabilities are easy to calculate, to avoid this problem.

The pdfs of the Sampford and Pareto designs (Equations (**??**) and (**??**)) differ only in the last term of the product. Bondesson et al (2006) shows that if

$$D = \sum_{i=1}^{N} p(i)(1 - p(i)) \text{ is large and } \sum_{i=1}^{N} p(i) = n, \quad (11)$$

then the constants $c(i)$ in (**??**) are approximately equal to $1-p(i)$, which is the corresponding term in (**??**). This implies that the Pareto and Sampford designs are almost identical in this case. The condition that D be large is generally equivalent to requiring that n and $N-n$ are not small. More importantly, if (11) holds then the inclusion probabilities of the Pareto design are approximately $\{p(i)\}_{i=1}^{N}$.

We normalize the size variables to sum to n, simulate directly from the Pareto design and assume that the inclusion probabilities are the normalized size variables. This choice has very significant computational advantages. It allows for fast sampling *and* fast computation of the inclusion probabilities.

In theory this approximation to the inclusion probabilities will introduce bias into our algorithms, but empirically this bias is found to be negligible. We emphasize that it is the approximation of the *inclusion probabilities* that is important. The fact that the designs themselves are almost identical is only a means of obtaining this approximation for the inclusion probabilities.

In general the condition

$$\sum_{i=1}^{N} p(i) = n, \qquad \text{and} \qquad 0 < p(i) < 1, \ \forall 1 \leq i \leq N \tag{12}$$

required by the Sampford design will not hold, and this cannot always be fixed by rescaling the $\{p(i)\}$. In these cases we take the approach outlined in Section 3.1.2. We deterministically select the unit corresponding to the largest size variable $p(i)$. If the $\{p(i)\}$ for the remaining units (suitably rescaled to sum to $n-1$) lie between 0 and 1 then the remaining $n-1$ units are selected according to the Pareto design. Otherwise, units are chosen deterministically until these conditions are met, and the design can be applied. The units chosen deterministically will have inclusion probability 1.

Example 2. We let $N = 1000$ and simulated the size variables $\{p(i)\}_{i=1}^{N}$ uniformly on $[0,1]$. For a fixed value of n, these size variables were rescaled to sum to n and used as the size variables for Pareto and Sampford designs. The inclusion probabilities $\left\{\pi_n^{\text{Pareto}}(i)\right\}_{i=1}^{N}$ of the Pareto design were computed. Recalling that the inclusion probabilities of the Sampford design are $\{p(i)\}_{i=1}^{N}$, we calculated

$$\max_{1 \leq i \leq N} \frac{\left|p(i) - \pi_n^{\text{Pareto}}(i)\right|}{\pi_n^{\text{Pareto}}(i)}. \tag{13}$$

This was repeated for different values of n, and the results are shown in Figure 2. It is clear that the inclusion probabilities for the Pareto design and the Sampford design are extremely close. Calculating the Pareto inclusion probabilities out to $n = 200$ required 1000 base-10 digits of accuracy. As a result these calculations were extremely slow. ■

It remains to specify the size variables $\{p(i)\}$ for the design. If we wish to use an importance sampling density g to specify the size variables, then for sampling

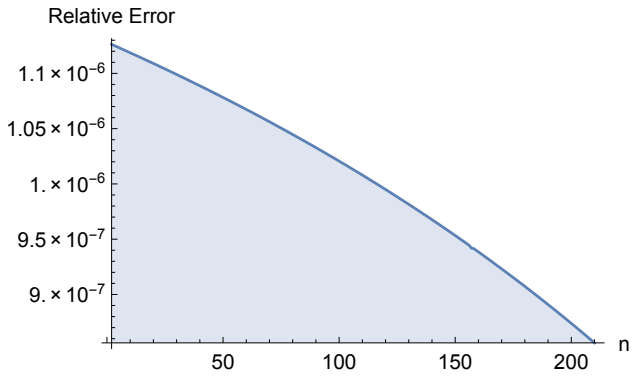


Figure 2: Maximum relative error (as measured by (13)) when approximating the Pareto inclusion probabilities by $\{p(i)\}_{i=1}^{N}$. The x-axis is the sample size n.

at step t we propose (with a slight abuse of notation) to use size variables

$$p(\mathbf{x}_t) = \frac{g(\mathbf{x}_t)}{\prod_{i=1}^{t-1} \pi^i(\mathbf{x}_{t-1})}. \tag{14}$$

The size variables can also be written recursively as

$$p(\mathbf{x}_t) = p(\mathbf{x}_{t-1}) \frac{g(x_t \mid \mathbf{x}_{t-1})}{\pi^{t-1}(\mathbf{x}_{t-1})}. \tag{15}$$

Equation (15) is similar to (4).

These size variables give a straightforward method for converting an importance sampling algorithm into a sequential Monte Carlo without replacement algorithm, shown in Algorithm 2. For simplicity, Algorithm 2 omits the details relating to the deterministic inclusion of some units if (12) fails to hold. If the sample size n is greater than the number N of units, then the entire population is sampled and every inclusion probability is 1.

4.3 Merging of Equivalent Units

When applying without-replacement sampling algorithms, there are often multiple values which will have identical contributions to the final estimator. Let $h^*(\mathbf{x}_t) = \mathbb{E}[h(\mathbf{X}_d) \mid \mathbf{X}_t = \mathbf{x}_t]$. That is, when the sample is taken on Line 3 of Algorithm 1, there may be values $\mathbf{x}_t$ and $\mathbf{x}'_t$ in $\mathscr{S}_t(\mathbf{S}_{t-1})$, for which $h_t^*(\mathbf{x}'_t) = h_t^*(\mathbf{x}_t)$. In such a case the units can be merged, reducing the set of units to which the sampling design is applied. Before continuing, we give a short example illustrating how this idea works.

Example 3. Consider again the example shown in Figure 1 of a random vector taking values in $\{0, 1, 2\}^3$. For simplicity we use the conditional Poisson

Algorithm 2: Sequential Monte Carlo without replacement, using an approximate Sampford design and an importance density

input : Density f, function h, importance density g, sample size n
output: Estimate of ℓ

$\mathbf{S}_0 \leftarrow \emptyset$
for $t = 1$ **to** d **do**
  Compute $\{p(\mathbf{x}_t) : \mathbf{x}_t \in \mathscr{S}_t(\mathbf{S}_{t-1})\}$ and normalize to sum to n
  $\mathbf{S}_t \leftarrow$ Pareto sample of $\min\{n, |\mathscr{S}_t(\mathbf{S}_{t-1})|\}$ from $\mathscr{S}_t(\mathbf{S}_{t-1})$ with size variables $\{p(\mathbf{x}_t)\}$
  `// The approx. inclusion probability of`
  `//` $\mathbf{x}_t \in \mathbf{S}_t$ `is` $\pi^t(\mathbf{x}_t) = p(\mathbf{x}_t)$ `or` $\pi^t(\mathbf{x}_t) = 1$
return $\sum_{\mathbf{x}_d \in \mathbf{S}_d} h(\mathbf{x}_d) f(\mathbf{x}_d) \prod_{t=1}^{d} \pi^t(\mathbf{x}_d)^{-1}$

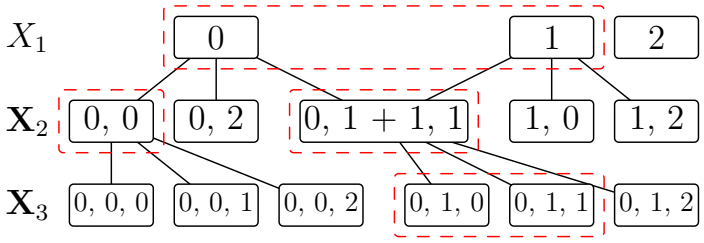


Figure 3: Illustration of merging of units in Example 3. Here $d = 3$ and $\mathbf{X}_3$ is a random vector in $\{0, 1, 2\}^3$. The merged unit is represented by $(0, 1)$, but could also be represented by $(1, 1)$. The marked subsets of $X_1, \mathbf{X}_2$ and $\mathbf{X}_3$ are $\mathbf{S}_1$, $\mathbf{S}_2$ and $\mathbf{S}_3$.

sampling design. Let

$$h(0,1,0) = 6, h(0,1,1) = h(0,1,2) = 0.1,$$
$$h(1,1,0) = 2, h(1,1,1) = h(1,1,2) = 2.1,$$

and let h be equal to 2 for all other values of $\mathbf{X}_3$. Assume that f is the uniform density on $\{0,1,2\}^3$, so that the value we aim to estimate is 2.015. Let $g(x_1) = \frac{1}{3}$, $g(\mathbf{x}_2) = \frac{1}{9}$ and $g(\mathbf{x}_3) = \frac{1}{27}$. This implies that the inclusion probabilities at iteration $t = 1$ are $\frac{2}{3}$, and the inclusion probabilities of all the units in $\mathscr{S}_2(\mathbf{S}_2)$ are $\frac{1}{3}$.

At iteration $t = 2$ the sampling design is applied to $\mathscr{S}_2(\mathbf{S}_2)$, which includes $(0,1)$ and $(1,1)$. In this example we have

$$h_3^*(0,1) = h_3^*(1,1) = \frac{62}{30}.$$

Both units have the same expected contribution to the final estimator, and if this was known, we could replace the pair of units by a single unit $(0,1)+(1,1)$, where the merged unit is *represented* by $(0,1)$ or $(1,1)$. After the merging we have the situation shown in Figure 3, where we have chosen to represent the merged unit as $(0,1)$. We could choose to represent the merged unit by $(1,1)$, in which case the units underneath the merged unit would be $(1,1,0), (1,1,1)$ and $(1,1,2)$. The value of the size variable for the merged unit is

$$\frac{g(0,1)}{\pi^1(0)} + \frac{g(1,1)}{\pi^1(1)} = \frac{1}{3}.$$

We must also double the contribution of the merged unit to the final estimator, as it represents two units.

If units $(0,1,0)$ and $(0,1,1)$ are chosen in the third step, the value of the estimator is

$$\frac{\mathbf{12}}{\mathbf{27}} \left(\pi^1(1)\,\pi^2((0,1)+(1,1))\,\pi^3(0,1,0)\right)^{-1}$$
$$+ \frac{\mathbf{0.2}}{\mathbf{27}} \left(\pi^1(1)\,\pi^2((0,1)+(1,1))\,\pi^3(0,1,0)\right)^{-1}.$$

The bolded values are $2h(0,1,0)\,f(0,1,0)$ and
$2h(0,1,1)\,f(0,1,1)$, where the factor of 2 accounts for the merging.

Assume that units 0 and 1 are initially selected. If no merging is performed, then the variance of estimator is 0.52. If the merging step is performed, and the merged unit is represented by $(0,1)$, then the variance of the estimator is 1.04. If the merged unit is represented by $(1,1)$, then the variance of the estimator is 0.0048. ■

As in Section 4.2, let g be the importance function, for simplicity assumed to be normalized. In order to formalize the idea of merging equivalent units, we add additional information to all the sample spaces and the samples chosen from them. The new units will be triples, where the first entry $\mathbf{x}_t$ represents

the value of the unit, the second entry w can be interpreted as the importance weight, and the third entry p can be interpreted as the size variable.

With slight abuse of notation, we redefine the sets $\mathbf{S}_0, \ldots, \mathbf{S}_d$ to account for this extra structure. Let

$$\mathscr{T}_1 = \mathscr{T}_1(\emptyset) = \{(x_1, f(x_1), g(x_1)) : x_1 \in \mathscr{S}_1\}.$$

The initial sample $\mathbf{S}_0$ is chosen from $\mathscr{T}_1$, with probability proportional to the third component. Assume that sample $\mathbf{S}_{t-1}$ has been chosen, and let

$$\mathscr{T}_t(\mathbf{S}_{t-1}) = \left\{ \left(\mathbf{x}_t, w\frac{f(x_t \mid \mathbf{x}_{t-1})}{\pi^{t-1}(\mathbf{x}_{t-1})}, p\frac{g(x_t \mid \mathbf{x}_{t-1})}{\pi^{t-1}(\mathbf{x}_{t-1})} \right) : (\mathbf{x}_{t-1}, w, p) \in \mathbf{S}_{t-1}, \mathbf{x}_t \in \text{Support}(\mathbf{X}_t \mid \mathbf{X}_{t-1} = \mathbf{x}_{t-1}) \right\}. \tag{16}$$

Note that (16) incorporates the recursive equations in (9) and (15). Using these definitions, we can sample $\mathbf{S}_2$ from $\mathscr{T}_2(\mathbf{S}_1)$, $\mathbf{S}_3$ from $\mathscr{T}_3(\mathbf{S}_2)$, etc. We can now state Algorithm 3. If the merging step on Line 4 is omitted, then this algorithm is in fact a restatement of Algorithm 1 using different notation. The merging rule on Line 4 is given in Proposition 4.1.

Algorithm 3: Sequential Monte Carlo without replacement, with merging

input : Density f, function h, sampling designs
output: Estimate of ℓ

1 $\mathbf{S}_0 \leftarrow \emptyset$
2 **for** $t = 1$ **to** d **do**
3 $\quad U \leftarrow \mathscr{T}_t(\mathbf{S}_{t-1})$
4 $\quad$ Modify U by merging pairs according to Proposition 4.1
5 $\quad \mathbf{S}_t \leftarrow$ Sample from U according to some design, with size variables $\{p\colon (\mathbf{x}_t, w, p) \in U\}$
6 $\quad \forall \mathbf{x}_t \in \mathbf{S}_t$ compute the inclusion probability $\pi^t(\mathbf{x}_t)$ of $\mathbf{x}_t$
7 **return** $\sum_{(\mathbf{x}_d, w, p) \in \mathbf{S}_d} \frac{h(\mathbf{x}_d) w}{\pi^d(\mathbf{x}_d)}$

Proposition 4.1. *If units* $(\mathbf{x}_t, w, p)$ *and* $(\mathbf{x}'_t, w', p')$ *in* $\mathscr{T}_t(\mathbf{S}_{t-1})$ *satisfy* $h^*(\mathbf{x}_t) = h^*(\mathbf{x}'_t)$*, they can be removed and replaced by the unit*

$$(\mathbf{x}_t, w + w', p + p') \qquad \textit{or} \qquad (\mathbf{x}'_t, w + w', p + p').$$

The final estimator is still unbiased.

Proof. See Appendix 8. □

The value $p+p'$ in the third component of the merged unit can be replaced by any positive value, without biasing the resulting estimator. We gave an example of this type of merging in Example 3. Example 3 is unusual, as it merges units

for which the function h takes very different values. A more common way for $h^*(\mathbf{x}_t) = h^*(\mathbf{x}'_t)$ to occur is if

$$h(\mathbf{X}_d) \mid \mathbf{X}_t = \mathbf{x}_t \quad \stackrel{\mathrm{d}}{=} \quad h(\mathbf{X}_d) \mid \mathbf{X}_t = \mathbf{x}'_t \,. \tag{17}$$

Example 4. We now continue Example 3, using the new definitions of $\mathscr{T}_1$ and $\mathscr{T}_t(\mathbf{S}_{t-1})$. As shown in Figure 3, the six units in $\mathscr{T}_2(\mathbf{S}_1)$ become five after the merging step. Of these, two units are chosen to be in $\mathbf{S}_2$; these units are

$$\left((0,0), \frac{f(0,0)}{\pi^1(0)}, \frac{g(0,0)}{\pi^1(0)}\right) = \left((0,0), \frac{1}{6}, \frac{1}{6}\right)$$

and

$$\left((0,1), \frac{f(0,1)}{\pi^1(0)} + \frac{f(1,1)}{\pi^1(1)}, \frac{g(0,1)}{\pi^1(0)} + \frac{g(1,0)}{\pi^1(1)}\right) = \left((0,1), \frac{1}{3}, \frac{1}{3}\right).$$

The other possible value for the merged unit is $\left((1,1), \frac{1}{3}, \frac{1}{3}\right)$. ■

Algorithm 3 does not specify a sampling design. We suggest the use of a Pareto design, with the inclusion probabilities being approximated by those of a Sampford design, as discussed in Section 4.2. However, these types of merging step can applied with any sampling design, including the systematic sampling suggested in Fearnhead and Clifford (2003).

4.4 Links with the work of Fearnhead and Clifford (2003)

Carpenter et al (1999) and Fearnhead and Clifford (2003) propose a resampling method which they name "stratified sampling". This method is *systematic sampling* (Section 3.1.1) with probability proportional to size, with large units included deterministically. This method has a long history in sampling theory (Madow and Madow, 1944; Madow, 1949; Hartley and Rao, 1962; Iachan, 1982). That large units must be included deterministically in a PPS design is well known in the sampling theory literature (Sampford, 1967; Rosén, 1997b; Aires, 2000).

From a sampling theory point of view, the optimality result of Fearnhead and Clifford (2003) can be paraphrased as "sampling with probability proportional to size is optimal". As the optimality criteria relates only to the inclusion probabilities, the Sampford design satisfies this condition just as well as systematic sampling. The conditional Poisson and Pareto designs will approximately satisfy this condition, especially when n is large.

In the approach of Fearnhead and Clifford (2003), units with large weights are included deterministically, and their weights are unchanged by the sampling step. All other units are selected stochastically, and are assigned the same weight if they are chosen.

This can be interpreted as an application of the Horvitz–Thompson estimator. With these observations, the approach of Fearnhead and Clifford (2003) can be interpreted as an application of Algorithm 1 using systematic sampling.

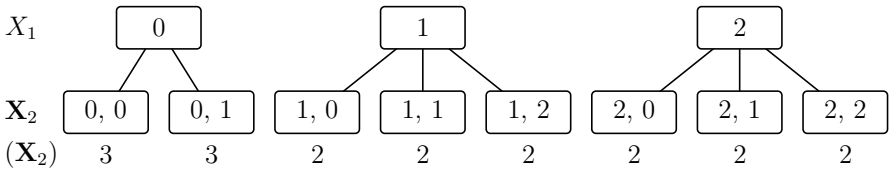


Figure 4: A pathological example, where increasing the sample size from 1 to 2 increases the variance.

4.5 Advantages and Disadvantages

Like many methods that involve interacting particles (e.g., multinomial resampling algorithms), the sample size used to generate the estimator is fixed at the start and cannot be increased without recomputing the entire estimator. By contrast, additional samples can be added to an importance sampling estimator and some sequential Monte Carlo estimators (Brockwell et al, 2010; Paige et al, 2014), if a lower variance estimator is desired.

Without–replacement sampling allows the use of particle merging steps, which can dramatically improve the variance of the resulting estimators, *while also* lowering the simulation effort required. Such merging steps are not possible with more classical types of resampling.

If particle merging is used then the resulting estimator is specialized to the particular function h, as the units that can be merged depend on the function h. By contrast the weighted sample generated by an importance sampling estimator can, in theory, be used to estimate the expectation of a different function h. In practice, even importance sampling estimators can be optimized by discarding particles as soon as they are known to make a contribution of zero to the final estimator. In such cases even the importance sampling algorithm is specialized to the function h.

The choice of the sample size is far more complicated than for traditional importance sampling algorithms. A large enough sample size will return a zero–variance estimator, but this sample size is generally impractical. However, it is unclear whether the variance of the estimator *must* decrease as n decreases. This is particularly true when merging steps are added to the algorithm. The following simple example illustrates this.

Example 5. Take the example shown in Figure 4, where $\mathbf{X}_2$ takes on eight values and the values of $h(\mathbf{x}_2)$ are as given. Assume that $f(\mathbf{x}_2) = \frac{1}{8}$ for each of these values. Let the size variables be $p(x_1) = p(\mathbf{x}_2) = 1$. if $n = 1$ the estimator has zero variance. However with $n = 2$ the estimator has non-zero variance; the value to be estimated is $\frac{18}{8}$, but if units $(0, 0)$ and $(0, 1)$ are selected, the estimator is $2.8125 \neq \frac{18}{8}$. So increasing the sample size has increased the variance from zero to some non-zero value.

Despite the previous remarks about choice of sample size, in practice the variance of the estimator decreases as n increases. As the variance of the estimator will reach 0 for finite n, it must be possible to observe a better than

n^{-1} decay in the variance of the estimator. This is in some sense a trivial statement, as there exists a sample size k, such that the estimator has non-zero variance with this sample size, but for sample size $k+1$ the estimator has zero variance. However, we observe more rapid decreases in practical applications of these types of estimators. For an example see the simulation results in Section 5.2.

5 Examples

In our examples we compare estimators using their *work-normalized variance*, defined as

$$\mathrm{WNV}\left(\widehat{\ell}\right) = T\mathrm{Var}\left(\widehat{\ell}\right),$$

where T is the simulation time to compute the estimator. In practice the terms in the definitions of WNRV are replaced by their estimated values.

5.1 Change Point Detection

We consider the discrete-time change-point model used in the example in Section 5 of Fearnhead and Clifford (2003). In this model there is some underlying real-valued signal $\{U_t\}_{t=1}^{\infty}$. At each time-step, this signal may maintain its value from the previous time, or change to a new value. The observations $\{Y_t\}_{t=1}^{\infty}$ combine $\{U_t\}_{t=1}^{\infty}$ with some measurement error. This measurement error will sometimes generate outliers, in which case Y_t is conditionally independent of U_t. This model is a type of hidden Markov model.

Let $X_t = (C_t, O_t)$ be the underlying Markov chain, where both C_t and O_t take values in $\{1, 2\}$, and let $\{V_t\}_{t=1}^{\infty}$ and $\{W_t\}_{t=1}^{\infty}$ be independent sequences of standard normal random variables. Let

$$U_t = \begin{cases} U_{t-1} & \text{if } C_t = 1, \\ \mu + \sigma V_t & \text{if } C_t = 2. \end{cases}$$

If $C_t = 2$, the signal changes to a new value, distributed according to $N\left(\mu, \sigma^2\right)$. Otherwise, the signal maintains the previous value. Let

$$Y_t = \begin{cases} U_{t-1} + \tau_1 W_t & \text{if } O_t = 1, \\ \nu + \tau_2 W_t & \text{if } O_t = 2. \end{cases}$$

If $O_t = 2$, the observed value is an outlier and is distributed according to $N\left(\nu, \tau_2^2\right)$. Otherwise, the measurement reflects the underlying signal, with error distributed according to $N\left(0, \tau_1^2\right)$.

It remains to specify the distribution of the Markov chain $\{X_t\}_{t=1}^{\infty}$. In the example given in Fearnhead and Clifford (2003), the $\{C_t\}_{t=1}^{\infty}$ are assumed iid, and $\{O_t\}_{t=1}^{\infty}$ is a Markov chain, with

$$\mathbb{P}\left(O_t = 2 \mid O_t = 2\right) = 0.75, \mathbb{P}\left(O_t = 2 \mid O_t = 1\right) = 1/250$$
$$\mathbb{P}\left(C_t = 2\right) = 1/250.$$

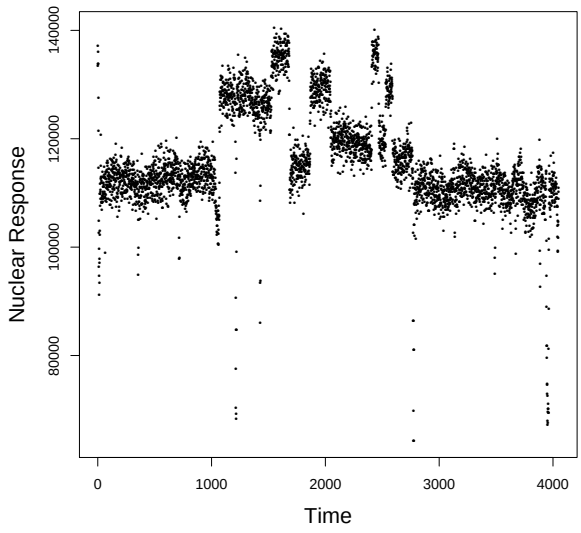


Figure 5: The well-log data from Ó Ruanaidh and Fitzgerald (1996).

In this example there is some integer $d > 1$, and the aim is to estimate the marginal distributions of $\{C_t\}_{t=1}^d$ and $\{O_t\}_{t=1}^d$, conditional on $\mathbf{Y}_d = \{Y_t\}_{t=1}^d$.

For the purposes of this example we apply a version of Algorithm 3 that involves some minor changes. See Appendix 9 for further details. The final algorithm is given as Algorithm 6 in Appendix 9. This algorithm contains the merging steps outlined in Fearnhead and Clifford (2003), which operate on principles similar to those described in Section 4.3.

For this example we used the well-log data from Ó Ruanaidh and Fitzgerald (1996); Fearnhead and Clifford (2003), and aimed to estimate the posterior probabilities

$$\mathbb{P}\left(C_t = 2 \mid \mathbf{Y}_d = \mathbf{y}_d\right) \qquad \text{and} \qquad \mathbb{P}\left(O_t = 2 \mid \mathbf{Y}_d = \mathbf{y}_d\right),$$

which are the posterior probabilities that there is a change or an outlier at time t, respectively. For this dataset $d = 4050$. The data are shown in Figure 5.

We applied two methods to this problem. The first was the method of Fearnhead and Clifford (2003), and the second was our without-replacement sampling method, using a Pareto design as an approximation to the Sampford design. Both of these methods can be viewed as specializations of Algorithm 6, where the method of Fearnhead and Clifford (2003) uses systematic sampling. Both methods were applied 1000 times with $n = 100$. Each run of either method produces 4050 outlier probability estimates and 4050 change-point probability estimates, so we provide a summary of the results. Note that the sample size required to produce a zero-variance estimator is on the order of 2^{4050} in this

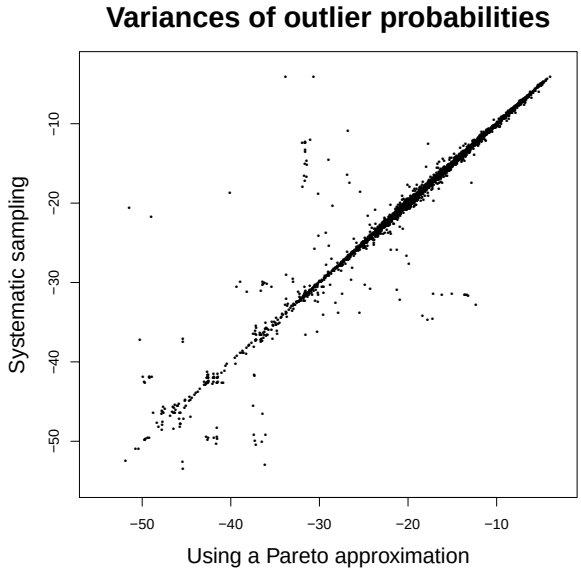


Figure 6: The variances of the estimated posterior outlier probabilities, using both methods.

case, which is clearly infeasible.

For the 4050 outlier probabilities, our method had a lower variance for 1656 estimates, and a higher variance for 2393 estimates. For the 4050 change-point estimates, our method had a lower variance for 1915 estimates, and a higher variance for 2121. This suggests that systematic sampling performs better than our approximation. Figure 6 shows the variances of every outlier probability estimate, under both methods. This plot suggests that if systematic sampling performs better, the improvement is small. The results for the change-points are similar.

Recall from Section 4.5 that the optimality condition of Fearnhead and Clifford (2003) can be paraphrased as "sampling with probability proportional to size is optimal". So, to the extent that the approximation for the inclusion probabilities of the Pareto design (See Section 4.2) holds, we expect that both methods should have similar performance. This is reflected in the simulation results. There is some discrepancy for estimates of the outlier probabilities, where systematic sampling performs slightly better. This may be due to the somewhat small sample size.

Fearnhead and Clifford (2003) also applied the mixture Kalman filter (Chen and Liu, 2000) and a multinomial resampling algorithm. They showed that the without-out replacement sampling approach significantly outperformed the alternatives. As our approach has equivalent performance to the method of Fearnhead and Clifford (2003), we do not consider these alternatives further.

5.2 Network Reliability

5.2.1 Without Particle Merging

We now give an application of without-replacement sampling to the $\mathcal{K}$-terminal network reliability estimation problem. Assume we have some known graph $\mathcal{G}$ with m edges, which are enumerated as $e_1, \ldots, e_m$. We define a random subgraph $\mathbf{X}$ of $\mathcal{G}$, with the same vertex set. Let $X_1, \ldots, X_m$ be independent binary random variables representing the states of the edges of $\mathcal{G}$. With probability θ_i variable $X_i = 1$, in which case edge e_i of $\mathcal{G}$ is included in $\mathbf{X}$. For a fixed set $\mathcal{K} = \{v_1, \ldots, v_k\}$ of vertices of $\mathcal{G}$, the $\mathcal{K}$-terminal network unreliability is the probability ℓ that these vertices are not connected; that is, they do not all lie in the same connected component of $\mathbf{X}$. As computation of this quantity is in general $\#P$ complete, it often cannot be computed and must be estimated. If the probabilities $\{\theta_i\}$ are close to 1 then the unreliability is close to zero, and the problem is one of estimating a rare-event probability.

One of the best methods currently available for estimating the unreliability ℓ is *approximate zero–variance importance sampling* (L'Ecuyer et al, 2011). This method is based on *mincuts*. In the $\mathcal{K}$-terminal reliability context a *cut* of a graph g is a set $\mathbf{c}$ of edges of g such that the vertices in $\mathcal{K}$ do not all lie in the same component of $g \setminus \mathbf{c}$. A *mincut* is a cut $\mathbf{c}$ such that no proper subset of $\mathbf{c}$ is also a cut.

In L'Ecuyer et al (2011) the states of the edges are simulated sequentially using state-dependent importance sampling. Assume that the values $x_1, \ldots, x_t$ of $X_1, \ldots, X_t$ are already known. Let $\mathcal{G}(x_1, \ldots, x_t)$ be the subgraph of $\mathcal{G}$ obtained by removing all edges e_i where $i \leq t$ and $x_i = 0$. Let $\mathcal{C}(x_1, \ldots, x_t)$ be the set of all mincuts of $\mathcal{G}(x_1, \ldots, x_t)$ that do not contain edges $e_1, \ldots, e_t$. Let $E(\cdot)$ be the event that a set of edges is missing from $\mathbf{X}$. Define

$$\gamma^+ = \max\left\{\mathbb{P}\left(E\left(\mathbf{c}\right)\right) : \mathbf{c} \in \mathcal{C}\left(x_1, \ldots, x_t, 1\right)\right\},$$
$$\gamma^- = \max\left\{\mathbb{P}\left(E\left(\mathbf{c}\right)\right) : \mathbf{c} \in \mathcal{C}\left(x_1, \ldots, x_t, 0\right)\right\}.$$

Under the importance sampling density, $X_{t+1} = 1$ with probability

$$\frac{\theta_{t+1}\gamma^+}{\theta_{t+1}\gamma^+ + (1 - \theta_{t+1})\,\gamma^-},$$

instead of θ_{t+1} under the original distribution. We add a without-replacement resampling step to this importance sampling algorithm by implementing Algorithm 2. We refer to this algorithm as WOR. As this algorithm is a fairly straightforward specialization of Algorithm 2 we do not describe the details of the algorithm here.

5.2.2 With Particle Merging

In order to apply Algorithm 3, we only need to specify the particle merging step. We do this by marking some of the missing edges in each unit as present, once

it has been determined that this change makes no difference to the connectivity properties of the graph.

An example of this situation is shown in Figure 7. In this case edge $\{3, 8\}$ is known to be missing, but vertices 3 and 8 are already known to be connected. So whether edge $\{3, 8\}$ is present or absent cannot change the connectivity properties of the final graph, regardless of the states of the remaining edges.

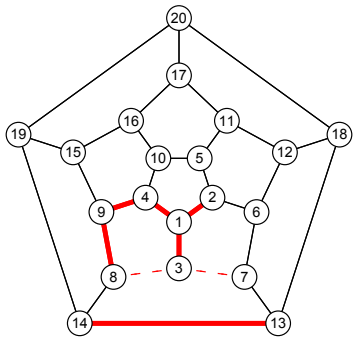


Figure 7: Example of the merging approach for network reliability. Thick edges are known to be present. Dashed edges are known to be absent. The states of all other edges are unknown.

Assume that we have some unit $(\mathbf{x}_t, w, p)$, and for some $1 < i < t$, $x_i = 0$. Let $\{v, v'\} = e_i$. Assume that v and v' are in the same connected component of $\mathcal{G}(x_1, \ldots, x_t)$, so that these vertices are already connected by a path that does not include edge e_i. Regardless of the states $x_{t+1}, \ldots, x_m$ of the remaining edges, setting $x_i = 1$ will never change whether the vertices in $\mathcal{K}$ to lie in the same connected component. So if

$$\mathbf{x}'_i = (x_1, \ldots, x_{i-1}, 1, x_{i+1}, \ldots, x_t),$$

it can be shown that $h^*(\mathbf{x}_t) = h^*(\mathbf{x}'_t)$. This observation leads to the particle merging step in Algorithm 4.

It is interesting to note that this algorithm is in some sense similar to the *turnip* (Lomonosov, 1994), which is a variation on *permutation Monte Carlo* (Elperin et al, 1991). In the case of the turnip, the states of some edges are ignored. In our case the merging step also tends to ignore the states of certain edge.

5.2.3 Results

We performed a simulation study to compare four different methods, all based on the importance sampling scheme of L'Ecuyer et al (2011). This importance sampling scheme by itself is method IS. Adding without-replacement sampling (Algorithm 2) is method WOR. Adding without-replacement sampling and particle merging (Algorithm 3) is method WOR-Merge. Adding the resampling

Algorithm 4: Merging step for network reliability example.

input : Set U of units of the form $(\mathbf{x}_t, w, p)$.
output: Set M of merged units

$W, M \leftarrow \emptyset$
for $(\mathbf{x}_t, w, p) \in U$ **do**
  for $i = 1$ **to** t **do**
    $\{v, v'\} \leftarrow e_i$
    if $x_i = 0$ *and* v, v' *are in the same component of* $\mathcal{G}(x_1, \ldots, x_t)$ **then**
      $x_i \leftarrow 1$ `// Modify entry` i `of` $\mathbf{x}_t$
  Add $(\mathbf{x}_t, w, p)$ to W `// Store modified values`
$W' \leftarrow \{\mathbf{x}_t \colon (\mathbf{x}_t, w, p) \in W\}$ `// Extract unique values of the first component`
for $\mathbf{x}_t \in W'$ **do**
  $w \leftarrow \sum_{(\mathbf{x}'_t, w', p') \in W, \mathbf{x}'_t = \mathbf{x}_t} w'$
  $p \leftarrow \sum_{(\mathbf{x}'_t, w', p') \in W, \mathbf{x}'_t = \mathbf{x}_t} p'$
  Add $(\mathbf{x}_t, w, p)$ to M

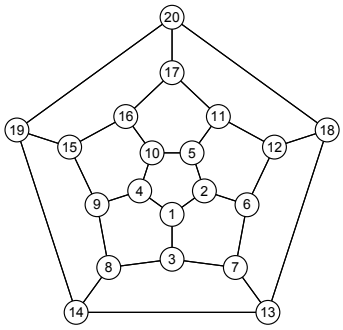
method of Fearnhead and Clifford (2003) is method Fearnhead. We used sample sizes $10, 20, 100$, 1000 and 10000.

We also implemented a residual resampling method (Carpenter et al, 1999). However, this method was found to perform uniformly worse than vanilla importance sampling on all the network reliability examples tested. The resampling step has the affect of "negating" the importance sampling scheme. The results for this method are not shown in the figures for this section, as they cannot reasonably be shown on the same scale.

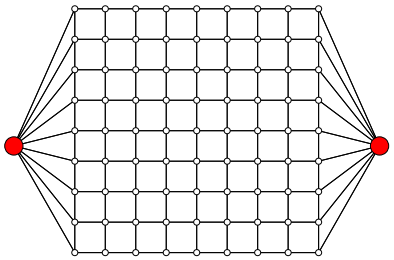
The first graph tested was the dodecahedron graph (Figure 8a), with $\mathcal{K} = \{1, 20\}$ and $\theta_i = 0.99$. Results are given in Figure 8c. In this case the true value of ℓ is known to be 2.061891×10^{-6}. All the without-replacement sampling methods have the property that the WNRV decreases as the sample size increases. Method WOR-Merge clearly outperforms the other methods. Application of a residual resampling algorithm to this problem resulted in an estimator with a work normalized variance on the order of 10^{-9}, many orders of magnitude worse than the results for the other four methods.

The second graph tested was a modification of the 9×9 grid graph (Figure 8b), where $\mathcal{K}$ contains the highlighted vertices. The modified grid graph is a somewhat pathological case for this importance sampling density, as in the limit as $p \to 1$ one of the 9 minimum cuts has a very low probability of being selected. Results in Figure 8d show that the WOR-Merge estimator significantly outperforms the other estimators.

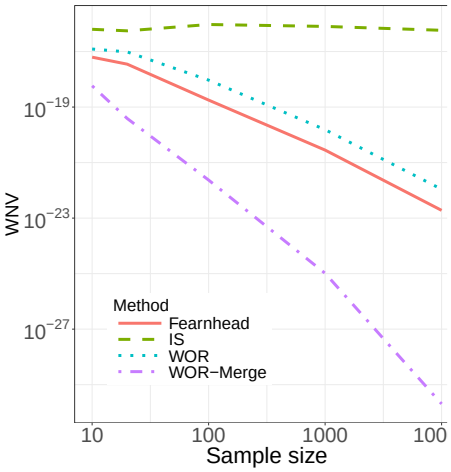
The third graph tested was three dodecahedron graphs arranged in parallel (Figure 9), with $\theta_i = 0.9999$. Simulation results are shown in Figure 10. It is interesting to see that the performance of method WOR-Merge does not change



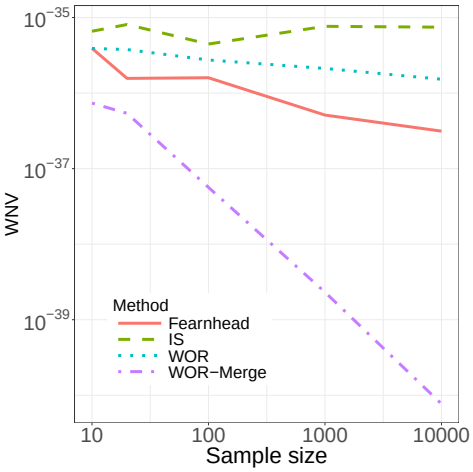
(a) Dodecahedron graph



(b) Modified 9×9 grid graph. The vertices in $\mathcal{K}$ are highlighted.



(c) Work normalized variance results for the dodecahedron graph, with edge reliability $\theta_i = 0.99$.



(d) Work normalized variance results for the modified 9×9 grid graph with edge reliability $\theta_i = 0.99$.

significantly when increasing the sample size from 20 to 100, or from 1000 to 10000.

The fourth graph tested was three dodecahedron graphs arranged in series (Figure 11), with $\theta_i = 0.9999$. Simulation results are given in Figure 12.

6 Concluding Remarks

This article has described the incorporation of ideas from sampling theory into sequential Monte Carlo methods. Taking a sampling theory approach provides a new perspective on the use of without-replacement sampling methods. It shows how the inclusion probabilities of the sampling designs take the place of the importance density in a standard importance resampling algorithm.

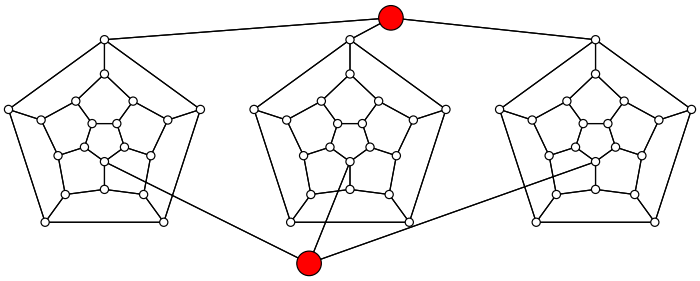


Figure 9: Three dodecahedron graphs arranged in parallel. The vertices in $\mathcal{K}$ are highlighted.

This article shows that the sampling method of Fearnhead and Clifford (2003) is systematic sampling, and that the optimality result of Fearnhead and Clifford (2003) relates to probability proportional to size sampling. The *stochastic enumeration algorithm* of Vaisman and Kroese (2015) is also a special case of the methods described in this paper. It uses simple random sampling without importance sampling, and introduces some merging ideas, which they term *tree reductions*.

Adding a resampling step to an importance sampling algorithm has the *potential* to increase the variance of the resulting estimator. If the importance sampling density is sufficiently different from the zero–variance density, adding a without-replacement resampling step can result in significant improvement. We illustrated this with reference to the $\mathcal{K}$-terminal network reliability problem, and a hidden Markov model.

In the case of the network reliability example, adding a without-replacement sampling step improved the variance of the importance sampling estimator proposed by L'Ecuyer et al (2011) by an order of magnitude. The without-replacement algorithms have the property that the work-normalized variance decreases as the sample size increases; the importance sampling algorithm on which they are based do not have this property. In our experience the importance sampling estimator was (previously) the best known estimation method for this problem.

We also applied a residual resampling method to the network reliability example, and found that its performance was an order of magnitude *worse* than the original importance sampling scheme. This is because in this case the resampling step tends to "negate" the importance sampling step. This highlights an important distinction between the without-replacement sampling methods we describe, and more traditional forms of resampling. In the methods we propose, the true density f does not enter into the resampling step. This works extremely well, where the importance density is well designed. The true density could be incorporated into the sampling step by changing the definition

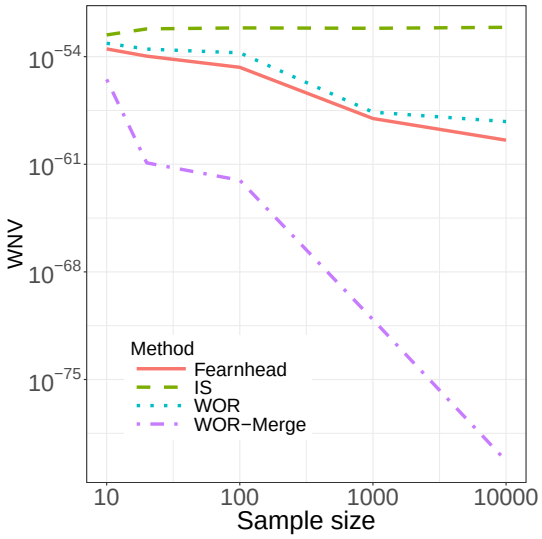


Figure 10: Work normalized variance results for three dodecahedron graphs in parallel, with edge reliability $\theta_i = 0.9999$.

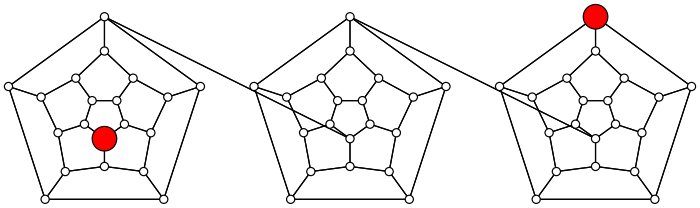


Figure 11: Three dodecahedron graphs arranged in series. The vertices in $\mathcal{K}$ are highlighted.

of the size variables.

In this article we suggested the use of a Pareto sampling design, where the inclusion probabilities are approximated by those of a Sampford design. This results in a design which is easy to simulate from, and which has inclusion probabilities which are easy to calculate. In this sense the proposed design is similar to systematic sampling. The performance of the Pareto approximation to the Sampford design is found to be similar to the systematic sampling design suggested by Fearnhead and Clifford (2003).

The approximation we suggest has the advantage that the joint inclusion probability of every pair of units is positive. This condition is known to be desirable in the sampling design literature, as it allows the estimation of the variance of the Horvitz–Thompson estimator. In future, this may allow the es-

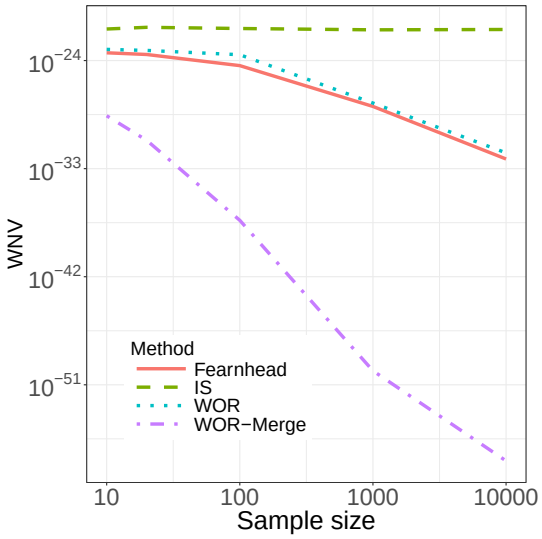


Figure 12: Work normalized variance results for three dodecahedron graphs in series, with $\theta_i = 0.9999$.

timation of the variance of the without-replacement sampling estimator, without the need to construct independent estimates.

Caution is potentially needed when applying an approximation within an iterative procedure, as it is possible that approximation errors will accumulate. Our numerical results suggest that such an accumulation of errors is not significant. Ultimately, such concerns must be balanced against the advantages of the proposed methods. Further work in the field of sampling design may make the use of approximations unnecessary.

When using without-replacement sampling, the merging of equivalent units can significantly reduce the variance of the resulting estimators. However, this also has disadvantages. When equivalent units are merged, it becomes even less certain that the variance of the resulting estimator always decreases as n increases. Merging also specializes the resulting algorithm to the function h; in general, it is not possible to use the final sample generated by the algorithm to estimate the expectation of a different function.

Particle merging is a way of incorporating problem-specific information into a particle filtering algorithm, in a way that is similar to the design of an importance sampling density. Proposition 4.1 is not the only way this can occur. Another possibility is that $h^*(\mathbf{x}_t) = m(\mathbf{x}_t)\, h^*(\mathbf{x}'_t)$, where $m(\mathbf{x}_t)$ is some known function, but both $h^*(\mathbf{x}_t)$ and $h^*(\mathbf{x}'_t)$ are unknown.

Our examples also illustrated the extreme flexibility of without-replacement sampling algorithms; the importance density, sampling design and merging steps can all be changed. In both our examples, this allowed us to use problem-specific information in the resulting algorithm. The downside is that customizing these

algorithms to this extent is non-trivial; it essentially requires the design of an entirely new algorithm every for every application.

The sampling theory approach is a promising direction for future work on without-replacement sampling methods; there is a large literature which probably contains many more relevant ideas. As an example, recall that in order to apply these types of methods, it must be possible to enumerate all the possible values of the particles at the next step. In some cases this set may so large that this is impractical. In the field of sampling theory, this is referred to as a case where the *sampling frame* (set of all possible units) is missing. These types of problems have been studied in the relevant literature, so solutions to this problem may already exist.

The sampling theory approach also provides some insight into the optimality result of Fearnhead and Clifford (2003). The optimality result is given for only a *single* sampling step. When multiple such resampling steps are performed, the variance of the resulting estimator will depend in a complicated way on the *joint* inclusion probabilities of the sampling designs which are applied. These joint inclusion probabilities do not enter into the optimality result. While the result of Fearnhead and Clifford (2003) is the strongest statement that can be made in the general case, there may be specific cases where sampling designs that *do not* satisfy the optimality condition result in a lower variance estimator.

Appendix

7 Unbiasedness of Sequential Without-Replacement Monte Carlo

Let $h^{*}\left(\mathbf{x}_{t}\right)=\mathbb{E}\left[h\left(\mathbf{X}_{d}\right)\mid\mathbf{X}_{t}=\mathbf{x}_{t}\right]$. Note that

$$\sum_{\mathbf{x}_{t}\in\mathscr{S}_{t}\left(\mathbf{x}_{t-1}\right)}h^{*}\left(\mathbf{x}_{t}\right)f\left(\mathbf{x}_{t}\right)=h^{*}\left(\mathbf{x}_{t-1}\right)f\left(\mathbf{x}_{t-1}\right).$$

Consider the expression

$$\sum_{\mathbf{x}_{t}\in\mathbf{S}_{t}}\frac{h^{*}\left(\mathbf{x}_{t}\right)f\left(\mathbf{x}_{t}\right)}{\prod_{i=1}^{t}\pi^{i}\left(\mathbf{x}_{t}\right)},\tag{18}$$

where $1\leq t<d$. Let $I\left(\mathbf{x}_{t}\right)$ be a binary variable, where $I\left(\mathbf{x}_{t}\right)=1$ indicates the inclusion of element $\mathbf{x}_{t}$ of $\mathscr{S}_{t}\left(\mathbf{S}_{t-1}\right)$ in $\mathbf{S}_{t}$. We can rewrite (18) as

$$\sum_{\mathbf{x}_{t}\in\mathscr{S}_{t}\left(\mathbf{S}_{t-1}\right)}I_{t}\left(\mathbf{x}_{t}\right)\frac{h^{*}\left(\mathbf{x}_{t}\right)f\left(\mathbf{x}_{t}\right)}{\prod_{i=1}^{t}\pi^{i}\left(\mathbf{x}_{t}\right)}.\tag{19}$$

Recall that $\mathbb{E}\left[I_t\left(\mathbf{x}_t\right) \mid \mathbf{S}_{t-1}\right] = \pi^t\left(\mathbf{x}_t\right)$. So the expectation of (19) conditional on $\mathbf{S}_1, \ldots, \mathbf{S}_{t-1}$ is

$$\begin{aligned}\sum_{\mathbf{x}_t \in \mathscr{S}_t(\mathbf{S}_{t-1})} \frac{h^*\left(\mathbf{x}_t\right) f\left(\mathbf{x}_t\right)}{\prod_{i=1}^{t-1} \pi^i\left(\mathbf{x}_t\right)} &= \sum_{\mathbf{x}_{t-1} \in \mathbf{S}_{t-1}} \frac{\sum_{\mathbf{x}_t \in \mathscr{S}_t(\mathbf{x}_{t-1})} h^*\left(\mathbf{x}_t\right) f\left(\mathbf{x}_t\right)}{\prod_{i=1}^{t-1} \pi^i\left(\mathbf{x}_{t-1}\right)} \\ &= \sum_{\mathbf{x}_{t-1} \in \mathbf{S}_{t-1}} \frac{h^*\left(\mathbf{x}_{t-1}\right) f\left(\mathbf{x}_{t-1}\right)}{\prod_{i=1}^{t-1} \pi^i\left(\mathbf{x}_{t-1}\right)}.\end{aligned}$$

So

$$\mathbb{E}\left[\sum_{\mathbf{x}_t \in \mathbf{S}_t} \frac{h^*\left(\mathbf{x}_t\right) f\left(\mathbf{x}_t\right)}{\prod_{i=1}^{t} \pi^i\left(\mathbf{x}_t\right)} \,\middle|\, \mathbf{S}_1, \ldots, \mathbf{S}_{t-1}\right] = \sum_{\mathbf{x}_{t-1} \in \mathbf{S}_{t-1}} \frac{h^*\left(\mathbf{x}_{t-1}\right) f\left(\mathbf{x}_{t-1}\right)}{\prod_{i=1}^{t-1} \pi^i\left(\mathbf{x}_{t-1}\right)} \tag{20}$$

Applying equation (20) d times to

$$\widehat{\ell} = \sum_{\mathbf{x}_d \in \mathbf{S}_d} \frac{h\left(\mathbf{X}_d\right) f\left(\mathbf{X}_d\right)}{\prod_{i=1}^{d-1} \pi^i\left(\mathbf{X}_d\right)} = \sum_{\mathbf{x}_d \in \mathbf{S}_d} \frac{h^*\left(\mathbf{X}_d\right) f\left(\mathbf{X}_d\right)}{\prod_{i=1}^{d-1} \pi^i\left(\mathbf{X}_d\right)}.$$

shows that $\mathbb{E}\left[\widehat{\ell}\right] = \ell$.

8 Unbiasedness of Sequential Without-Replacement Monte Carlo, with merging

The proof is similar to Appendix 7. In this case all the sample spaces and samples are sets of triples. Consider any expression of the form

$$\sum_{(\mathbf{x}_t, w, p) \in \mathscr{T}_t(\mathbf{S}_{t-1})} h^*\left(\mathbf{x}_t\right) w. \tag{21}$$

It is clear that if the proposed merging rule is applied to $\mathscr{T}_t\left(\mathbf{S}_{t-1}\right)$, then the value of (21) is unchanged. Using the definition of $\mathscr{T}_t\left(\mathbf{S}_{t-1}\right)$, equation (21) can be written as

$$\begin{aligned}&\sum_{(\mathbf{x}_{t-1}, w, p) \in \mathbf{S}_{t-1}} w \sum_{\mathbf{x}_t \in \mathscr{S}_t(\mathbf{x}_{t-1})} h^*\left(\mathbf{x}_t\right) \frac{f\left(x_t \mid \mathbf{x}_{t-1}\right)}{\pi^{t-1}\left(\mathbf{x}_{t-1}\right)} \\ &= \sum_{(\mathbf{x}_{t-1}, w, p) \in \mathbf{S}_{t-1}} \frac{\mathbb{E}\left[h^*\left(\mathbf{X}_t\right) \mid \mathbf{X}_{t-1} = \mathbf{x}_{t-1}\right] w}{\pi^{t-1}\left(\mathbf{x}_{t-1}\right)} \\ &= \sum_{(\mathbf{x}_{t-1}, w, p) \in \mathbf{S}_{t-1}} \frac{h^*\left(\mathbf{x}_{t-1}\right) w}{\pi^{t-1}\left(\mathbf{x}_{t-1}\right)}.\end{aligned} \tag{22}$$

The expectation of (22) conditional on $\mathbf{S}_{t-2}$ is

$$\sum_{(\mathbf{x}_{t-1}, w, p) \in \mathscr{T}_{t-1}(\mathbf{S}_{t-2})} h^*\left(\mathbf{x}_{t-1}\right) w. \tag{23}$$

So

$$\mathbb{E}\left[\sum_{(\mathbf{x}_t,w,p)\in\mathscr{T}_t(\mathbf{S}_{t-1})} h^*(\mathbf{x}_t)\, w \;\middle|\; \mathbf{S}_{t-2}\right] = \sum_{(\mathbf{x}_{t-1},w,p)\in\mathscr{T}_{t-1}(\mathbf{S}_{t-2})} h^*(\mathbf{x}_{t-1})\, w. \quad (24)$$

Applying equation (24) $d-1$ times to

$$\mathbb{E}\left[\widehat{\ell}\;\middle|\;\mathbf{S}_{d-1}\right] = \sum_{(\mathbf{x}_d,w,p)\in\mathscr{T}_d(\mathbf{S}_{d-1})} h^*(\mathbf{x}_d)\, w$$

shows that $\widehat{\ell}$ is unbiased.

9 Without-replacement sampling for the change point example

We now give the details of the application of without-replacement sampling to the change-point example in Section 9. Recall that $\mathbf{X}_d = \{X_t\}_{t=1}^d$ is a Markov chain and $\mathbf{Y}_d = \{Y_t\}_{t=1}^d$ are the observations. Let f be the joint density of $\mathbf{X}_d$ and $\mathbf{Y}_d$. Note that

$$f(\mathbf{x}_t \mid \mathbf{y}_t) = c_t f(\mathbf{x}_{t-1} \mid \mathbf{y}_{t-1}) f(x_t \mid \mathbf{x}_{t-1}) f(y_t \mid x_t), \quad (25)$$
$$f(\mathbf{x}_1 \mid \mathbf{y}_1) = c_1 f(x_1) f(y_1 \mid x_1), \quad (26)$$

for some unknown constants $\{c_t\}_{t=1}^d$. Define the size variables recursively as

$$p(\mathbf{x}_t) = p(\mathbf{x}_{t-1}) \frac{f(x_t \mid \mathbf{x}_{t-1}) f(y_t \mid x_t)}{\pi^{t-1}(\mathbf{x}_{t-1})}, \quad (27)$$
$$p(x_1) = f(x_1) f(y_1 \mid x_1). \quad (28)$$

This updating rule is slightly different from that given in (15). Equations (28) and (25) require an initial distribution for $X_1 = (C_1, O_1)$, which we take to be

$$\mathbb{P}(C_1 = 2, O_1 = 2) = \frac{1}{250}, \mathbb{P}(C_1 = 2, O_1 = 2) = \frac{249}{250}.$$

Define

$$\mathscr{U}_1 = \mathscr{U}_1(\emptyset) = \{(x_1, f(x_1) f(y_1 \mid x_1)) : x_1 \in \mathscr{S}_1\},$$

and let $\mathbf{S}_1$ be a sample chosen from $\mathscr{U}_1$, with probability proportional to the last component. Assume that sample $\mathbf{S}_{t-1}$ has been chosen, and let

$$\mathscr{U}_t(\mathbf{S}_{t-1}) = \left\{ \left(\mathbf{x}_t, w \frac{f(x_t \mid \mathbf{x}_{t-1}) f(y_t \mid x_t)}{\pi^{t-1}(\mathbf{x}_{t-1})} \right) : \right.$$
$$\left. (\mathbf{x}_{t-1}, w) \in \mathbf{S}_{t-1}, \mathbf{x}_t \in \text{Support}(\mathbf{X}_t \mid \mathbf{X}_{t-1} = \mathbf{x}_{t-1}) \right\}.$$

We account for the unknown normalizing constants in (25) by using an estimator of the form (10). This results in Algorithm 5.

Algorithm 5: Sequential Monte Carlo without replacement, for the change-point problem

input : Density f, function h, sampling designs
output: Estimate of $\mathbb{E}\left[h\left(\mathbf{X}_d\right) \mid \mathbf{Y}_d\right]$

$\mathbf{S}_0 \leftarrow \emptyset$
for $t = 1$ **to** d **do**
  $\mathbf{S}_t \leftarrow$ Sample from $\mathscr{U}_t\left(\mathbf{S}_{t-1}\right)$ according to some design, with size variables $\{w\colon\ (\mathbf{x}_t, w) \in \mathscr{U}_t\left(\mathbf{S}_{t-1}\right)\}$
  $\forall \mathbf{x}_t \in \mathbf{S}_t$ compute the inclusion probability $\pi^t\left(\mathbf{x}_t\right)$ of $\mathbf{x}_t$
return $\left(\sum_{(\mathbf{x}_d, w) \in \mathbf{S}_d} \frac{h(\mathbf{x}_d) w}{\pi^d(\mathbf{x}_d)}\right)\left(\sum_{(\mathbf{x}_d, w) \in \mathbf{S}_d} \frac{w}{\pi^d(\mathbf{x}_d)}\right)^{-1}$

Proposition 9.1. *The set $\mathbf{S}_d$ generated by Algorithm 5 has the property that*

$$\mathbb{E}\left[\sum_{(\mathbf{x}_d, w) \in \mathbf{S}_d} \frac{h\left(\mathbf{x}_d\right) w}{\pi^d\left(\mathbf{x}_d\right)}\right] = \mathbb{E}\left(h\left(\mathbf{X}_d\right) \mid \mathbf{Y}_d\right) \prod_{t=1}^{d} c_t^{-1}.$$

Proof. Define

$$H\left(\mathbf{x}_t\right) = \frac{\mathbb{E}\left[h\left(\mathbf{X}_d\right) \mid \mathbf{X}_t = \mathbf{x}_t, \mathbf{Y}_d = \mathbf{y}_d\right] f\left(\mathbf{x}_t \mid \mathbf{y}_d\right)}{f\left(\mathbf{x}_t \mid \mathbf{y}_t\right) \prod_{i=t+1}^{d} c_i}.$$

Using (25),

$$\begin{aligned}
&\sum_{\mathbf{x}_t \in \mathscr{S}_t\left(\mathbf{x}_{t-1}\right)} H\left(\mathbf{x}_t\right) f\left(x_t \mid \mathbf{x}_{t-1}\right) f\left(y_t \mid x_t\right) \\
&= \sum_{\mathbf{x}_t \in \mathscr{S}_t\left(\mathbf{x}_{t-1}\right)} \frac{\mathbb{E}\left[h\left(\mathbf{X}_d\right) \mid \mathbf{X}_t = \mathbf{x}_t, \mathbf{Y}_d = \mathbf{y}_d\right] f\left(\mathbf{x}_t \mid \mathbf{y}_d\right)}{f\left(\mathbf{x}_{t-1} \mid \mathbf{y}_{t-1}\right) \prod_{i=t}^{d} c_i} \\
&= \frac{\mathbb{E}\left[h\left(\mathbf{X}_d\right) \mid \mathbf{X}_{t-1} = \mathbf{x}_{t-1}, \mathbf{Y}_d = \mathbf{y}_d\right] f\left(\mathbf{x}_{t-1} \mid \mathbf{y}_d\right)}{f\left(\mathbf{x}_{t-1} \mid \mathbf{y}_{t-1}\right) \prod_{i=t}^{d} c_i} \\
&= H\left(\mathbf{x}_{t-1}\right).
\end{aligned}$$

Consider any expression of the form

$$\sum_{(\mathbf{x}_t, w) \in \mathscr{U}_t\left(\mathbf{S}_{t-1}\right)} H\left(\mathbf{x}_t\right) w. \tag{29}$$

Equation (29) can be written as

$$\begin{aligned}
&\sum_{(\mathbf{x}_{t-1}, w) \in \mathbf{S}_{t-1}} \sum_{\mathbf{x}_t \in \mathscr{S}_t\left(\mathbf{x}_{t-1}\right)} H\left(\mathbf{x}_t\right) w \frac{f\left(x_t \mid \mathbf{x}_{t-1}\right) f\left(y_t \mid x_t\right)}{\pi^{t-1}\left(\mathbf{x}_{t-1}\right)} \\
&= \sum_{(\mathbf{x}_{t-1}, w) \in \mathbf{S}_{t-1}} \frac{w H\left(\mathbf{x}_{t-1}\right)}{\pi^{t-1}\left(\mathbf{x}_{t-1}\right)}.
\end{aligned} \tag{30}$$

The expectation of (30) conditional on $\mathbf{S}_{t-2}$ is

$$\sum_{(\mathbf{x}_{t-1},w)\in\mathscr{U}_{t-1}(\mathbf{S}_{t-2})} H\left(\mathbf{x}_{t-1}\right) w.$$

So

$$\mathbb{E}\left[\sum_{(\mathbf{x}_t,w)\in\mathscr{U}_t(\mathbf{S}_{t-1})} H\left(\mathbf{x}_t\right) w \,\middle|\, \mathbf{S}_{t-2}\right] = \sum_{(\mathbf{x}_{t-1},w)\in\mathscr{U}_{t-1}(\mathbf{S}_{t-2})} H\left(\mathbf{x}_{t-1}\right) w. \tag{31}$$

Applying equation (31) $d-1$ times to

$$\begin{aligned}\mathbb{E}\left[\sum_{(\mathbf{x}_d,w)\in\mathbf{S}_d} \frac{h\left(\mathbf{x}_d\right) w}{\pi^d\left(\mathbf{x}_d\right)} \,\middle|\, \mathbf{S}_{d-1}\right] &= \sum_{(\mathbf{x}_d,w)\in\mathscr{U}_d(\mathbf{S}_{d-1})} h\left(\mathbf{x}_d\right) w \\ = \sum_{(\mathbf{x}_d,w)\in\mathscr{U}_d(\mathbf{S}_{d-1})} H\left(\mathbf{x}_d\right) w.\end{aligned}$$

completes the proof. ■ □

We now describe the merging step outlined in Fearnhead and Clifford (2003), applied to the estimation of the posterior change-point probabilities

$$\left\{\mathbb{P}\left(C_t = 2 \mid \mathbf{Y}_d = \mathbf{y}_d\right)\right\}_{t=1}^{d}.$$

The method we describe here can be extended fairly trivially to also estimate $\left\{\mathbb{P}\left(O_t = 2 \mid \mathbf{Y}_d = \mathbf{y}_d\right)\right\}_{t=1}^{d}$.

In order to perform this merging, we must add more information to all the sample spaces and the samples chosen from then. The extended space will have $\mathbf{x}_t$ as the first entry, the particle weight w as the second entry, and a vector $\mathbf{m}_t$ of t values as the third entry. The last entry will be an estimate of $\left\{\mathbb{P}\left(C_i = 2 \mid \mathbf{y}_t\right)\right\}_{i=1}^{t}$. Let

$$\mathscr{V}_1 = \left\{\left(x_1, f\left(x_1\right) f\left(y_1 \mid x_1\right), \mathbb{P}\left(C_1 = 2 \mid x_1\right)\right) : x_1 \in \mathscr{S}_1\right\}.$$

Note that the third component of every element of $\mathscr{V}_1$ is either 0 or 1. Let $\mathbf{S}_1$ be a sample drawn from $\mathscr{V}_1$, with probability proportional to the second element. Assume that sample $\mathbf{S}_{t-1}$ has been chosen, and let $\mathscr{V}_t\left(\mathbf{S}_{t-1}\right)$ be

$$\begin{aligned}&\left\{\left(\mathbf{x}_t, w\frac{f\left(x_t \mid \mathbf{x}_{t-1}\right) f\left(y_t \mid \mathbf{x}_t\right)}{\pi^{t-1}\left(\mathbf{x}_{t-1}\right)}, \left(\mathbf{m}_{t-1}, \mathbb{P}\left(C_t = 2 \mid \mathbf{X}_t = \mathbf{x}_t, \mathbf{Y}_d = \mathbf{y}_d\right)\right)\right) : \right. \\ &\left.\left(\mathbf{x}_{t-1}, w, \mathbf{m}_{t-1}\right) \in \mathbf{S}_{t-1}, \mathbf{x}_t \in \mathscr{S}_t\left(\mathbf{x}_{t-1}\right)\right\}.\end{aligned}$$

We can now define Algorithm 6, which uses the merging step outlined in Proposition 9.3.

Algorithm 6: Sequential Monte Carlo without replacement, for the change-point problem, for the marginal distributions of $\{C_t\}_{t=1}^d$.

input : Density f, function h, sampling designs
output: Estimate of $\{\mathbb{P}(C_t = 2 \mid \mathbf{y}_d)\}_{t=1}^d$.

$\mathbf{S}_0 \leftarrow \emptyset$
for $t = 1$ **to** d **do**
  $U \leftarrow \mathscr{V}_t(\mathbf{S}_{t-1})$
  Merge elements in U according to Proposition 9.3
  $\mathbf{S}_t \leftarrow$ Sample from $\mathscr{V}_t(\mathbf{S}_{t-1})$ according to some design, with size variables $\{w\colon (\mathbf{x}_t, w) \in \mathscr{V}_t(\mathbf{S}_{t-1})\}$
  $\forall \mathbf{x}_t \in \mathbf{S}_t$ compute the inclusion probability $\pi^t(\mathbf{x}_t)$ of $\mathbf{x}_t$
return $\left(\sum_{(\mathbf{x}_d, w, \mathbf{m}_d) \in \mathbf{S}_d} \frac{\mathbf{m}_d w}{\pi^d(\mathbf{x}_d)}\right)\left(\sum_{(\mathbf{x}_d, w, \mathbf{m}_d) \in \mathbf{S}_d} \frac{w}{\pi^d(\mathbf{x}_d)}\right)^{-1}$

Proposition 9.2. *If the merging step is omitted, then the set $\mathbf{S}_d$ generated by Algorithm 6 has the property that*

$$\mathbb{E}\left[\sum_{(\mathbf{x}_d, w, \mathbf{m}_d) \in \mathbf{S}_d} \frac{\mathbf{m}_d w}{\pi^d(\mathbf{x}_d)}\right] = \frac{\{\mathbb{P}(C_t = 2 \mid \mathbf{Y}_d = \mathbf{y}_d)\}_{t=1}^d}{\prod_{t=1}^d c_t}.$$

Proof. Define

$$G(\mathbf{x}_t, \mathbf{m}_t) = (\mathbf{m}_t, \mathbb{P}(C_{t+1} = 2 \mid \mathbf{X}_t = \mathbf{x}_t, \mathbf{Y}_d = \mathbf{y}_d), \ldots, \mathbb{P}(C_d = 2 \mid \mathbf{X}_t = \mathbf{x}_t, \mathbf{Y}_d = \mathbf{y}_d)) \times \frac{f(\mathbf{x}_t \mid \mathbf{y}_d)}{f(\mathbf{x}_t \mid \mathbf{y}_t) \prod_{i=t+1}^d c_i}.$$

It can be shown that

$$\sum_{\mathbf{x}_t \in \mathscr{S}_t(\mathbf{x}_{t-1})} G(\mathbf{x}_t, (\mathbf{m}_{t-1}, \mathbb{P}(C_t = 2 \mid \mathbf{X}_t = \mathbf{x}_t, \mathbf{Y}_d = \mathbf{y}_d))) f(x_t \mid \mathbf{x}_{t-1}) f(y_t \mid x_t) = G(\mathbf{x}_{t-1}, \mathbf{m}_{t-1}).$$

Consider any expression of the form

$$\sum_{(\mathbf{x}_t, w, \mathbf{m}_t) \in \mathscr{V}_t(\mathbf{S}_{t-1})} G(\mathbf{x}_t, \mathbf{m}_t) w. \tag{32}$$

Equation (32) can be written as

$$\sum_{(\mathbf{x}_{t-1}, w, \mathbf{m}_{t-1}) \in \mathbf{S}_{t-1}} w \sum_{\mathbf{x}_t \in \mathscr{S}_t(\mathbf{x}_{t-1})} G(\mathbf{x}_t, (\mathbf{m}_{t-1}, \mathbb{P}(C_t = 2 \mid \mathbf{X}_t = \mathbf{x}_t, \mathbf{Y}_d = \mathbf{y}_d))) \times \frac{f(x_t \mid \mathbf{x}_{t-1}) f(y_t \mid \mathbf{x}_t)}{\pi^{t-1}(\mathbf{x}_{t-1})}$$
$$= \sum_{(\mathbf{x}_{t-1}, w, \mathbf{m}_{t-1}) \in \mathbf{S}_{t-1}} w \frac{G(\mathbf{x}_{t-1}, \mathbf{m}_{t-1})}{\pi^{t-1}(\mathbf{x}_{t-1})}. \tag{33}$$

The expectation of (33) conditional on $\mathbf{S}_{t-2}$ is

$$\sum_{(\mathbf{x}_{t-1},w,\mathbf{m}_{t-1})\in\mathscr{V}_{t-1}(\mathbf{S}_{t-2})} wG\left(\mathbf{x}_{t-1},\mathbf{m}_{t-1}\right).$$

So

$$\begin{aligned}\mathbb{E}\left[\sum_{(\mathbf{x}_t,w,\mathbf{m}_t)\in\mathscr{V}_t(\mathbf{S}_{t-1})} G\left(\mathbf{x}_t,\mathbf{m}_t\right)w \;\middle|\; \mathbf{S}_{t-2}\right] = \\ = \sum_{(\mathbf{x}_{t-1},w,\mathbf{m}_{t-1})\in\mathscr{V}_{t-1}(\mathbf{S}_{t-2})} wG\left(\mathbf{x}_{t-1},\mathbf{m}_{t-1}\right). \qquad (34)\end{aligned}$$

Applying equation (34) $d-1$ times to

$$\mathbb{E}\left[\sum_{(\mathbf{x}_d,w,\mathbf{m}_d)\in\mathbf{S}_d} \frac{\mathbf{m}_d w}{\pi^d\left(\mathbf{x}_d\right)} \;\middle|\; \mathbf{S}_{d-1}\right] = \sum_{(\mathbf{x}_d,w,\mathbf{m}_d)\in\mathscr{V}_d(\mathbf{S}_{d-1})} wG\left(\mathbf{x}_d,\mathbf{m}_d\right)$$

completes the proof. ■ □

Proposition 9.3. *Assume we have two units* $(\mathbf{x}_t, w, \mathbf{m}_t)$ *and* $(\mathbf{x}'_t, w', \mathbf{m}'_t)$*, both corresponding to paths of the Markov chain with* $C_t = 2$ *and* $O_t = 2$*. Then we can remove these units, and replace them with the single unit*

$$\left(\mathbf{x}_t, w+w', \frac{w\mathbf{m}_t + w'\mathbf{m}'_t}{w+w'}\right).$$

This rule also applies if both units correspond to $C_t = 2$ *and* $O_t = 1$*.*

Proof. Under the specified conditions on $\mathbf{x}_t$ and $\mathbf{x}'_t$,

$$\begin{aligned}\mathbb{P}\left(C_i = 2 \mid \mathbf{X}_t = \mathbf{x}_t, \mathbf{Y}_d = \mathbf{Y}_d\right) = \mathbb{P}\left(C_i = 2 \mid \mathbf{X}_t = \mathbf{x}'_t, \mathbf{Y}_d = \mathbf{Y}_d\right), &\quad \forall t+1 \le i \le d, \\ f\left(\mathbf{x}_t \mid \mathbf{y}_t\right) = f\left(\mathbf{x}_t \mid \mathbf{y}_d\right), &\quad \forall t+1 \le i \le d, \\ f\left(\mathbf{x}'_t \mid \mathbf{y}_t\right) = f\left(\mathbf{x}'_t \mid \mathbf{y}_d\right), &\quad \forall t+1 \le i \le d.\end{aligned}$$

This shows that

$$\left(w+w'\right) G\left(\mathbf{x}_t, \frac{w\mathbf{m}_t + w'\mathbf{m}'_t}{w+w'}\right) = wG\left(\mathbf{x}_t,\mathbf{m}_t\right) + w'G\left(\mathbf{x}'_t,\mathbf{m}'_t\right).$$

So replacement of this pair of units by the specified single unit does not bias the resulting estimator. □

References

Aires N (2000) Comparisons between conditional Poisson sampling and Pareto πps sampling designs. Journal of Statistical Planning and Inference 88(1):133 – 147

Bondesson L, Traat I, Lundqvist A (2006) Pareto sampling versus Sampford and conditional Poisson sampling. Scandinavian Journal of Statistics 33(4):pp. 699–720

Brewer KRW, Hanif M (1983) Sampling with unequal probabilities, vol 15. Springer, New York

Brockwell A, Del Moral P, Doucet A (2010) Sequentially interacting markov chain monte carlo methods. Ann Statist 38(6):3387–3411

Carpenter J, Clifford P, Fearnhead P (1999) Improved particle filter for nonlinear problems. IEE Proceedings - Radar, Sonar and Navigation 146(1):2–7

Chen R, Liu JS (2000) Mixture kalman filters. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 62(3):493–508

Chen Y, Diaconis P, Holmes SP, Liu JS (2005) Sequential Monte Carlo methods for statistical analysis of tables. Journal of the American Statistical Association 100(469):109–120

Cochran WG (1977) Sampling techniques, 3rd edn. Wiley, New York

Del Moral P, Doucet A, Jasra A (2006) Sequential Monte Carlo samplers. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 68(3):411–436

Douc R, Cappé O, Moulines E (2005) Comparison of resampling schemes for particle filtering. In: ISPA 2005. Proceedings of the 4th International Symposium on Image and Signal Processing and Analysis, 2005., pp 64–69

Doucet A, de Freitas N, Gordon N (eds) (2001) Sequential Monte Carlo methods in practice. Statistics for Engineering and Information Science, Springer New York

Elperin TI, Gertsbakh I, Lomonosov M (1991) Estimation of network reliability using graph evolution models. IEEE Transactions on Reliability 40(5):572–581

Fearnhead P (1998) Sequential monte carlo methods in filter theory. PhD thesis, University of Oxford

Fearnhead P, Clifford P (2003) On-line inference for hidden markov models via particle filters. Journal of the Royal Statistical Society Series B (Statistical Methodology) 65(4):887–899

Gerber M, Chopin N (2015) Sequential quasi monte carlo. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 77(3):509–579

Gilks WR, Berzuini C (2001) Following a moving target–Monte Carlo inference for dynamic Bayesian models. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 63(1):127–146

Gordon N, Salmond D, Smith A (1993) Novel approach to nonlinear/non-gaussian bayesian state estimation. Radar and Signal Processing, IEE Proceedings F 140(2):107–113

Hammersley JM, Morton KW (1954) Poor man's monte carlo. Journal of the Royal Statistical Society Series B (Methodological) 16(1):pp. 23–38

Hartley HO, Rao JNK (1962) Sampling with unequal probabilities and without replacement. Ann Math Statist 33(2):350–374

Horvitz DG, Thompson DJ (1952) A generalization of sampling without replacement from a finite universe. Journal of the American Statistical Association 47(260):pp. 663–685

Iachan R (1982) Systematic sampling: A critical review. International Statistical Review / Revue Internationale de Statistique 50(3):293–303

Kong A, Liu JS, Wong WH (1994) Sequential imputations and bayesian missing data problems. Journal of the American statistical association 89(425):278–288

Kou SC, McCullagh P (2009) Approximating the α-permanent. Biometrika 96(3):635–644

L'Ecuyer P, Rubino G, Saggadi S, Tuffin B (2011) Approximate zero-variance importance sampling for static network reliability estimation. IEEE Transactions on Reliability 60(3):590–604

Liu JS (2001) Monte Carlo strategies in scientific computing. Springer, New York

Liu JS, Chen R (1995) Blind deconvolution via sequential imputations. Journal of the American Statistical Association 90(430):567–576

Liu JS, Chen R, Logvinenko T (2001) A theoretical framework for sequential importance sampling with resampling. In: Doucet A, de Freitas N, Gordon N (eds) Sequential Monte Carlo Methods in Practice, Statistics for Engineering and Information Science, Springer New York, pp 225–246

Lomonosov M (1994) On Monte Carlo estimates in network reliability. Probability in the Engineering and Informational Sciences 8:245–264

Madow WG (1949) On the theory of systematic sampling, ii. Ann Math Statist 20(3):333–354

Madow WG, Madow LH (1944) On the theory of systematic sampling, i. Ann Math Statist 15(1):1–24

Marshall A (1956) The use of multi-stage sampling schemes in monte carlo computations, ha meyer. In: Symposium on Monte Carlo Methods, Wiley, Hoboken, NJ

Ó Ruanaidh JJK, Fitzgerald WJ (1996) Numerical Bayesian Methods applied to Signal Processing. Springer, New York

Paige B, Wood F, Doucet A, Teh YW (2014) Asynchronous anytime sequential monte carlo. In: Ghahramani Z, Welling M, Cortes C, Lawrence N, Weinberger K (eds) Advances in Neural Information Processing Systems 27, Curran Associates, Inc., pp 3410–3418

Rosén B (1997a) Asymptotic theory for order sampling. Journal of Statistical Planning and Inference 62(2):135 – 158

Rosén B (1997b) On sampling with probability proportional to size. Journal of Statistical Planning and Inference 62(2):159 – 191

Rosenbluth MN, Rosenbluth AW (1955) Monte carlo calculation of the average extension of molecular chains. The Journal of Chemical Physics 23(2):356–359

Rubinstein RY, Kroese DP (2017) Simulation and the Monte Carlo Method, 3rd edn. John Wiley & Sons, New York

Sampford MR (1967) On sampling without replacement with unequal probabilities of selection. Biometrika 54(3-4):499–513

Tillé Y (2006) Sampling algorithms. Springer, New York

Vaisman R, Kroese DP (2015) Stochastic enumeration method for counting trees. Methodology and Computing in Applied Probability pp 1–43

Wall FT, Erpenbeck JJ (1959) New method for the statistical computation of polymer dimensions. The Journal of Chemical Physics 30(3):634–637