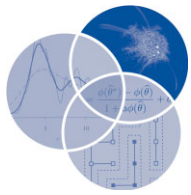
Advanced Review



Importance sampling: a review

Surya T. Tokdar[1] and Robert E. Kass[2]*

We provide a short overview of importance sampling—a popular sampling tool used for Monte Carlo computing. We discuss its mathematical foundation and properties that determine its accuracy in Monte Carlo approximations. We review the fundamental developments in designing efficient importance sampling (IS) for practical use. This includes parametric approximation with optimization-based adaptation, sequential sampling with dynamic adaptation through resampling and population-based approaches that make use of Markov chain sampling. *WIREs Comp Stat* 2010 2 54–60

Importance sampling (IS) refers to a collection of Monte Carlo methods where a mathematical expectation with respect to a *target* distribution is approximated by a weighted average of random draws from another distribution. Together with Markov Chain Monte Carlo methods, IS has provided a foundation for simulation-based approaches to numerical integration since its introduction as a variance reduction technique in statistical physics.[1,2] Nowadays, IS is used in a wide variety of application areas and there have been recent developments involving *adaptive* versions of the methodology.

The appeal of IS lies in a simple probability result. Let $p(x)$ be a probability density for a random variable X and suppose we wish to compute an expectation $\mu_f = \mathbb{E}_p[f(X)]$, with

$$\mu_f = \int f(x)p(x)dx. \tag{1}$$

Then for any probability density $q(x)$ that satisfies $q(x) > 0$ whenever $f(x)p(x) \neq 0$, one has

$$\mu_f = \mathbb{E}_q[w(X)f(X)] \tag{2}$$

where $w(x) = \frac{p(x)}{q(x)}$ and now $\mathbb{E}_q[\cdot]$ denotes the expectation with respect to $q(x)$. Therefore, a sample of independent draws $x^{(1)}, \cdots, x^{(m)}$ from $q(x)$ can be used to estimate μ_f by

$$\hat{\mu}_f = \frac{1}{m}\sum_{j=1}^{m} w(x^{(j)})f(x^{(j)}). \tag{3}$$

In many applications the density $p(x)$ is known only up to a normalizing constant. Here one has $w(x) = cw_0(x)$ where $w_0(x)$ can be computed exactly but the multiplicative constant c is unknown. In this case one replaces $\hat{\mu}_f$ with the ratio estimate

$$\tilde{\mu}_f = \frac{\sum_{j=1}^{m} w(x^{(j)})f(x^{(j)})}{\sum_{j=1}^{m} w(x^{(j)})}. \tag{4}$$

It follows from the strong law of large numbers that $\hat{\mu}_f \to \mu$ and $\tilde{\mu}_f \to \mu_f$ as $n \to \infty$ almost surely; see Geweke.[3] Moreover a central limit theorem yields that $\sqrt{m}(\hat{\mu}_f - \mu_f)$ and $\sqrt{m}(\tilde{\mu}_f - \mu_f)$ are asymptotically normal with mean zero and respective variances $\mathbb{E}_q[(w(X)f(X) - \mu_f)^2]$ and $\mathbb{E}_q[w(X)^2(f(X) - \mu_f)^2]$—whenever these quantities are finite. These asymptotic variances can be consistently estimated by re-using the sampled $x^{(j)}$ values as $\frac{1}{m}\sum_j[w(x^{(j)})f(x^{(j)}) - \hat{\mu}_f]^2$ and $\sum_j[w(x^{(j)})^2(f(x^{(j)}) - \tilde{\mu}_f)^2]/[\sum_j w(x^{(j)})]^2$, respectively.

The approximation accuracy offered by IS depends critically on the choice of the trial density $q(x)$. Suppose $f(x) = 1$ for all x, and consequently $\mu_f = 1$, but we still want to estimate this by using IS with a trial density $q(x)$. In this case the variance of $\hat{\mu}_f$ is

$$V_p(\hat{\mu}_f) = \mathbb{E}_q[(w(X) - 1)^2]/m. \tag{5}$$

*Correspondence to: kass@stat.cmu.edu

[1]Department of Statistical Science, Duke University, Durham, NC 27708, USA

[2]Department of Statistics and Center for the Neural Basics of Cognition, Carnegie Mellon University, Pittsburgh, PA 15213, USA

DOI: 10.1002/wics.56

For IS to be accurate (with a limited number m of draws) this variance must be small, which requires $q(x)$ be approximately proportional to $p(x)$ for most x.

In a general Monte Carlo problem, where very little is known about the structural properties of the target density $p(x)$, it could be challenging to identify a $q(x)$ that is easy to sample from and yet provides a good approximation to $p(x)$. Usually, this problem intensifies with the dimension of x, as the relative volume of x where $p(x)$ is high becomes extremely small. There are, however, special cases where a reasonable choice of $q(x)$, or a class of such distributions, present itself. This article provides an overview of these cases and the related IS algorithms.

In many applications, theoretical properties of $p(x)$ are used to determine approximation within a family of $q(x)$ indexed by a low-dimensional vector-valued parameter. A final choice of $q(x)$ from the chosen family is made by numerically optimizing some prespecified measure of efficiency. Evaluating this measure can itself require a pilot IS or a recursive scheme of IS—details are discussed in Section 'Adaptive Parametric Importance Sampling'.

When x is high dimensional and possibly non-Euclidean, a parametric approximation to $p(x)$ is hard to obtain. In such cases one strategy is to break the task of approximating $p(x)$ into a series of low dimensional approximations. In many interesting Monte Carlo problems, $p(x)$ leads to a natural chain-like decomposition of x allowing a sequential construction of $q(x)$ that takes advantage of this decomposition. The resulting IS is discussed in Section 'Sequential Importance Sampling' (SIS). In the absence of a natural decomposition, it is still possible to apply the SIS framework by extending the Monte Carlo problem to an augmented space. A specific implementation of this strategy is presented in Section 'Annealed Importance Sampling'.

In Section 'SIS with Resampling', we review the use of resampling in SIS to adapt dynamically from an initial candidate $q(x)$ to the target $p(x)$ without requiring any numerical optimizations. This adaptability of SIS, which takes full advantage of its parallel computing structure, gives it a competitive edge against Monte Carlo methods that rely solely on Markov chain sampling (MCS). In Section 'SIS and Markov Chain Sampling', we end with a brief discussion of the current developments in IS research, especially its combination with MCS.

ADAPTIVE PARAMETRIC IMPORTANCE SAMPLING

In Bayesian statistics and econometric applications, $p(x)$ often represents an un-normalized posterior density $p(x) = cg(x)$ over a Euclidean parameter space, known only up to a multiplicative constant $c > 0$. In many cases, such a $p(x)$ can be asymptotically well approximated by a multivariate normal distribution with mean given by the posterior mode $\hat{x} = \text{argmin}_x[-\log p(x)]$ and variance matrix given by the inverse of the Hessian of $-\log p(x)$ at $\hat{x}$ (see Section 4 of Ghosh *et al.*[4] and the references therein). It is rather cheap to obtain stable and accurate approximation to these quantities through standard optimization routines. Hence the corresponding multivariate normal density serves as a good candidate for $q(x)$. In practice, a multivariate Student density with similar characteristics may be preferred to the multivariate normal choice.[3,5] This is because a multivariate Student density, with its heavy tails, provides a higher assurance of the finiteness $\mathbb{E}_q[w(X)^2]$ and thus that of the variance of $\tilde{\mu}_f$.

Oh and Berger[6] and previously Kloek and van Dijk[7] extended the above approach by reducing its dependence on the exact asymptotic approximation of $p(x)$. They took $q(x)$ to be given by the density $q_\lambda(x) = t_\nu(x \mid \lambda)$—the multivariate Student density with a fixed degrees of freedom ν and a location-scale $\lambda = (\mu, \Sigma)$ chosen to minimize

$$cv^2(\lambda) = \int \frac{p(x)^2}{t_\nu(x \mid \lambda)} dx - 1. \tag{6}$$

Note that $cv^2(\lambda)$ equals $\mathbb{E}_q[(w(X) - 1)^2]$ and hence, as noted earlier, a small magnitude of this quantity ensures high accuracy in estimating μ_f for all $f(x)$ that are relatively flat with respect to $p(x)$. Many authors take the related quantity $\mathbb{E}_q[w(X)^2]$ as a *rule-of-thumb* measure of efficiency of IS based on a trial density $q(x)$; see Liu[8] for a discussion. In the study by Oh and Berger[9], the candidate set for $q(x)$ was further extended to a finite mixture of the form $q_\lambda(x) = \sum_{i=1}^k \pi_i t_\nu(x \mid \mu_i, \Sigma_i)$ with $\lambda = \{(\pi_i, \mu_i, \Sigma_i)\}$, to cover the case of multimodal posterior densities. As the quantity in Eq. (6) cannot be computed in closed form or minimized analytically, Oh and Berger[9] suggested the following approximate optimization. Start with an initial guess λ^{init} for λ and sample $x^{(1)}, \cdots, x^{(m)}$ from $q_{\lambda^{\text{init}}}(x)$. Compute $\lambda^{\text{opt}} = \text{argmin}_\lambda \widehat{cv}^2(\lambda; \lambda^{\text{init}})$ where

$$\widehat{cv}^2(\lambda; \lambda') = \frac{\frac{1}{m}\sum_{j=1}^m \frac{[g(x^{(j)})/q_{\lambda'}(x^{(j)})]^2}{[q_\lambda(x^{(j)})|)/q_{\lambda'}(x^{(j)})]}}{\left[\frac{1}{m}\sum_{j=1}^m g(x^{(j)})/q_{\lambda'}(x^{(j)})\right]^2} - 1 \tag{7}$$

is an IS approximation to $cv^2(\lambda)$ based on the sample drawn from $q_{\lambda^{\text{init}}}(x)$.

A variation of the above idea was proposed by Richard and Zhang.[10] For a family of candidates $q_\lambda(x)$, they suggested choosing $\lambda = \lambda^{\text{opt}}$ where $(\alpha^{\text{opt}}, \lambda^{\text{opt}})$ minimizes the pseudo divergence

$$d(\alpha, \lambda) = \int (\log g(x) - \alpha - \log q_\lambda(x))^2 p(x) dx \quad (8)$$

over (α, λ). Note that if $p(x) = q_{\lambda_0}(x)$ for some λ_0, then the above is minimized at $(-\log c, \lambda_0)$. As in Oh and Berger,[9] Eq. (8) too has to be solved numerically. Richard and Zhang[10] proposed the following iterative scheme for this. Start with an initial estimate of $\lambda = \lambda^{(0)}$. For $t = 1, 2, \cdots$ compute

$$(\alpha^{(t+1)}, \lambda^{(t+1)}) = \text{argmin}_{\alpha,\lambda} \sum_{j=1}^{m} (\log g(x_t^{(j)}) - \alpha - \log q_\lambda(x_t^{(j)}))^2 \frac{g(x_t^{(j)})}{q_{\lambda^{(t)}}(x_t^{(j)})} \quad (9)$$

where $x_t^{(j)}$ are drawn from $q_{\lambda^{(t)}}(x)$. The attractive feature of this program is that the minimization in Eq. (9) is a generalized, weighted least squares minimization problem for which global solutions are often easy to find. In particular, if q_λ is chosen from an exponential family, then the above reduces to a simple least squares problem.

A different adaptive parametric IS was proposed by Owen and Zhou[11] who combined IS with the method of control variates.[12] They worked with a single choice of $q(x)$ but adapted their Monte Carlo method by optimizing over a parametric choice of the control variates. In particular, they took $q(x) = \sum_{i=1}^{k} \alpha_i q_i(x)$, with fixed densities q_i's and a fixed probability vector $\alpha = (\alpha_1, \cdots, \alpha_k)$, but proposed to estimate μ_f by

$$\hat{\mu}_{f,\beta} = \frac{1}{m} \sum_{j=1}^{m} \frac{f(x^{(j)})p(x^{(j)}) - \sum_{i=1}^{k} \beta_i q_i(x^{(j)})}{q(x^{(j)})} + \sum_{i=1}^{k} \beta_i \quad (10)$$

with $\beta = (\beta_1, \cdots, \beta_k)$ minimizing the asymptotic variance of $\hat{\mu}_{f,\beta}$ given by

$$\sigma^2(\beta) = \int \left(\frac{f(x)p(x) - \sum_i \beta_i q_i(x)}{q(x)} - \mu_f + \sum_i \beta_i \right)^2 q(x) dx. \quad (11)$$

A consistent estimate of an optimal β can be found by minimizing

$$\hat{\sigma}^2(\beta, \beta_0) = \sum_{j=1}^{m} \left(\frac{f(x^{(j)})p(x^{(j)})}{q(x^{(j)})} - \sum_i \beta_i \frac{q_i(x^{(j)})}{q(x)} - \beta_0 \right)^2 \quad (12)$$

through least squares methods. Note that $\hat{\mu}_{f,\beta}$ requires exact knowledge of $p(x)$. When $p(x) = cg(x)$ with c unknown, one can modify the estimate to become

$$\tilde{\mu}_{f,\beta} = \frac{\sum_{j=1}^{m} \frac{f(x^{(j)})g(x^{(j)}) - \sum_{i=1}^{k} \beta_i q_i(x^{(j)})}{q(x^{(j)})} + \sum_{i=1}^{k} \beta_i}{\sum_{j=1}^{m} \frac{g(x^{(j)}) - \sum_{i=1}^{k} \beta_i q_i(x^{(j)})}{q(x^{(j)})} + \sum_{i=1}^{k} \beta_i}. \quad (13)$$

It is also possible to use two different sets of β in the numerator and the denominator above. This approach is particularly attractive when more than one $(p(x), f(x))$ are of interest, and at least one of the chosen $q_i(x)$ is expected to lead to an efficient IS for each pair (see Theorem 2 in Ref 11).

SEQUENTIAL IMPORTANCE SAMPLING

In many Monte Carlo problems with a high-dimensional x, the target density $p(x)$ induces a chain-like decomposition of $x = (x_1, \cdots, x_d)$, paving the way for generating x sequentially as $x_{[1:t]} = (x_1, \cdots, x_t)$, $1 \le t \le d$. Such decompositions occur naturally in state-space models (finance, signal-tracking), evolutionary models (molecular physics and biology, genetics) and others (see Section 3 of Liu[8]). Writing $p(x)$ as

$$p(x) = p(x_1) \prod_{t=2}^{d} p(x_t \mid x_{[1:t-1]}) \quad (14)$$

it is easy to see that an efficient IS can be built by using a $q(x)$ of the form

$$q(x) = q_1(x_1) \prod_{t=2}^{d} q_t(x_t \mid x_{[1:t-1]}), \quad (15)$$

where $q_t(x_t \mid x_{[1:t-1]})$ mimics $p(x_t \mid x_{[1:t-1]})$ well. For such a scheme, the importance weight $w(x) = p(x)/q(x)$, too, can be computed sequentially as $w(x) = w_d$ where

$$w_t = w_{t-1} \frac{p(x_t \mid x_{[1:t-1]})}{q_t(x_t \mid x_{[1:t-1]})} \quad (16)$$

and $w_0 = 1$. The sequence w_t can be used to check on the fly the importance of the sample being generated, and one can possibly discard a sample half-way if w_t starts getting very small. We shall make this idea more precise in the next section.

To facilitate the construction of $q_t(x_t \mid x_{[1:t-1]})$, Liu[12] presented the above sequential importance sampling (SIS) scheme in a slightly more general form. Liu introduced a sequence of auxiliary distributions $p_t(x_{[1:t]})$, $1 \le t \le d$, with $p_d(x_{[1:d]}) = p(x)$ and rewrote the updating Eq. (16) as

$$w_t = w_{t-1} \frac{p_t(x_{[1:t]})}{p_{t-1}(x_{[1:t-1]}) q_t(x_t \mid x_{[1:t-1]})}. \qquad (17)$$

The auxiliary densities $p_t(x_{[1:t]})$ could be chosen to approximate the marginal densities $p(x_{[1:t]})$ with $p_t(x_t \mid x_{[1:t-1]})$ serving as a guideline for constructing $q_t(x_t \mid x_{[1:t-1]})$. This general definition accommodates the possibility that there could be various ways of obtaining a good approximation. We shall illustrate this with two examples of historical and practical relevance.

Consider the task of simulating a length-d *self-avoiding walk* (SAW, see Liu[8]) on the two-dimensional integer lattice starting from $(0, 0)$. Here $x = (x_1, \cdots, x_d)$ denotes a chain of d integer coordinates $x_t = (i_t, j_t)$, $1 \le t \le d$ such that

$$(i_t, j_t) \in \{(i_{t-1} - 1, j_{t-1}), (i_{t-1} + 1, j_{t-1}), (i_{t-1}, j_{t-1} - 1), (i_{t-1}, j_{t-1} + 1)\} \qquad (18)$$

with $x_t \ne (0, 0)$ and $x_t \ne x_s$ for any $1 \le s \ne t \le d$. Suppose the target distribution $p(x)$ is the uniform distribution over all length-d SAWs. A reasonable choice of an auxiliary $p_t(x_{[1:t]})$ in this case is the uniform distribution on $x_{[1:t]}$. Bear in mind that $p_t(x_{[1:t]}) \ne p(x_{[1:t]})$. By taking $q_t(x_t \mid x_{[1:t-1]}) = p_t(x_t \mid x_{[1:t-1]})$ it is easy to see that given the first $t - 1$ coordinates $x_{[1:t-1]}$, one samples x_t uniformly from the unoccupied neighbors of x_{t-1}. Alternatively one can take $p_t(x_{[1:t]})$ as the marginal distribution of $x_{[1:t]}$ under a uniform distribution on $x_{[1:t+1]}$. In this case $q_t = p_t(x_t \mid x_{[1:t-1]})$ leads to a two-step look ahead sampling of x_t given $x_{[1:t-1]}$ where a neighbor of x_{t-1} is selected with probability proportional to the number of unoccupied neighbors it currently has.

In statistical missing data problems, for example, the observables z_t, $1 \le t \le n$, are partitioned into $z_t = (x_t, y_t)$ with only the y_t components being actually observed. For maximum likelihood or Bayesian inference on such problems $p(x)$ often represents the conditional distribution $f(x \mid y)$ derived from a joint model $f(x, y)$ on the complete data z. When $f(x, y)$ specifies independence or a simple chain structure across z_t's, a useful choice of $p_t(x_{[1:t]})$ is given by $f(x_{[1:t]} \mid y_{[1:t]})$ with $q_t(x_t \mid x_{[1:t-1]}) = p_t(x_t \mid x_{[1:t-1]}) = f(x_t \mid z_{[1:t-1]}, y_t)$. The corresponding updates of w_t can be written more compactly as $w_t = w_{t-1} f(y_t \mid z_{1:t-1})$. Because the method fills in the missing components sequentially it is called *sequential imputation*.

ANNEALED IMPORTANCE SAMPLING

The appealing feature of SIS is that it achieves an approximation to $p(x)$ through a series of simpler approximations of $p(x_t \mid x_{[1:t-1]})$ by $q_t(x_t \mid x_{[1:t-1]})$. In *annealed importance sampling* (AIS), Neal[13] introduced a similar construction to handle cases where x does not admit a natural chain-like decomposition. Like SIS, a sequence of distributions $p_t(x)$, $0 \le t \le d$ is used, with $p_d(x) = p(x)$. But each of these densities is defined on the same space on which $p(x)$ is defined. Here the sequence $\{p_t(x)\}_t$ forms a bridge of successive approximations from $p_0(x)$ to $p(x) = p_d(x)$. The initial density $p_0(x)$ is taken to be diffuse and easy to sample from. It is required that at every step, sampling from $p_{t-1}(x)$ leads to an efficient IS for the immediate target $p_t(x)$. This can be achieved for some $p(x)$ when one defines $p_t(x) = p_0(x)^{1-b_t} p(x)^{b_t}$ with $0 = b_0 < b_1 < \cdots < b_d = 1$. This gradual morphing of a diffuse $p_0(x)$ to a possibly well concentrated $p(x)$ is reminiscent of the cooling schedules applied in *simulated annealing* (SANN) for function optimization. In fact, Neal[13] introduced AIS as an IS-augmented version of SANN fit for Monte Carlo approximations.

In AIS, a random draw of x is made by sequentially drawing $x_{(t)}$, $0 \le t \le d$, and equating $x = x_{(d)}$ as follows. One starts by drawing $x_{(0)}$ from $p_0(x)$ and sets $w_0 = 1$. Then, for $t = 1, \cdots, d$

1. Sample $x_{(t)}$ from $g_t(\cdot \mid x_{(t-1)})$.
2. Set $w_t = w_{t-1} \frac{p_t(x_{(t-1)})}{p_{t-1}(x_{(t-1)})}$

where $g_t(x' \mid x)$ is a transition kernel that leaves $p_t(x)$ invariant:

$$g_t(x' \mid x) \ge 0, \int g_t(x' \mid x) dx' = 1, \quad \int p_t(x) g_t(x' \mid x) dx = p_t(x'). \qquad (19)$$

Taking $\tilde{g}_t(x' \mid x) = g_t(x \mid x') p_t(x') / p_t(x)$ – the reversal of g_t – it can be shown that w_d gives the proper

importance weight $p^*(x^*)/q^*(x^*)$ for the target density

$$p^*(x^*) = p_d(x_{(d)}) \times \tilde{g}_d(x_{(d-1)} \mid x_{(d)}) \times \cdots \times \tilde{g}_1(x_{(0)} \mid x_{(1)}) \quad (20)$$

on the augmented variable $x^* = (x_{(0)}, x_{(1)}, \cdots, x_{(d)})$ with

$$q^*(x^*) = p_0(x_{(0)}) \times g_1(x_{(1)} \mid x_{(0)}) \times \cdots \times g_d(x_{(d)} \mid x_{(d-1)}) \quad (21)$$

as determined by the AIS sampling scheme. The marginal distribution of $x_{(d)}$ determined by $p^*(x^*)$ is simply $p_d(x_{(d)}) = p(x_{(d)})$.

Note that $g_t(x' \mid x)$ is left completely unspecified beyond the requisite invariance property (Eq. (19)). One can tap into the vast literature of Markov Chain Monte Carlo to construct a suitable transition kernel $g_t(x' \mid x)$. A simple choice is a few Metopolis or Gibbs updates of x with $p_t(x)$ as the target density. AIS also offers complete flexibility in the choice of the number of steps d and the intermediate densities $p_d(x)$. This choice can have a major impact on the performance of the algorithm; see Lyman and Zuckerman[14] and Godsill and Clapp[15].

SIS WITH RESAMPLING

In all of the SIS implementations detailed above, the m samples $x^{(1)}, \cdots, x^{(m)}$ are drawn in a noninteractive parallel manner. In such schemes, most of the corresponding weights $w^{(j)} = w(x^{(j)})$ are very small and contribute only a little to the computation of $\hat{\mu}_f$ or $\tilde{\mu}_f$. This becomes particularly problematic when x is high dimensional. For example, in SIS, it can be shown that the weight sequence w_t forms a martingale and hence its coefficient of variation $cv_t^2 = \mathbb{V}(w_t)/\mathbb{E}(w_t)^2$ explodes to infinity as t increases.[16] Consequently, a very small proportion of the final draws $x^{(j)}$ carry most of the weight—making the SIS estimate rather inefficient.

A simple fix of this, known as the enrichment method, was proposed by Wall and Erpenbeck.[17] Their idea was to grow all the $x_{[1:t]}^{(j)}$, $j \le 1 \le m$—each called a stream—simultaneously, and at intermediate check points $1 \le t_1 \le \cdots \le t_k \le d$, replace the streams with small current weights $w_t^{(j)}$ with replicates of the streams with large current weights. A simple re-weighting of the resulting streams makes the whole process a valid IS scheme. Grassberger[18] further suggested making the check points dynamic. In his *pruned-enriched Rosenbluth method* (PERM) each current stream is either removed or replicated (or grown according to the original SIS scheme) based on its current weight being smaller than a lower cutoff c_t or larger than an upper cutoff C_t (or otherwise). Note that the total number of streams may not remain the same.

A related idea of replicating the 'good' streams was explored by Gordon *et al.*[19] for the special IS method known as the *bootstrap filter* (also *particle filter*) for nonlinear/non-Gaussian state-space models. The setting here is similar to the missing data problem discussed in the previous section with the following important differences: (1) the model on $f(y, x)$ is assumed to have the following Markov structure

$$f(y, x) = \prod_{t=1}^{n} [f_{\text{state}}(x_t \mid x_{t-1}) f_{\text{obs}}(y_t \mid x_t)] \quad (22)$$

and (2) $f(x_t \mid x_{t-1}, y_t)$ is not assumed to be easy to sample from, making the choice $q_t(x_t \mid x_{[1:t-1]}) = f(x_t \mid z_{[1:t-1]}, y_t)$ infeasible. The choice of $q_t(x_t \mid x_{t-1}) = f_{\text{state}}(x_t \mid x_{t-1})$ is assumed feasible, but it often leads to an extremely large cv_t^2. Gordon *et al.*[19] improved upon this through an extra resampling stage as follows. At stage t, draw $x_t^{(*j)}$, $1 \le j \le m$ from $f_{\text{state}}(x_t \mid x_{t-1}^{(j)})$ and weight each draw by $w_t^{(j)} \propto f_{\text{obs}}(y_t \mid x_t^{(*j)})$. Resample from $\{x_t^{(*1)}, \cdots, x_t^{(*m)}\}$ with weights $w_t^{(j)}$ to produce the next stage draws $\{x_t^{(1)}, \cdots, x_t^{(m)}\}$. At completion, each stream gets weighed equally ($w_d^{(j)} = 1/m$) for evaluating the estimate $\tilde{\mu}_f$. This estimate is guaranteed to converge to μ_f as the number of streams (also known as *particles* in state-space literature) tends to infinity.

Liu and Chen[20] introduced sequential importance sampling with resampling (SISR) by combining across-stream resampling with a dynamic choice of check points. In SISR, a decision of resampling at state t is made by checking the current coefficient of variation cv_t^2 of the weights $\{w_t^{(j)}, \cdots, w_t^{(j)}\}$ against a prespecified cutoff c_t (typically growing with t at a polynomial rate). If cv_t^2 exceeds c_t, then $x_{[1:t]}^{(j)}$'s are updated by resampling from $\{x_{[1:t]}^{(1)}, \cdots, x_{[1:t]}^{(m)}\}$ with probability proportional to $w_t^{(j)}$. Each resampled stream is then assigned a weight $\sum_j w_t^{(j)}/m$. A check on the coefficient of variation guards against unwanted pruning when all streams have similar weights. Chen *et al.*[21] further modified SISR to allow resampling along a different time measurement than the original stage index t. The modified algorithm, called sequential importance sampling with stopping time resampling (SISSTR)

determines check points by applying a stopping rule on each stream separately. Once all streams have reached their first stop, they are pooled together and a resampling is done if the coefficient of variation of the current weights exceeds a prespecified cutoff. The streams then grow in parallel until the next stop is reached by each, and so on. An interesting application of this was presented by Chen *et al.*[21] to the coalescent model of Kingman[22] where SISSTR remarkably improved a naive SIS due to Griffiths and Tavaré.[23]

SIS AND MARKOV CHAIN SAMPLING

The introduction of resampling and the use of transition kernels have largely expanded the scope of SIS algorithms. The latter development has brought these algorithms closer to Monte Calro methods that use Markov chain sampling to generate a sequence of dependent draws from a target density by exploring it through appropriate transition kernels. A consensus is rapidly emerging that much can be gained by combining these two approaches together. The AIS algorithm, in which MCS facilitates SIS on an artificially augmented space, gave the first formal exploration of such a combination. The *population Monte Carlo* (PMC) algorithm of Cappé et al.[24]—inspired by a resampling enriched AIS due to Hukushima and Iba[25]—went in the reverse direction. In PMC, a resampling-based SIS facilitates MCS by adaptively choosing transition kernels that lead to most efficient exploration of the target distribution. Del Moral et al.[26] proposed an extremely flexible theoretical framework for an effective symbiosis of IS and MCS in population-based simulation methods for sequential Monte Carlo problems; see also Jasra et al.[27] and Fernhead.[28] In these methods, a pool of draws is generated sequentially in an interactive, parallel manner. MCS guides local exploration by each stream in the pool and importance sampling enables the pool to decide how to efficiently redistribute its streams in the vast space it is trying to explore. Such confluence of MCS and IS will be an important direction for future Monte Carlo research.

REFERENCES

1. Hammersely JM, Morton KW. Poor man's Monte Carlo. *J Roy Stat Soc., Ser B* 1954, 16:23–38.
2. Rosenbluth MN, Rosenbluth AW. Monte Carlo calculations of the average extension of molecular chains. *J Chem Phys* 1955, 23:356–359.
3. Geweke J. Bayesian inference in econometric models using Monte Carlo integration. *Econometrica* 1989, 57:1317–1339.
4. Ghosh JK, Delampady M, Samanta T. *An Introduction to Bayesian Analysis: Theory and Methods*. New York: Springer; 2006.
5. Evans M, Swartz T. Methods for approximating integrals in statistics with special emphasis on Bayesian integration problems. *Stat Sci* 1995, 10:254–272.
6. Oh M-S, Berger JO. Adaptive importance sampling in Monte Carlo integration. *J Stat Comput Simulation* 1992, 41:143–168.
7. Kloek K, van Dijk HK. Bayesian estimates of equation system parameters: an application of integration by Monte Carlo. *Econometrica* 1978, 46:1–20.
8. Liu, JS. *Monte Carlo Strategies in Scientific Computing*. New York: Springer-Verlag; 2001.
9. Oh M-S, Berger JO. Integration of multimodal functions by Monte Carlo Importance Sampling. *J Am Stat Assoc* 1993, 88:450–456.
10. Richard J-F, Zhang W. Efficient high-dimensional importance sampling. *J Econometrics* 2007, 141: 1385–1411.
11. Owen A, Zhou Y. Safe and effective importance sampling. *J Am Stat Assoc* 2000, 95:135–143.
12. Hammersley JM, Handscomb DC. *Monte Carlo Methods*. Methuen's Monographs on Applied Probability and Statistics. Methuen, London: John Wiley & Sons; 1964.
13. Neal RM. Annealed importance sampling. *Stat Comput* 2001, 11:125–139.
14. Lyman E, Zuckerman DM. Annealed importance sampling of peptides. *J Chem Phys* 2007, 126:65101–65101.
15. Godsill SJ, Clapp TC. Improvement Strategies for Monte Carlo Particle Filters. In: Doucet A, De Freitas JFG, Gordon NJ, eds. *Sequential Monte Carlo Methods in Practice*. New York: Springer-Verlag; 2001.
16. Kong A, Liu JS, Wong WH. Sequential imputations and Bayesian missing data problems. *J Am Stat Assoc* 1994, 89:278–288.
17. Wall FT, Erpenbeck JJ. New methods for the statistical computation of polymer dimensions. *J Chem Phys* 1959, 30:634–637.
18. Grassberger P. Pruned-enriched Rosenbluth method: simulations of θ polymers of chain length up to 1,000,000. *Phys Rev E* 1997, 56:3682–3693.

19. Gordon NJ, Salmond DJ, Smith AFM. Novel approach to nonlinear non-Gaussian Bayesian state estimation. *IEEE Proc Radar Signal Processing* 1993, 140:107–113.
20. Liu JS, Chen R. Sequential Monte Carlo methods for dynamic systems. *J Am Stat Assoc* 1998, 93:1032–1044.
21. Chen Y, Xie J, Liu JS. Stopping-time resampling for sequential Monte Carlo methods. *J Roy Stat Soc., Ser B* 2005, 67:199–217.
22. Kingman JFC. On the genealogy of large populations. *J Appl Probab A* 1982, 19:27–43.
23. Griffiths RC, Tavaré S. Simulating probability distributions in the coalescent. *Theor Population Biol* 1994, 46:131–159.
24. Cappé O, Guillin A, Marin JM, Robert CP. Population Monte Carlo. *J Comput Graph Stat* 2004, 13:907–929.
25. Hukushima K, Iba Y. Population annealing and its application to a spin glass. *AIP Conf Proc* 2003, 690:200–206 (Monte Carlo Method in the Physical Sciences: J. E. Gubernatis Editor).
26. Del Moral P, Doucet A, Jasra A. Sequential Monte Carlo samplers. *J Roy Stat Soc., Ser B* 2006, 68: 411–436.
27. Jasra A, Stephens DA, Holmes CC. On population based simulation for static inference. *Stat Computing* 2007, 17:263–279.
28. Fernhead P. Computational methods for complex stochastic systems: a review of some alternatives to MCMC. *Stat Computing* 2008, 18:151–171.

