Truncated importance sampling

E. L. Ionides†

University of Michigan, Department of Statistics
Technical report # 424, August 2005. Revised October 2005.

Summary. Importance sampling is a fundamental Monte Carlo technique. It involves generating a sample from a proposal distribution in order to estimate some property of a target distribution. Importance sampling can be highly sensitive to the choice of proposal distribution, and fails if the proposal distribution does not sufficiently well approximate the target. Truncated importance sampling is shown theoretically to improve on standard importance sampling by being less sensitive to the proposal distribution and having lower mean squared estimation error. Consistency is shown under weak conditions, and optimal truncation rates found under more specific conditions. Truncation at rate $n^{1/2}$ is shown to be a good general choice. An example demonstrates that truncated importance sampling is effective as a component of a sequential importance sampling scheme for a continuous time population disease model.

Keywords: Importance sampling; Truncation; Monte Carlo; Sequential Monte Carlo

1. Introduction

Importance sampling is a basic Monte Carlo tool (Liu, 2001; Bernardo and Smith, 1994). A typical goal is to estimate $H = E_f[h(X)] = \int h(x)f(x)\,dx$ by importance sampling using $X_1, \ldots, X_n$ drawn from a density $g(x)$. Here, $f(x)$ is called the target density and $g(x)$ is the proposal density. The standard unbiased estimate of H is

$$H_n = \frac{1}{n}\sum_{i=1}^{n} h(X_i)w_i$$

where $w_i = w(X_i) = f(X_i)/g(X_i)$. This estimate may have infinite variance if the tail behaviour of $g(x)$ is not sufficiently similar to $h(x)f(x)$. To avoid this, $g(x)$ must be chosen carefully so that $h(x)f(x)/g(x)$ does not get too large. Standard methods to choose $g(x)$ (Bernardo and Smith, 1994, Section 5.5.3) can struggle in complex, high dimensional importance sampling situations. We introduce truncated importance sampling estimate as readily applicable, theoretically justifiable method which reduces the sensitivity of importance sampling to the choice of $g(x)$. Truncation is also found to reduce the Monte Carlo mean square error for importance sampling in a broad class of situations. The truncated importance sampling estimate is

$$H'_n = \frac{1}{n}\sum_{i=1}^{n} h(X_i)w'_i$$

†*Address for correspondence*: Department of Statistics, The University of Michigan, 439 West Hall, 1085 South University Ave, Ann Arbor MI 48109-1107, USA.
E-mail: ionides@umich.edu

where $w_i' = w_i \wedge \tau_n$, the minimum of w_i and τ_n.

There are two distinct motivations for importance sampling. Firstly, drawing from $f(x)$ may be hard, in which case $g(x)$ may be chosen to be convenient for simulation. Secondly, $g(x)$ can be chosen to reduce Monte Carlo variability. This second motivation can be in opposition to the first: drawing from $f(x)$ may be easy, but result in prohibitive Monte Carlo variability. Choosing $g(x) \propto h(x)f(x)$ results in $\mathrm{Var}(H_n) = 0$, but calculating $w(x) = f(x)/g(x)$ then requires knowledge of the normalizing constant $\int h(x)f(x)\,dx$. For the first motivation h is not of primary interest for the choice of g, whereas for the second the particular h is critical. Truncation can be applied when importance sampling is motivated by either concern. A feature of H_n' is that h plays no direct role in the truncation, which depends only on how well g can approximate the tail of f.

Truncation is shown in Section 2 to give mean square consistency under weak conditions. Section 3 determines optimal truncation rates, requiring more assumptions. A recommended rate for routine use is $\tau_n = n^{1/2}$. Section 4 discusses a toy example, for which bias and variance can be calculated analytically. Section 5 demonstrates that truncation can be of practical interest, when truncation allows the use of a linearization to give a proposal distribution. Section 5 presents a fairly elaborate epidemiological model: one could study simpler examples, but the more complex model demonstrates a type of situation where truncation is expected to be particularly useful. For complex models it may be relatively easy to find a function $g(x)$ which well approximates the center of $h(x)f(x)$, but harder to approximate the tails. As increasing computational capabilities lead to the consideration of increasingly complex stochastic models and Monte Carlo inference techniques, truncation methods for importance sampling may have an increasing role to play. Recent applications of importance sampling to complex stochastic models include population genetics (Stephens and Donnelly, 2000), finance (Glasserman et al., 1999) and signal processing (Arulampalam et al., 2002). The intuition behind the effectiveness of truncation is that it gives less weight to the part of the space that $g(x)$ cannot approximate effectively based on a sample of size n, which otherwise leads to large Monte Carlo variability. This heuristic is discussed further in Section 6.

2. Consistency of truncated importance sampling

Let b_n and V_n be the bias and variance of H_n'. Supposing that $h(x)f(x) = 0$ whenever $g(x) = 0$, the bias may be calculated as

$$\begin{aligned} b_n = E_g[H_n'] - H &= \int_{x:g(x)>0} h(x)((w(x) \wedge \tau_n) - w(x))g(x)dx \\ &= \int_{x:f(x)>\tau_n g(x) \text{ and } g(x)>0} h(x)(\tau_n g(x) - f(x))dx. \end{aligned}$$

Since $|h(x)(\tau_n g(x) - f(x))| \le |h(x)f(x)|$, dominated convergence gives $b_n \to 0$ if $\tau_n \to \infty$ as long as $E_f[|h(X)|] < \infty$. To bound the variance,

$$\begin{aligned} E_g[(h(X)w'(X))^2] &= \int_{x:g(x)>0} h(x)^2(w(x) \wedge \tau_n)^2 g(x)\,dx \\ &\le \tau_n \int_{x:g(x)>0} h(x)^2 w(x) g(x)\,dx \\ &\le \tau_n E_f[h(X)^2]. \end{aligned}$$

Thus, $V_n = \mathrm{Var}_g(H'_n) \leq \tau_n E_f[h(X)^2]/n$ and so $V_n \to 0$ as long as $E_f[h(X)^2] < \infty$ and $\tau_n/n \to 0$. This gives very general conditions for the truncated importance sampling to give mean square consistent estimators, when there is no such guarantee for the standard version. If we know more about the tail behaviour of $f(x)$, $g(x)$ and $h(x)$ we can get optimal rates for τ_n, as shown in Section 3 below.

A consistency argument similar to the above applies for the estimator

$$H''_n = \frac{1}{n}\sum_{i=1}^{n}(h(X_i)w(X_i) \wedge \tau_n) \vee (-\tau_n), \tag{1}$$

where $h(X_i)w(X_i)$ is truncated, rather than $w(X_i)$. A substantial advantage of using H'_n over H''_n is that the weights have a natural unit scale ($E_g[w(X)] \leq 1$), which is used in Sections 4 and 5 for selecting τ_n in practice. Arguing heuristically, it is also undesirable that H''_n tends to introduce bias by truncating extreme values of $h(x)$ which contribute disproportionately to H. If $g(x)$ is a reasonable proposal distribution then extreme values of $w(x)$ should typically correspond to small values of $h(x)$, allowing the truncation to reduce Monte Carlo variability without introducing excessive bias.

3. Optimal rates

To get optimal rates, we require rather more assumptions than the consistency argument. The resulting theoretical investigation still leads to some useful and possibly surprising findings. For X drawn from g, let $Z = w(X)$ and suppose that Z has a density $f_Z(z)$. For truncated and standard importance sampling it is not necessary that Z should have a density, but the assumption is convenient for the analysis of this section. We also assume, as in Section 2, that $h(x)f(x) = 0$ whenever $g(x) = 0$. To study the tail behaviour of Z we suppose that $f_Z(z) \sim z^{-(\alpha+2)}$, meaning that there exist some z_0, a and b such that $az^{-(\alpha+2)} < f_Z(z) < bz^{-(\alpha+2)}$ for all $z > z_0$. The property $E_g[Z] < \infty$ implies that $\alpha > 0$. Suppose initially that h is bounded; this may arise when using importance sampling for integrating out unobserved variables to calculate a likelihood, as in the example of Section 5. The bias may be calculated as

$$\begin{aligned} b_n &= \int_{\tau_n}^{\infty} E_g[h(X)|Z = z](\tau_n - z) f_Z(z)\, dz \\ &\sim \tau_n^{-\alpha}. \end{aligned} \tag{2}$$

For $\alpha < 1$, $\mathrm{Var}_g(Z) = \infty$. This leads us to look at two separate cases for bounding V_n.

Case (i) $\alpha < 1$. We find that $V_n \sim n^{-1}\tau_n^{1-\alpha}$ since

$$\begin{aligned} E_g[(h(X)w'(X))^2] &= \int_0^{\infty} E_g[h(X)^2|Z = z](z \wedge \tau_n)^2 f_Z(z)\, dz \\ &\sim \tau_n^{1-\alpha}. \end{aligned} \tag{3}$$

A bias-variance trade off, to minimize $b_n^2 + V_n$, suggests $\tau_n \sim n^{1/(1+\alpha)}$. This gives a mean square convergence rate of $b_n^2 + V_n \sim n^{-2\alpha/(1+\alpha)}$.

Case (ii) $\alpha > 1$. Now $E_g[(h(X)w'(X))^2]$ is no longer determined by the tails, and we find the usual importance sampling rate $V_n \sim n^{-1}$. We can still show that truncation gives a

higher order reduction in mean square error. The change in variance due to truncation is

$$\begin{aligned} V_n - \mathrm{Var}_g(H_n) &= n^{-1}\Big\{H^2 - (H+b_n')^2 + \int_{\tau_n}^{\infty} E_g[h(X)^2|Z=z](z^2-\tau_n^2) f_Z(z)\,dz\Big\} \\ &\sim n^{-1}\tau_n^{1-\alpha} \end{aligned} \tag{4}$$

Since $V_n - \mathrm{Var}_g(H_n) \le 0$, a bias-variance trade off to minimize mean square error suggests the same truncation rate $\tau_n \sim n^{1/(1+\alpha)}$ as for $\alpha > 1$.

One slightly strange feature of these rate calculations is that the more pathological cases (α small) require less truncation (i.e. a higher τ_n) than with larger α. This is because the b_n^2 term dominates for small α. As α increases, b_n^2 decreases faster than V_n so the optimal rate is obtained by decreasing τ_n to control V_n. From a practical point of view, setting $\tau_n = n^{1/2}$ is an attractive choice since it gives the optimal first order rate and an advantageous higher order correction with $\alpha > 1$, and more generally assures consistency. There is a hazard associated with using $\tau_n \sim n^{\beta}$ with $\beta < 1/2$: although this will give a good convergence rate for $1/(1+\alpha) \le \beta$, one risks losing a possible rate $b_n^2 + V_n \sim n^{-1}$ if $1 < \alpha < 1/2\beta$. It would be unfortunate to lose first order optimality in pursuit of higher order optimality.

The calculations in this section can be generalized, and remain essentially unchanged, for $h(x)$ unbounded but sufficiently slowly varying. For example, if $h(x)$ is a polynomial in x and $w(x)$ increases exponentially then $E_g[h(X)^k|Z=z]$ increases logarithmically with z. One can then replace (2) by $b_n \sim \tau_n^{-\alpha+\epsilon}$ for any $\epsilon > 0$, with similar adjustments required to (3) and (4). Importance sampling with polynomially bounded h plays a role, for example, in pricing financial options (Glasserman et al., 1999). The rate calculations in this section can also be modified to apply to H_n'' in (1) without a polynomial bound on h, setting $Z = h(X)w(X)$. Although this favours the use of H_n'', H_n' is still preferred for the reasons given in Section 2.

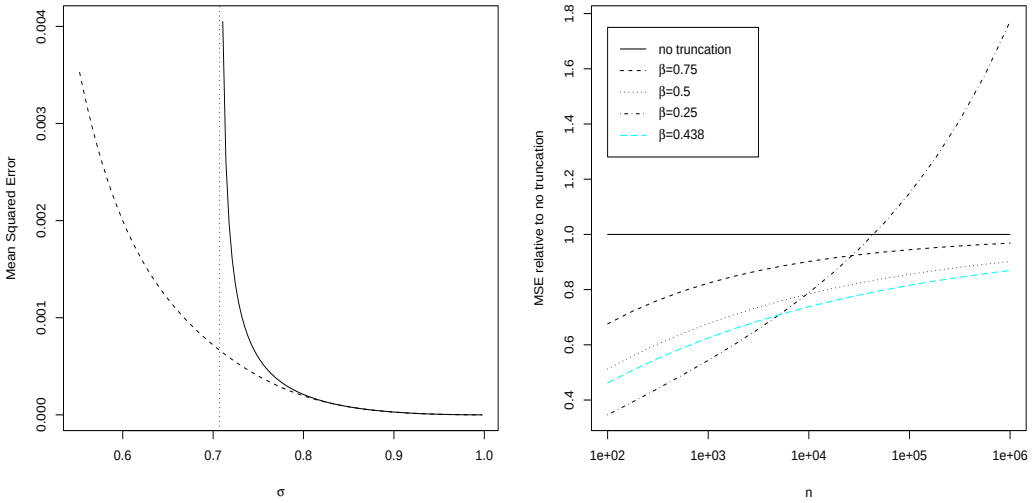
The rates calculated in this section correspond to a worst case scenario, and in specific cases improved rates may be possible. For example, if $\alpha > 1$ then $\mathrm{Var}_g(H_n) \sim n^{-1}$ but if $g(x) \propto h(x)f(x)$ then $\mathrm{Var}_g(H_n) = 0$. The rate calculations in (2), (3) and (4) are still formally correct in such special cases (in this example, $0 \sim n^{-1}$) but the bias-variance tradeoff argument no longer gives the truncation rate minimising mean square error.

4. A toy example

We consider a toy example, with $f(x) = (1/\sqrt{2\pi})e^{-x^2/2}$ and $g(x) = (1/\sqrt{2\pi\sigma^2})e^{-x^2/2\sigma^2}$ for $\sigma < 1$. In the notation of Section 3, $Z = \sigma(g(X)\sqrt{2\pi\sigma^2})^{\sigma^2-1}$ and

$$f_Z(z) \propto (\log(z/\sigma))^{-1/2} z^{-\left(2+\sigma^2/(1-\sigma^2)\right)}.$$

Ignoring the slowly varying term $(\log(z/\sigma))^{-1/2}$ this corresponds to $\alpha = \sigma^2/(1-\sigma^2)$. One would not intentionally get oneself into the kind of situation caricatured here, where the proposal distribution has shorter tails than the target. In higher dimensional situations, this can occur more readily by accident. One reason for this is that, if $f(x)$ and $g(x)$ are densities on $\mathbb{R}^d$, the tails of the proposal distribution cannot be uniformly larger than $|x|^{-d}$ in order that $\int g(x)\,dx = 1$. A related explanation invokes the "curse of dimensionality": the difficulty of importance sampling typically increases exponentially with the dimension d, so in higher dimensions it is necessary to take increasing care in the choice of proposal



(a) **(b)**

Fig. 1. **(a)** The solid line shows the MSE of H_n with $n = 1000$ plotted as a function of σ, with an asymptote at $\sigma = \sqrt{2}/2$ (dotted line). The dashed line shows the MSE of H'_n for $\tau_n = \sqrt{n}$. **(b)** The ratio of the MSE with truncation at $\tau_n = n^\beta$ to the MSE without truncation, as a function of n for $\sigma = 0.75$, for values of β given in the key. The asymptotically optimal rate is $\beta = 1/(1+\alpha) = 0.438$.

distribution. Choosing a relatively flat proposal distribution, insensitive to the particular target distribution, is not computationally viable as d becomes large.

When carrying out truncation at rate $\tau_n \sim n^\beta$, we use $\tau_n = Cn^\beta$. The weights have a natural unit scale ($E_g[w(X)] = 1$), so we take $C = 1$ as a default value. Figure 1 plots mean square error (MSE), as n and σ vary, for the case $h(x) = 1$. This example is analytically tractable, and the formulae used for Figure 1 are presented in Appendix A. Figure 1(a) shows that large gains are available using truncation in situations where standard importance sampling becomes unstable. Figure 1(b) takes a value $\sigma = 0.75$ where truncation is starting to make a marked difference, and looks at the effect of varying n. Figure 1(b) shows that, for this example, the choice of β becomes more important than the choice of C by $n \approx 10^4$. For smaller values of n, Fig 1(b) suggests that, in this example, small gains in the MSE would be possible by taking $\tau_n = Cn^{0.438}$ with $C < 1$.

Resampling techniques could be developed to determine a good value of C and/or β. For example, one could generate a large number n of importance samples. Treating this as the true (discrete) distribution, the target integral (which becomes a finite sum) could be approximated by resampling with a sample size of m for various values of $m < n$. Truncation levels to minimise MSE could be calculated for resampling with sample size m, and then extrapolated to recommend a truncation level for a sample of size n. We do not investigate this possibility here. Our opinion (supported by the examples of Sections 4 and 5) is that most of the advantages of truncation can be achieved with $C = 1$ and β determined by asymptotic considerations. Computational effort can then be focused on making n as large as possible.

Table 1. Monte Carlo variance of the estimates of the expected log likelihood per unit time, using SIS as described in the text with parameter values from Figure 2. The last column corresponds to a SIS algorithm using a standard particle filter method. These results are based on 10 realizations from the model, each with 480 time points (40 years). For each realization, SIS was repeated 5 times with an importance sample of size n. The Monte Carlo variance was estimated by the sample variance between SIS repetitions, within each realization and time point. Each realization was made approximately stationary by discarding a "burn in" of 60 time points (5 years). Simulation error was present in the last significant figure shown. Comparisons between the variances have reduced error since the same model realizations were used throughout.

n	$\tau_n = n^{1/4}$	$\tau_n = n^{1/2}$	$\tau_n = n^{3/4}$	$\tau_n = n$	$\tau_n = \infty$	PF
100	0.0136	0.0155	0.0192	0.0244	0.08	0.59
200	0.0072	0.0085	0.0120	0.022	0.13	0.22
400	0.0037	0.0047	0.0101	0.027	0.13	0.10
800	0.0022	0.0034	0.0094	0.041	0.19	0.05

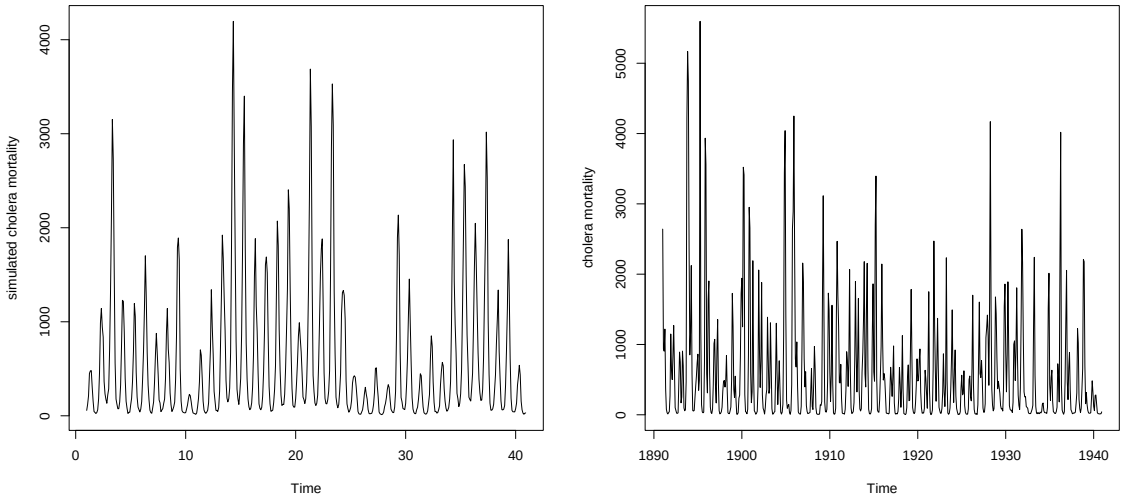
5. Application to a population model for cholera

Importance sampling can be used as a tool for nonlinear filtering (Arulampalam et al., 2002) and thus for likelihood based inference concerning nonlinear time series models. Here we consider one particular model arising from population disease dynamics. A more general framework in which this example falls is given in Appendix B.

Cholera is endemic to India and Bangladesh, and has recently become established in Africa, south Asia and South America (Sack et al., 2004). *Vibrio cholerae* can flourish in warm coastal waters. The role of ecosystems and climate are not fully understood. Challenges in unravelling the epidemiology/ecology include the non-linear dynamics of the disease and the uncertain role of immunity. We consider a model for cholera dynamics that is a continuous time version of a discrete time compartment model considered by Koelle and Pascual (2004), following a similar discrete time model for measles by Finkenstädt and Grenfell (2000). Discrete time models have some features that are accidents of the discretization. Working in continuous time avoids this, and also in principle allows inclusion of covariates measured at various time intervals.

A basic compartment models has N_t individuals grouped as $S_t = \#$ susceptible; $I_t = \#$ infected (and infectious); $R_t = \#$ recovered or removed. Compartment models (Bartlett, 1956) can be discrete time or continuous time, deterministic or stochastic, discrete population or continuous population. The real world is stochastic with a discrete population and continuous time. Imagining a continuous-valued population allows an approach of writing down stochastic differential equations, which have interpretable coefficients and allow a flexible modeling framework: the method allows covariates, or other modelling features such as additional compartments, to be added. We consider the following model:

$$\begin{pmatrix} dS_t \\ dI_t \\ dR_t \end{pmatrix} = \begin{pmatrix} -1 & 0 & 1 \\ 1 & -1 & 0 \\ 0 & 1 & -1 \end{pmatrix} \begin{pmatrix} \mu_t^{SI} dt + \sigma_t^{SI} dB_t^{SI} \\ \mu_t^{IR} dt + \sigma_t^{IR} dB_t^{IR} \\ \mu_t^{RS} dt + \sigma_t^{RS} dB_t^{RS} \end{pmatrix} \tag{5}$$



(a) **(b)**

Fig. 2. **(a)** A realization from (5) with $N_t = 2.5 \times 10^6$, $\alpha = 0.2$, $b_0 = 1.2$, $b_1 = 0.8$, $\gamma = 1$, $D = 1$, $\theta = 25$, $m = 1/30$, $\rho = 0.1$. The parameters were chosen by analogy with the discrete time model of Koelle and Pascual (2004), combined with some trial and error. **(b)** Historical data for Dhaka, Bangladesh. The model captures the key features of seasonal variation with immunity-driven inter-annual variation. Features present in the data that could be incorporated into the model include two peaks per year (the model has one) and non-stationarity (the population of Dhaka increased from 2.5×10^6 to 4×10^6 over this time period, indicating corresponding changes in infection rate and/or the chance of an infection leading to death).

$$\begin{aligned} \mu_t^{SI} &= (\beta_t I_t + \theta) S_t / N_t & \sigma_t^{SI} &= \alpha \mu_t^{SI} \\ \mu_t^{IR} &= \gamma I_t & \sigma_t^{IR} &= \sqrt{\mu_t^{IR}} \\ \mu_t^{RS} &= m R_t & \sigma_t^{RS} &= \sqrt{\mu_t^{RS}} \\ \beta_t &= b_0(1 + b_1 \cos(2\pi t/12)) \end{aligned}$$

Only N_t and the number of cases, C_t, are observed. C_t is modelled as

$$C_t \sim N[\rho I_t, D\rho(1-\rho) I_t].$$

Here, time is measured in months and C_t is observed at integer values of t. S_t, I_t and R_t are unobserved, and are defined in continuous time. A fundamental issue for inference concerning population dynamics is calculating the likelihood of the data C_t. One approach to doing this is via sequential Monte Carlo (Gordon et al., 1993), otherwise known as sequential importance sampling (SIS). Using this SIS estimate of the likelihood for Bayesian or classical inference is beyond the scope of this paper, and the reader is referred to Hürzeler and Künsch (2001) and Liu and West (2001). SIS consists of repeating many importance sampling steps, and at each step truncation may be appropriate. For efficiency of SIS, the choice of the proposal distribution can be critical. Locally in time and space, a diffusion process has a good linear, Gaussian approximation. Using a local linear Gaussian approximation to (5) and C_t, we calculate an approximation to the conditional diffusion

of S_t, I_t and R_t given the following observation, $C_{\lceil t \rceil}$. SIS was then applied using the approximate conditional diffusion as the proposal. The details of the linearization and the implementation of SIS are given in Appendix B.

Figure 2 illustrates (5) applied to a historical cholera mortality time series analysed at greater length by Koelle and Pascual (2004). Table 1 shows the Monte Carlo variances when calculating the log likelihood by SIS at different truncation levels. The last column compares with a standard SIS algorithm, which we call PF for "particle filter" (Arulampalam et al., 2002, Algorithm 4), where the proposal is the unconditional diffusion given by (5). For PF, the issue of truncating extremely large weights does not arise. The conditional diffusion approximation took approximately 3 times longer to compute than PF. When a $n^{1/2}$ truncation was employed, the conditioning procedure decreased the Monte Carlo variance by a factor of about 20 compared to PF. Note that it is desirable to have

$$\text{variance per unit time} \times \#\text{ of time points} \ll 1$$

in order for the Monte Carlo variability to have negligible effect on later inference. Without truncation ($\tau_n = \infty$), or with too little truncation ($\tau_n = n$) the conditioning approximation fails to give a convergent importance sampling estimate. Similarly, $\tau_n = n^{3/4}$ leads to a slowly decreasing variance. The bias was too small to be detected reliably by a reasonably sized simulation experiment, in part because the unbiased estimators ($\tau_n = \infty$ and PF) have high variability. We consider that Table 1 supports the use of $\tau_n = n^{1/2}$, since the Monte Carlo variance was found to be well behaved and theory reassures us that the bias will be negligible for large enough n. We deduce that the linearization used for the proposal distribution may not be a good approximation for the tails of the target distribution, but that truncated SIS allows effective use of this imperfect approximation. For a model of this fairly modest complexity it is not clear how else to get a superior approximation to the target distribution to avoid the need for truncation. Even if one could do this for some specific model, linearization and truncated importance sampling is a general technique that makes it routine to add extra features to the model, such as a sequence of different compartments for decreasing levels of immunity.

6. Discussion

Heuristically, truncation is effective because it follows the principle of not trying to estimate that which cannot be estimated well. Being instructed to ignore difficulties is pleasant advice to follow, and so truncation of importance weights (perhaps routinely using a $n^{1/2}$ truncation level) should become a standard technique for those who practice Monte Carlo importance sampling. Other statistical examples of this principle are naïve Bayes (Bickel and Levina, 2004), wavelet thresholding (Donoho and Johnstone, 1994), and perhaps shrinkage techniques in general. An analogy with soft wavelet thresholding suggests using soft truncation for importance sampling, say $w' = \gamma w + (1-\gamma)\tau_n$ with $\gamma = 1/(1+e^{\delta(w-\tau_n)})$ for some $\delta > 0$. Although a hard threshold has the unnatural feature of not being continuously differentiable, and may be slightly inferior, we suspect that the small gains that may be possible by soft truncation will not usually be worth the trouble.

The minimum mean square error is not a perfect criterion for evaluating Monte Carlo procedures. For some purposes the variance is more critical than the bias, and in these situations truncation is particularly attractive. For example, when the goal is to find some parameter θ maximising $H(\theta) = \int h(x,\theta) f(x,\theta) dx$, the bias may be relatively unimportant

as long as it varies slowly with θ. This occurs when using importance sampling to evaluate and maximise a likelihood function. For parameter estimation, even if the size of the bias is unknown, the success of the method can be assessed by testing on simulated data with known parameter values. In other situations, such as, pricing financial options (Glasserman et al., 1999), bias is certainly relevant and assessing the success of truncation may be more problematic. Truncation makes the most difference when the unbiased, standard importance sampling estimator is unreliable. This is exactly the situation, as in the example of Section 5, where estimating the bias due to truncation is difficult. If the truncated importance sampling estimate were markedly different from the untruncated estimate, one might want to start looking for a better proposal distribution. Meanwhile, until a better proposal is found, the truncated importance sampling estimate should be more reliable as long as mean squared error is a relevant criterion.

Truncation is not a panacea that will enable successful importance sampling using a very poor choice of proposal distribution. Truncation does allow the successful use of some proposal distributions whose poor approximations to the tails of the target distribution would otherwise render them useless. This is particularly relevant in complex stochastic models, where finding a proposal distribution that is everywhere a good approximation to the target may be challenging. Sensitivity to the choice of proposal distribution has been cited as a major draw-back of importance sampling (Glasserman et al., 1999) and sometimes a considerable amount of work has gone into refining a proposal distribution to make it practically useful (Stephens and Donnelly, 2000). By reducing the sensitivity to the proposal distribution, truncation should make importance sampling more readily applicable to new problems.

7. Acknowledgements

The author acknowledges helpful comments from Kerby Shedden, Jun Liu and two anonymous referees. Mercedes Pascual gave advice on the cholera example. Manno Bouma provided the cholera data for Figure 2. The author was supported by National Science Foundation grant 0430120.

A. Details for the example in Section 4

Recall that $f(x) = (1/\sqrt{2\pi})e^{-x^2/2}$, $g(x) = (1/\sqrt{2\pi\sigma^2})e^{-x^2/2\sigma^2}$, $h(x) = 1$ and $\sigma < 1$. We calculate

$$\mathrm{Var}(H_n) = \frac{1}{n}\left(\frac{\sigma^2}{\sqrt{2\sigma^2 - 1}} - 1\right)$$

for $\sigma^2 > 1/2$, with $\mathrm{Var}(H_n) = \infty$ for $\sigma^2 \leq 1/2$. Setting $u = \sqrt{2\log\left(\tau_n/\sigma\right)/\left(1/\sigma^2 - 1\right)}$,

$$b_n = 2\left[\tau_n \Phi\left(-u/\sigma\right) - \Phi(-u)\right]$$

where $\Phi(.)$ is the standard normal cumulative distribution function. Then,

$$\mathrm{Var}\left(H_n'\right) = \frac{1}{n}\left(\frac{\sigma}{\sqrt{2\pi}}\int_{-u}^{u} \exp\left\{\frac{x^2}{2}\left(\frac{1}{\sigma^2} - 2\right)\right\} dx + 2\tau_n^2\Phi(-u/\sigma) - \left(1 + b_n\right)^2\right).$$

For $\sigma^2 > 1/2$, $\mathrm{Var}(H_n')$ has a closed form via

$$\frac{\sigma}{\sqrt{2\pi}}\int_{-u}^{u} \exp\left\{\frac{x^2}{2}\left(\frac{1}{\sigma^2} - 2\right)\right\} dx = \frac{\sigma^2}{\sqrt{2\sigma^2 - 1}}\left(2\Phi(u/\sqrt{2 - 1/\sigma^2}) - 1\right),$$

whereas for $\sigma^2 \leq 1/2$ this integral was computed numerically.

B. Details for the example in Section 5

A general setting for Section 5 involves a diffusion x_t in $\mathbb{R}^p$ given by the Itô solution to a stochastic differential equation (SDE)

$$dx_t = \mu(x_t, t)\, dt + \sigma(x_t, t)\, dB_t \tag{6}$$

where B_t is Brownian motion in $\mathbb{R}^q$ and σ is a $p \times q$ matrix. A set of observations at discrete times, say $k = 1, \ldots, K$, is denoted by $y_{1:K}$. We suppose that $y_k \in \mathbb{R}^r$, and that y_k given x_k consists of a draw from some density $f(y_k|x_k)$. We abuse notation to let the argument of $f(\cdot|\cdot)$ denote the density in question. We also assume that all required densities and conditional densities exist.

Sequential importance sampling (Arulampalam et al., 2002) involves iteratively using a sample drawn (approximately) from $f(x_{k-1}|y_{1:k-1})$ to generate a sample from (approximately) $f(x_k|y_{1:k})$. At each iteration, importance sampling can be used to calculate

$$H = f(y_k|y_{1:k-1}) = \int f(y_k|x_k) f(x_k|y_{1:k-1})\, dx_k \tag{7}$$

An attractive proposal distribution comes from conditioning $\{x_t, k-1 \leq t \leq k\}$ on y_k. This conditional diffusion solves an SDE with the same infinitesimal variance as (6) but with a modified drift term, say

$$dx_t = \hat{\mu}(x_t, t)\, dt + \sigma(x_t, t)\, dB_t.$$

Although $\hat{\mu}$ cannot usually be readily calculated, a linear approximation is available, namely

$$\tilde{\mu}(x_t, t) = \mu + \sigma\sigma^T C^T (C\sigma\sigma^T C^T (k-t) + \psi)^{-1} (y - C(x_t + (k-t)\mu) - D) \tag{8}$$

where $y_k \approx Cx_k + D + \zeta$ with $E[\zeta] = 0$ and $\mathrm{Var}(\zeta) = \psi(x_t)$. This corresponds to the zeroth order linearization of Roberts and Stramer (2001). Truncated importance sampling to estimate H in (7) can now be carried out as follows:

(i) Generate n sample paths $\{x_t^{(i)}, k-1 \leq t \leq k,\, i = 1, \ldots, n\}$ using the drift $\tilde{\mu}$ given in (8), supposing inductively the availability of starting points $\{x_{k-1}^{(i)}, i = 1, \ldots, n\}$ drawn from $f(x_{k-1}|y_{1:k-1})$.

(ii) Calculate importance weights using Girsanov's Theorem (Oksendal, 1998),

$$w_i = \exp\Big\{ \int_{k-1}^{k} (\mu - \tilde{\mu})^T (\sigma\sigma^T)^{-1}\, dx_t^{(i)} + \frac{1}{2} \int_{k-1}^{k} \left(\tilde{\mu}^T (\sigma\sigma^T)^{-1} \tilde{\mu} - \mu^T (\sigma\sigma^T)^{-1} \mu \right) dt \Big\}$$

(iii) Calculate $H_n' = (1/n) \sum_{i=1}^{n} f(y_k|x_k{=}x_k^{(i)})(w_i \wedge \tau_n)$.

(iv) Resample $\{x_k^{(i)}\}$ with weights $f(y_k|x_k{=}x_k^{(i)})(w_i \wedge \tau_n)$ to give the (approximate) sample from $f(x_k|y_{1:k})$ inductively required in (i). A systematic resample (Arulampalam et al., 2002, Algorithm 2) was used in preference to a simple weighted random sample.

An Euler approximation (Kloeden and Platen, 1999) was used for (i) in Section 5, with a step length of 1/5 month. For (ii), the integral was approximated numerically by a sum based on this same discretization. The results in this paper justify the use of truncated weights in (iii) but we applied the same truncation also in (iv) where similar motivation applies. The log likelihood of $y_{1:K}$ can be found using (7) via the identity $\log f(y_{1:K}) = \sum_{k=1}^{K} \log f(y_k|y_{1:k-1})$.

References

Arulampalam, M. S., Maskell, S., Gordon, N., and Clapp, T. (2002). A tutorial on particle filters for online nonlinear, non-Gaussian Bayesian tracking. *IEEE Trans. Sig. Proc.*, 50:174 – 188.

Bartlett, M. S. (1956). Deterministic and stochastic models for recurrent epidemics. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, Volume 4*, pages 81–109.

Bernardo, J. M. and Smith, A. F. M. (1994). *Bayesian Theory*. Wiley, Chichester.

Bickel, P. J. and Levina, E. (2004). Some theory for Fisher's linear discriminant function, "naive Bayes", and some alternatives when there are many more variables than observations. *Bernoulli*, 10:989–1010.

Donoho, D. L. and Johnstone, I. M. (1994). Ideal spatial adaptation by wavelet shrinkage. *Biometrika*, 81:425–455.

Finkenstädt, B. F. and Grenfell, B. T. (2000). Time series modelling of childhood diseases: A dynamical systems approach. *Appl. Statist.*, 49:187–205.

Glasserman, P., Heidelberger, P., and Shahabuddin, P. (1999). Asymptotically optimal importance sampling and stratification for pricing path-dependent options. *Mathematical Finance*, 9:117–152.

Gordon, N., Salmond, D. J., and Smith, A. F. M. (1993). Novel approach to nonlinear/non-Gaussian Bayesian state estimation. *IEE Proceedings–F*, 140(2):107–113.

Hürzeler, M. and Künsch, H. R. (2001). Approximating and maximising the likelihood for a general state-space model. In Doucet, A., de Freitas, N., and Gordon, N. J., editors, *Sequential Monte Carlo Methods in Practice*, pages 159–175. Springer, New York.

Kloeden, P. E. and Platen, E. (1999). *Numerical Soluion of Stochastic Differential Equations*. Springer, New York, 3rd edition.

Koelle, K. and Pascual, M. (2004). Disentangling extrinsic from intrinsic factors in disease dynamics: A nonlinear time series approach with an application to cholera. *The American Naturalist*, 163:901–913.

Liu, J. and West, M. (2001). Combining parameter and state estimation in simulation-based filtering. In Doucet, A., de Freitas, N., and Gordon, N. J., editors, *Sequential Monte Carlo Methods in Practice*, pages 197–224. Springer, New York.

Liu, J. S. (2001). *Monte Carlo Strategies in Scientific Computing*. Springer, New York.

Oksendal, B. (1998). *Stochastic Differential Equations*. Springer, New York, 5th edition.

Roberts, G. O. and Stramer, O. (2001). On inference for partially observed nonlinear diffusion models using the Metropolis-Hastings algorithm. *Biometrika*, 88:603–621.

Sack, D. A., Sack, R. B., Nair, G. B., and Siddique, A. K. (2004). Cholera. *Lancet*, 363:223–233.

Stephens, M. and Donnelly, P. (2000). Inference in molecular population genetics. *J. Roy. Statist. Soc. B*, 62:605–655.