Truncated importance sampling

Edward L. Ionides
Department of Statistics, The University of Michigan, Ann Arbor, MI 48109–1107
E-mail: ionides@umich.edu

Abstract

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. Procedures which involve truncation of large importance sampling weights are 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 adaptive truncation threshold, based on minimizing an unbiased risk estimate, is also presented. As an example, truncation is found to be effective for calculating the likelihood of partially observed multivariate diffusions. It is demonstrated as a component of a sequential importance sampling scheme for a continuous time population disease model. Truncation is most valuable for computationally intensive, multi-dimensional situations in which finding a proposal distribution that is everywhere a good approximation to the target distribution is challenging.

Key Words: Monte Carlo; Diffusion; 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

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

where $w_i = w(X_i) = f(X_i)/g(X_i)$. This estimate may have infinite variance if the tail behavior 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) may perform poorly in complex, high dimensional importance sampling situations. We introduce truncated importance sampling as a 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 squared error for importance sampling in a broad class of situations. The truncated importance sampling estimate is

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

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. For example, when h is non-negative, a choice of $g(x) \propto h(x)f(x)$ 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 used for either reason. 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. The rate $\tau_n = n^{1/2}$ is found to have good asymptotic properties, and also performs well in the examples presented. Section 4 describes conditions under which the distribution of the truncated importance sampling estimator is asymptotically normal; again, the rate $\tau_n = n^{1/2}$ is favorable in this context. Section 5 introduces an adaptive truncation threshold, based on minimizing an unbiased risk estimate. Section 6 discusses a toy example, for which bias and variance can be calculated analytically. Section 7 develops an importance sampling approach to inference for partially observed diffusion processes, where truncation assists the use of a linearization to give a proposal distribution. Results are shown in Section 7.1 for a population model of cholera infection and mortality. This fairly elaborate epidemiological 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). With or without truncation, a good choice of the proposal density $g(x)$ reduces Monte Carlo variability. Truncation, however, alleviates the sensitivity to the tail behavior of $g(x)$. 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 8.

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

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

The integrand is bounded by $|h(x)(\tau_n g(x) - f(x))| \le |h(x)f(x)|$, since $f(x) > \tau_n g(x)$ over the region of integration. Thus, 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, $V_n = \mathrm{Var}_g(H_n') \le \tau_n E_f[h(X)^2]/n$ 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, while there is no such guarantee for the standard version. If we know more about the tail behavior 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

$$H_n'' = \frac{1}{n}\sum_{i=1}^{n}(h(X_i)w(X_i)\wedge\tau_n)\vee(-\tau_n), \tag{2}$$

where $h(X_i)w(X_i)$ is truncated, rather than $w(X_i)$. Since $h(x)$ is not necessarily positive, two-sided truncation is required. Arguing heuristically, it is 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 more assumptions than used for 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 behavior 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 7. The bias may be calculated as

$$\begin{aligned} b_n &= \int_{\tau_n}^{\infty} E_g[h(X)|Z=z](\tau_n - z) f_Z(z)\,dz \\ &\sim \tau_n^{-\alpha}. \end{aligned} \tag{3}$$

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

$$\begin{aligned} E_g[(h(X)w'(X))^2] &= \int_0^{\infty} E_g[h(X)^2|Z=z](z\wedge\tau_n)^2 f_Z(z)\,dz \\ &\sim \tau_n^{1-\alpha}. \end{aligned} \tag{4}$$

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 squared error. The reduction in variance due to truncation, which is always non-negative, is

$$\begin{aligned} r_n &= \mathrm{Var}_g(H_n) - V_n \\ &= n^{-1}\Big\{ \int_{\tau_n}^{\infty} E_g[h(X)^2|Z=z](z^2-\tau_n^2) f_Z(z)\, dz - H^2 + (H+b_n)^2 \Big\} \\ &\sim n^{-1}\tau_n^{1-\alpha} \end{aligned} \tag{5}$$

The mean squared error of H_n' is $\mathrm{Var}_g(H_n) - r_n + b_n^2$. Since τ_n plays no role in $\mathrm{Var}_g(H_n)$, truncation can be chosen with the goal of minimizing $b_n^2 - r_n$. Suppose $\tau_n \sim n^{\epsilon+1/(1+\alpha)}$, with $\epsilon > 0$. Then, r_n dominates b_n^2 and so truncation leads to a reduction in mean squared error of order $n^{(1-\alpha)(\epsilon+1/(1+\alpha))-1}$. This reduction increases as $\epsilon \downarrow 0$. In the limit, with $\epsilon = 0$, truncation with $\tau_n \sim n^{1/(1+\alpha)}$ may either provide the mean square optimal rate or give rise to a truncation rule performing worse than H_n. This depends on the relative size of the two terms r_n and b_n^2 which are both of order $n^{-2\alpha/(1+\alpha)}$.

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 those 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 (3) by $b_n \sim \tau_n^{-\alpha+\epsilon}$ for any $\epsilon > 0$, with similar adjustments required to (4) and (5). 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 (2) without a polynomial bound on h, setting $Z = h(X)w(X)$. Although this favors 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 $h(x) \ge 0$ and $g(x) \propto h(x)f(x)$ then $\mathrm{Var}_g(H_n) = 0$. The rate calculations in (3) and (5) are still formally correct in such special cases but the bias-variance trade-off argument no longer gives the truncation rate minimizing mean squared error.

4 Asymptotic normality

The distribution of truncated importance sampling estimators is asymptotically normal, under some general conditions. Here we discuss two results, postponing further details and proofs to Appendix A.

Theorem 1. *Suppose that $\tau_n n^{-1/2}$ is bounded and $\lim_{n\to\infty}\tau_n = \infty$. The estimator H_n'' has a Gaussian limit. If $h(X)$ is bounded (or, more generally, if $E[h(X)|w(X)]$ is bounded) then the estimator H_n' has a Gaussian limit.*

Theorem 1 shows that a truncation rate $\tau_n \sim n^{1/2}$ has a privileged position of combining asymptotic normality with the favorable properties of Section 3. If $\mathrm{Var}(H_n) = \infty$ then $\lim_{n\to\infty}\sqrt{n}\mathrm{Var}(H_n') = \infty$, and the asymptotic normality occurs at a different rate from the usual $\sqrt{n}$ central limit theorem for averages of independent variables. The requirement that h be bounded is sufficient but not necessary to imply a Gaussian limit for H_n'. The methods used to prove Theorem 1 could be extended to investigate other tail behaviors, though, in latent variable likelihood calculations such as Section 7, it is usual for h to be bounded. The $n^{1/2}$ bound on the rate of truncation is also not necessary. For example, Theorem 2 considers more general truncation rates than Theorem 1, at the expense of additional assumptions.

Theorem 2. *Using the notation of Section 3, let X have density g, let $Z = w(X)$ and suppose that Z has a density $f_Z(z)$ with tail behavior $f_Z(z) \sim z^{-(2+\alpha)}$ for $0 < \alpha < 1$. If $h(X)$ is bounded (or, more generally, if $E_g[h(X)|w(X)]$ is bounded) with $\lim_{n\to\infty}\tau_n = \infty$ and $\lim_{n\to\infty}\tau_n n^{-1/(1+\alpha)} = 0$, then H_n' has a Gaussian limit.*

5 Truncation selection via unbiased risk estimation

We present a finite sample method to choose τ for a particular finite weighted sample $\{(h(X_i), w_i), i = 1, \ldots, n\}$, via minimizing an estimate of the risk function. Similar approaches have been applied to regression model selection (Mallows, 1973) and wavelet coefficient thresholding (Donoho and Johnstone, 1994). The risk function $r(\tau)$ is taken to be the mean squared error,

$$r(\tau) = b^2(\tau) + V(\tau).$$

Here, $b(\tau)$ and $V(\tau)$ are the bias and variance of H_n', with the dependence on τ made explicit and the dependence on n suppressed. An unbiased estimator of $b(\tau)$ is $\hat{b}(\tau) = (1/n)\sum_{i=1}^n [(\tau - w_i)\wedge 0]h(X_i)$. Unbiased variance estimators $\widehat{\mathrm{Var}}(\hat{b}(\tau)) = [n(n-1)]^{-1}\sum_{i=1}^n \left\{[(\tau - w_i)\wedge 0]h(X_i) - \hat{b}\right\}^2$ and $\hat{V}(\tau) = [n(n-1)]^{-1}\sum_{i=1}^n \left\{(\tau \wedge w_i)h(X_i) - H_n'\right\}^2$ then give rise to an unbiased estimator of $r(\tau)$, namely

$$\hat{r}(\tau) = \hat{b}^2(\tau) - \widehat{\mathrm{Var}}(\hat{b}(\tau)) + \hat{V}(\tau).$$

This motivates a minimum unbiased risk estimate (MURE) truncation rule, taking $\hat{\tau}$ to minimize $\hat{r}(\tau)$, i.e, $\hat{\tau} = \arg\min_\tau \hat{r}(\tau)$. It is necessary to store the entire weighted sample to evaluate $\hat{\tau}$ and the corresponding estimator H_n'. The additional computation of implementing MURE is approximately linear in the sample size. If the effort of generating each weighted pair $(h(X_i), w_i)$ is considerable, the computational effort of MURE may be negligible. Alternatively, if large sample sizes are available, it may be preferable to rely on asymptotic truncation results.

MURE may also be used to calibrate the constant for an asymptotically justified truncation scheme. For example, if $\hat{\tau}_n$ is calculated based on a sample $\{(h(X_i), w_i), i = 1, \ldots, n\}$, and the sample size is later increased to $m > n$, then one may adopt a truncation threshold $\tau_m = \hat{\tau}_n(m/n)^{1/2}$. This avoids the need to repeat the MURE calculation whenever the sample size is increased.

To investigate the effectiveness of $\hat{\tau}$, $E[r(\hat{\tau})]$ is compared in Sections 6 and 7 with $r(n^\beta)$ and with the risk $\min_\tau r(\tau)$ corresponding to truncation at $\tilde{\tau} = \arg\min_\tau r(\tau)$. Comparison with $\tilde{\tau}$ demonstrates how much improvement would be possible if the mean-square optimal threshold,

which is not normally available to the practitioner, were revealed by an oracle. Hence, following Donoho and Johnstone (1994), $\tilde{\tau}$ is referred to as the "oracle" truncation threshold.

For both the examples of Sections 6 and 7, $\hat{\tau}$, $\tau = n^{1/2}$ and $\tilde{\tau}$ all have fairly similar performance. In every situation encountered by the author, $\hat{r}(\tau)$ had a unique local minimum for τ in the interval $[0, \max_i w_i]$ and was readily minimized by numerical methods. The expectation of $\hat{r}(\tau)$ is undefined when untruncated importance sampling has infinite variance, in which case $\mathrm{Var}(\hat{b}(\tau)) = \infty$. This does not imply that $E[r(\hat{\tau})] = \infty$, and the MURE procedure was found to be effective for the examples in Sections 6 even in situations where $\mathrm{Var}(H_n) = \infty$.

6 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 be harder to diagnose and to avoid. 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 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 $\tau_n = Cn^\beta$. In this example (in common with many, but not all, importance sampling situations) $E_g[w(X)] = 1$. The natural unit scale of the weights suggests taking $C = 1$ as a default value. This example is analytically tractable for $h(x) = 1$, and Fig. 1 plots mean squared error (MSE), as n and σ vary.

Fig. 1A demonstrates the increasing value of truncation as the tail of the proposal distribution becomes an increasingly poor approximation to the tail of the target distribution. The choice of $\tau_n = n^{1/2}$ performs well in the finite sample situation investigated in Fig. 1A, with an MSE closely matching oracle truncation. However, to make a fair comparison of truncation rates, one must vary n, since the rate β is inseparable from the constant C when n is fixed.

Fig. 1B takes a value $\sigma = 0.75$, and looks at the effect of varying n. At $\sigma = 0.75$, truncation is starting to make a marked difference though the untruncated estimator still has finite variance. For $\beta = 0.25$ bias eventually overwhelms variance, producing poor results at large values of n. Selecting $\beta = 0.75$ results in too little variance reduction. The asymptotically optimal truncation rate, identified in Section 3 case (ii), is at or near $\beta = 1 - \sigma^2 = 0.438$. MURE, $\beta = 0.5$ and $\beta = 1 - \sigma^2$ all result in comparable MSEs, a little larger than that for oracle truncation.

7 Example: inference for diffusions

Likelihood based inference for diffusion processes observed at discrete time points has received considerable attention (Elerian et al., 2001; Roberts and Stramer, 2001; Durham and Gallant, 2002; Aït-Sahalia, 2002; Beskos et al., 2006). Much previous work has emphasized one-dimensional diffusions, with methods that are difficult to apply to a broad class of multivariate diffusions.

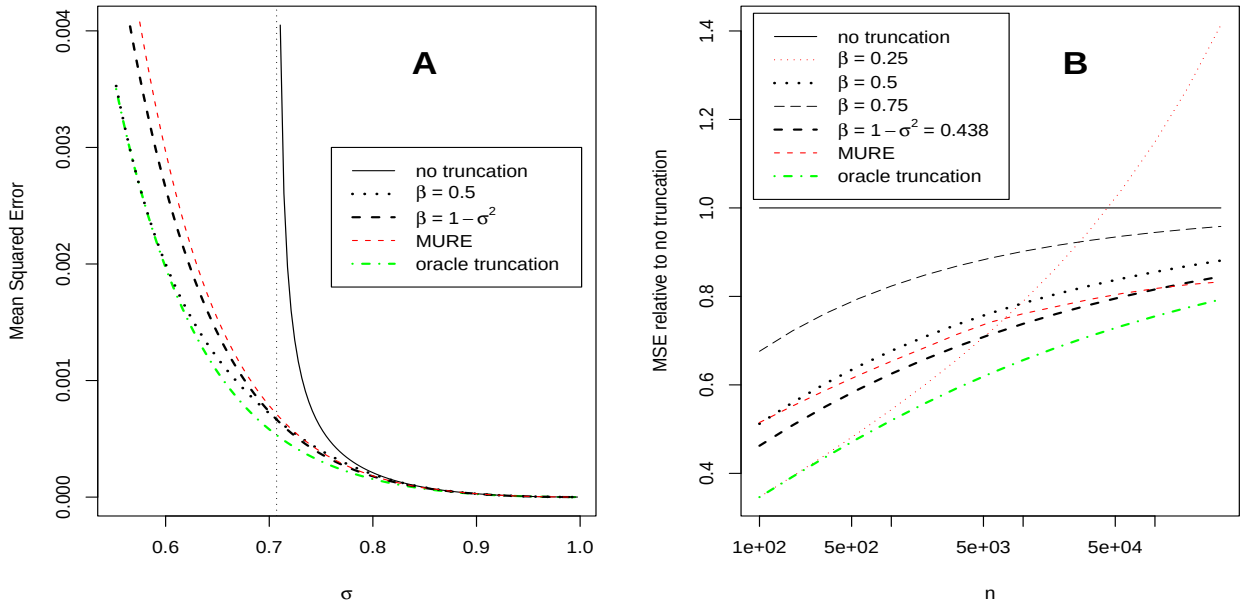

Figure 1: (A) The MSE of H_n for the example of Section 6, plotted as a function of σ with $n = 1000$. The solid line (no truncation) has an asymptote at $\sigma = \sqrt{2}/2$ (dotted line). The dashed lines represent truncation rules indicated in the key. Values of β correspond to truncation at $\tau_n = n^\beta$. Oracle truncation corresponds to the MSE minimized over all truncation levels. (B) The ratio of the MSE with truncation to the MSE without truncation, plotted as a function of n with $\sigma = 0.75$ for various truncation rules given in the key.

We consider an importance sampling approach to multivariate diffusions, also allowing for the possibility of unobserved components and/or measurement error. Our approach is essentially a multivariate extension of Durham and Gallant (2002). Truncated importance sampling facilitates the use of a proposal distribution arising from a linearization which is relatively immediate to apply to nonlinear multivariate diffusions with non-Gaussian measurement models. We first describe a general framework for inference from partially observed multivariate diffusions. In Section 7.1, we present a simulation study investigating the use of truncated importance sampling to evaluate the likelihood for a model arising from disease dynamics of cholera. In the fields of ecology and epidemiology, there is a need for development of methodology to permit inference for flexible classes of continuous time models (Bjørnstad and Grenfell, 2001).

Suppose x_t is a diffusion in $\mathbb{R}^p$ given by the Itô solution to the stochastic differential equation (SDE)

$$dx_t = \mu_t(x_t)\,dt + \sigma_t(x_t)\,dW_t \tag{6}$$

where W_t is Brownian motion in $\mathbb{R}^q$ and σ_t is a $p \times q$ matrix, giving rise to the infinitesimal variance $\Sigma_t = \sigma_t \sigma_t^T$. Observations $\{y_k\}$ occur at discrete times, say $k = 1, \ldots, K$. We suppose that $y_k \in \mathbb{R}^r$, and that y_k given x_k consists of a draw from some density $f_{Y|X}(y|x_k)$. To solve a nonlinear SDE we use the Euler method (Kloeden and Platen, 1999), where each time interval is discretized into N equal sub-intervals, each of length $\delta = 1/N$. The SDE is then replaced by a conditionally Gaussian stochastic difference equation,

$$X_{t+\delta} \sim N[X_t + \delta\mu_t(X_t), \delta\Sigma_t(X_t)]. \tag{7}$$

It may be appropriate to take (7), with sufficiently small δ, as the model of interest for a statistical application. This avoids the issue of how well (7) approximates (6), though the approximation is known to become increasingly accurate as $\delta \to 0$ under general conditions (Kloeden and Platen, 1999). An importance sampling approach to calculating the likelihood of a single observation y_k conditional on x_{k-1} (surveyed by Durham and Gallant, 2002) is to carry out Monte Carlo evaluation of

$$H = E_{\mu,\Sigma}[f_{Y|X}(y_k|X_k)|X_{k-1} = x_{k-1}], \tag{8}$$

where $E_{\mu,\Sigma}$ denotes expectation with X_t solving (7). The importance sampling paradigm involves sampling from a different difference equation, replacing $\mu_t(\cdot)$ and $\Sigma_t(\cdot)$ in (7) by $\nu_t(\cdot)$ and $\Psi_t(\cdot)$ respectively. The corresponding importance sampling identity is

$$H = E_{\nu,\Psi}\Big[\Big\{\prod_{j=0}^{N-1} \frac{\phi(X_{[j+1]} - X_{[j]}\,;\delta\mu_{[j]}, \delta\Sigma_{[j]})}{\phi(X_{[j+1]} - X_{[j]}\,;\delta\nu_{[j]}, \delta\Psi_{[j]})}\Big\} f_{Y|X}(y_k|X_k)\Big], \tag{9}$$

where $\phi(\,\cdot\,;\mu,\Sigma)$ is the multivariate Gaussian density function with mean μ and covariance matrix Σ, $X_{[j]} = X_{k-1+j\delta}$, $\nu_{[j]} = \nu_{k-1+j\delta}(X_{k-1+j\delta})$ and $\Psi_{[j]} = \Psi_{k-1+j\delta}(X_{k-1+j\delta})$. By choosing ν_t and Ψ_t carefully, one may hope for considerable gains over the naive proposal distribution given by $\nu_t = \mu_t$ and $\Psi_t = \Sigma_t$. A desirable choice of ν_t and Ψ_t would come from conditioning $\{x_t, k-1 \le t \le k\}$ on y_k. The resulting conditional diffusion solves an SDE with the same infinitesimal variance as (6) but with a modified drift term, say

$$dx_t = \hat{\mu}_t(x_t)\,dt + \sigma_t(x_t)\,dW_t.$$

Although $\hat{\mu}$ cannot usually be readily calculated, the conditioning can be carried out analytically if μ_t and σ_t are approximated to be constant and $f_{Y|X}(y|x)$ is approximated by a conditional

Gaussian density, $\phi(y; Ax + B, \Theta)$. The matrices A and Θ and the vector B can be functions of x_t and t. We take ν_t to be the drift of the resulting diffusion,

$$\nu_t(x_t) = \mu_t + \Sigma_t^T A^T (A\Sigma_t^T A^T (k-t) + \Theta)^{-1}(y_k - A(x_t + (k-t)\mu_t) - B). \tag{10}$$

This is a multivariate version of the zeroth order linearization of Roberts and Stramer (2001), and may be derived by standard calculations for multivariate normal random variables. If $\Psi_{[j]} = \Sigma_{[j]}$ then the importance weight in (9) is a Riemann sum discretization of the Radon-Nikodym derivative, given by Girsanov's Theorem (Øksendal, 1998), between two diffusion processes with infinitesimal drifts μ_t and ν_t. We take

$$\Psi_{[j]} = \begin{cases} \Sigma_{[j]} & \text{for } j = 0, \ldots, N-2, \\ \Sigma_{[N-1]} - \delta\Sigma_{[N-1]} A^T (A\Sigma_{[N-1]} A^T \delta + \Theta)^{-1} A\Sigma_{[N-1]} & \text{for } j = N-1. \end{cases} \tag{11}$$

The modified estimate in the last sub-interval is similar to a refinement introduced by Durham and Gallant (2002, Section 4); the substantial gains they reported from this refinement are in agreement with the author's experience.

The naive proposal distribution becomes arbitrarily inefficient as the measurement error of the observations becomes small (i.e., as the density $f_{Y|X}(y|x)$ approaches singularity). However, the linearization in (10) and (11) also becomes increasingly poor as x_t becomes increasingly nonlinear. Thus, the use of truncated importance sampling may become beneficial. Section 7.1 demonstrates one example of such a situation.

The importance sampling situation in (8) is an example of the second motivation given in Section 1, where the naive proposal may have intolerable Monte Carlo variability. Another problem of interest for this model is inference about unobserved variables given the data. One might look to find

$$E[h(X_k)|x_{k-1}, y_k] = \int h(x) f_{X_k|X_{k-1},Y_k}(x|x_{k-1}, y_k) dx \tag{12}$$

with, for example, $h(x) = x$ or $h(x) = x^2$. This falls into the first motivation, since sampling from $f_{X_k|X_{k-1},Y_k}(x|x_{k-1}, y_k)$ is difficult. The proposal distribution developed above for (8) would also be well suited for (12), unless $h(x)$ puts heavy emphasis on the tails of $f_{X_k|X_{k-1},Y_k}(x|x_{k-1}, y_k)$.

The joint likelihood of multiple observations is required for likelihood-based analysis of time series data, and this can be obtained by sequential importance sampling (Gordon et al., 1993; Liu and Chen, 1998; Liu, 2001; Arulampalam et al., 2002). Sequential importance sampling involves a sequence of dependent importance sampling calculations. Truncation of importance weights can be carried out at each step in the sequence, using the above method for each observation interval. Section 7.1 gives an illustration of inference carried out by sequential importance sampling with truncation.

7.1 A population model for cholera

Cholera is endemic to northeast India and Bangladesh, and has recently become established in Africa, South America, and elsewhere in south Asia (Sack et al., 2004). The disease is caused by a bacterium, *Vibrio cholerae*, which can flourish in warm coastal waters. The role of ecosystems and climate are not fully understood. Challenges in unraveling 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.

Compartment models are a basic tool for quantitative analysis of population dynamics (Kermack and McKendrick, 1927; Bartlett, 1960; Diekmann and Heesterbeek, 2000). A basic epidemiological compartment model has N_t individuals divided into three groups, with S_t denoting the number susceptible, I_t the number infected (and infectious), and R_t the number recovered or removed. Compartment models 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 permits an approach of writing down stochastic differential equations, which have interpretable coefficients and allow a flexible modeling framework: the method allows covariates, or other modeling features such as additional compartments, to be added. We consider the following model, with $x_t = (S_t, I_t, R_t)^T$

$$dx_t = \begin{pmatrix} dS_t \\ dI_t \\ dR_t \end{pmatrix} = \begin{pmatrix} -1 & 0 & 1 \\ 1 & -1 & 0 \\ 0 & 1 & -1 \end{pmatrix} \begin{pmatrix} \mu_t^{SI} dt + \sigma_t^{SI} dW_t^{SI} \\ \mu_t^{IR} dt + \sigma_t^{IR} dW_t^{IR} \\ \mu_t^{RS} dt + \sigma_t^{RS} dW_t^{RS} \end{pmatrix} \tag{13}$$

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

The population variables S_t, I_t and R_t are unobserved; N_t is treated as known, from census data; monthly case reports $\{y_k, k = 1, \ldots, K\}$ are observed. The observations process is taken to have over-dispersed binomial variation, modeled as $y_k = \rho I_k + (\rho(1-\rho)I_k)^{1/2}\xi_k$, where ξ_k is a Student's t random variable on τ degrees of freedom. A linear approximation to this observation equation for (10) and (11) is $A = (0, \rho, 0)$, $B = 0$ and $\Theta = I_t \rho(1-\rho)\tau/(\tau - 2)$.

A fundamental issue for inference concerning population dynamics is calculating the likelihood of the data $\{y_k\}$. Fig. 2 demonstrates likelihood calculation via truncated importance sampling for a single observation. Fig. 2A shows how importance sampling ensures that Monte Carlo simulations are consistent with the data. Fig. 2B compares the effect of different truncation rules. Similarly to Fig. 1, truncation at $\tau_n = n^{0.25}$ results in excessive bias for large n, whereas MURE and $\tau_n = n^{1/2}$ have an effect comparable to oracle truncation. For small sample sizes in Fig. 2B, $\hat{\tau}$ out-performs the oracle truncation level, $\tilde{\tau}$. This is possible because $\hat{\tau}$ is a random variable whereas $\tilde{\tau}$ only provides the minimum risk over non-random thresholds. From the evidence that truncation makes a marked improvement, 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 importance sampling allows more 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.

Fig. 3 illustrates likelihood based inference for time series data. Fig. 3A shows historical cholera mortality time series analyzed at greater length by Koelle and Pascual (2004). This is used to motivate the simulation results in Figs. 3B and 3C. Sequential importance sampling was carried out, with truncation applied to the importance weights at each step in the sequence. Based on the results in Fig. 2, a truncation level of $\tau_n = n^{1/2}$ was used. The unbiased risk estimate of Section 5 is not directly applicable due to the dependence introduced by sequential importance sampling. Perhaps the most direct strategy for inference via sequential importance sampling is to

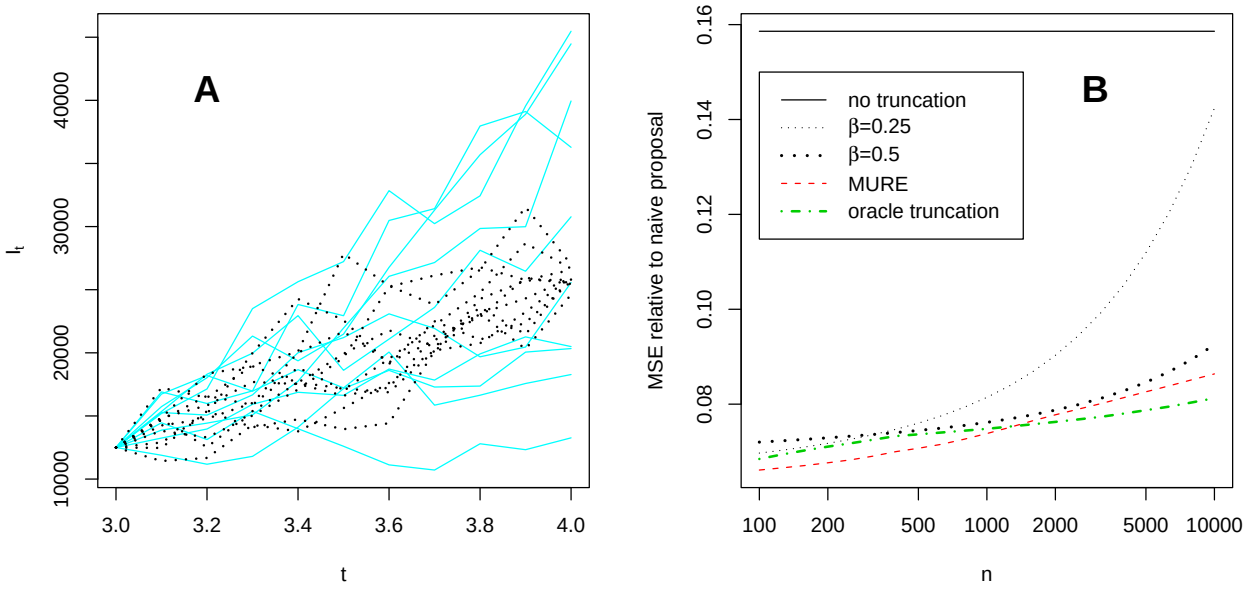

Figure 2: Investigation of a single observation interval, from March to April, for the model in (13) with $N_t = 2.5 \times 10^6$, $\alpha = 0.2$, $b_0 = 1.2$, $b_1 = 0.8$, $\gamma = 1$, $\theta = 25$, $m = 1/30$, $\rho = 0.1$, $\tau = 6$. The parameters were chosen by analogy with the discrete time model of Koelle and Pascual (2004), combined with some trial and error. The goal is to find the likelihood of $y_4 = 2600$ given $S_3 = 0.85N_3$ and $I_3 = 0.005N_3$. (A) Ten sample paths for the number of infected individuals, for the naive proposal distribution (solid lines) and the proposal distribution given by (10) and (11) (dotted lines). The discrete time step used to solve the SDE was $\delta = 0.1$. (B) Relative mean squared errors of various truncation rules for estimating the likelihood of y_4 via importance sampling. Values of β correspond to truncation at $\tau_n = n^\beta$. Results were calculated using an importance sample of size 2×10^6, which was treated as an entire population for the purposes of determining the "true" likelihood value and hence the MSE of a subsample of size n at a given truncation level. Oracle truncation corresponds to the MSE minimized over all truncation levels. The MSE for MURE was based on 5000 subsamples, each of size n, from this population.

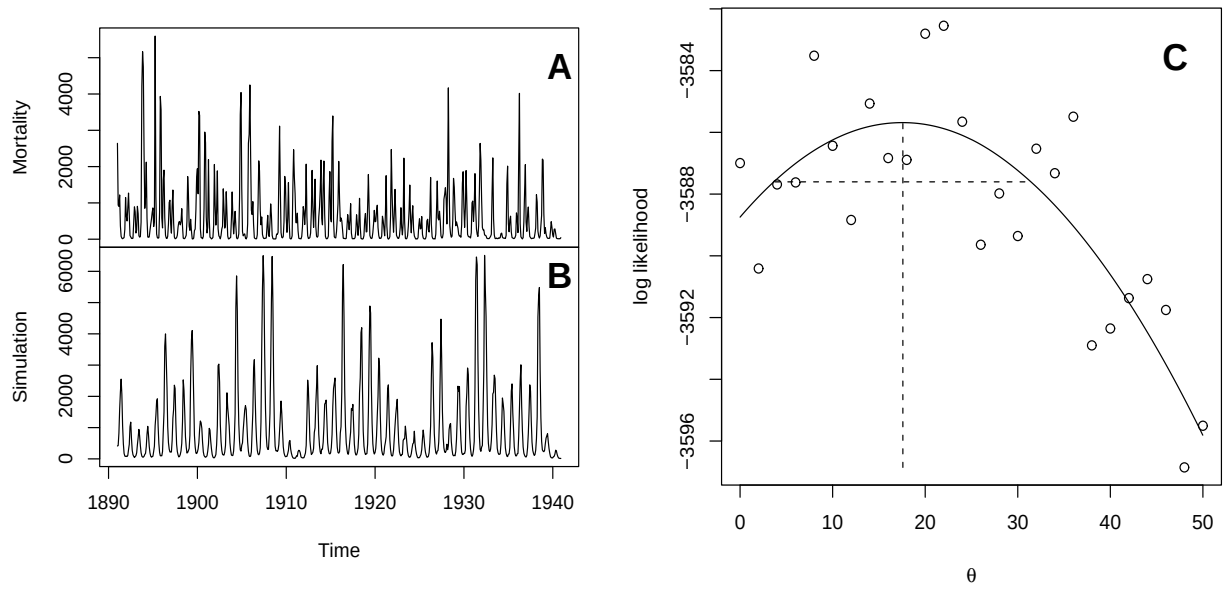

Figure 3: (A) Historical data on monthly cholera mortality for Dhaka, Bangladesh. (B) One realization from (13) with parameter values given in Fig. 2. The model captures the key features of seasonal variation with immunity-driven inter-annual variation. Features present in the data that could be incorporated 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, suggesting corresponding changes in infection rate and/or the chance of an infection leading to death). (C) The log likelihood, $\lambda(\theta)$, estimated for the simulated data by smoothing the likelihood evaluations (shown as points) arising from truncated importance sampling with $n = 1000$ and $\tau_n = n^{1/2}$. Local quadratic smoothing was recommended in this context by Ionides (2005), and was implemented here using the `loess` procedure in *R* (R Development Core Team, 2006). Dashed lines show the construction of the maximum likelihood estimate $\hat{\theta} = \arg\max \lambda(\theta)$ and an approximate 95% confidence interval given by $\{\theta : \lambda(\hat{\theta}) - \lambda(\theta) < (1/2)\chi^2_{0.95}(1)\}$ where $\chi^2_{0.95}(1)$ is the 0.95 quantile of a chi-squared random variable on 1 degree of freedom.

estimate the likelihood surface, as in Fig. 3C. A local quadratic approximation of the log likelihood (Ionides, 2005) can be used to construct confidence intervals, as shown here, or combined with a prior distribution to give a Bayesian analysis. The parameter θ investigated in Fig. 3C corresponds to a rate of infection independent of the current number of infected individuals, which may be interpreted as an environmental reservoir of the disease. The flat likelihood suggests that, in the context of the model (13), the data contain rather little information about the parameter θ. Monte Carlo variability much larger than one unit of log likelihood makes likelihood based inference difficult or impossible. Bias in estimating the log likelihood is of negligible importance for inference if it is a slowly varying function of θ, suggesting that a lower truncation level might be effective.

Previous methodology has been developed to avoid the need to reconstruct the likelihood surface from noisy Monte Carlo estimates. Many of these techniques are less broadly applicable than likelihood reconstruction (Hürzeler and Künsch, 2001; Cappé et al., 2005; Beskos et al., 2006), and cannot readily be applied to the general multivariate nonlinear diffusion framework of Section 7. Methods that employ basic sequential Monte Carlo techniques for parameter estimation by judiciously adding stochasticity to the parameters (Liu and West, 2001; Ionides et al., 2006) could be applied in situations such as Section 7, and could also potentially benefit from truncation of importance weights.

8 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 should become a standard technique for those who practice Monte Carlo importance sampling. Other statistical examples of this principle are naive 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 $\gamma = 1/(1+e^{\alpha(w-\tau_n)})$ for some $\alpha > 0$.

The minimum mean squared 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. One example is the likelihood surface estimation in Section 7.1. Maximum likelihood estimation is a specific case of the more general goal of finding some parameter θ maximizing $H(\theta) = \int h(x,\theta) f(x,\theta) dx$, for which the bias may be relatively unimportant as long as it varies slowly with θ. For parameter estimation, even if the size of the bias is unknown, the success of a method can be assessed by attempting to recover known parameters from simulated data. This could be used to investigate a good choice of C and/or β for truncation at $\tau_n = Cn^\beta$. Alternatively, the MSE of parameter estimation could replace the MSE of estimating H as a criterion for selecting a truncation level at a particular value of n. 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 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 quantity.

Sequential importance sampling, as employed in Section 7.1, is one example of a situation where truncation can be applied as an extension of the basic importance sampling problem in (1). Another

variation of importance sampling arises when only un-normalized weights $\tilde{w}_i = cw_i$ are available, with c an unknown constant. Standard practice is to self-normalize these weights, replacing (1) by $\sum_{i=1}^n h(X_i)\tilde{w}_i\{\sum_{j=1}^n \tilde{w}_j\}^{-1}$. Truncation is appropriate for both the numerator and denominator of this ratio estimator (see Appendix B for more details).

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.

9 Acknowledgments

The author thanks two referees, the editor and an associate editor for many helpful suggestions. The author acknowledges comments from Kerby Shedden, Stilian Stoev and Jun Liu. Mercedes Pascual gave advice on the cholera example. Menno Bouma provided the cholera data for Fig. 3. The author was supported by National Science Foundation grant 0430120.

A More on asymptotic normality

We present two propositions giving sufficient conditions for averages of independent truncated random variables to have a Gaussian limit distribution. Proposition 1 gives a general Gaussian limit result for truncation at rates less than or equal to $n^{1/2}$. Proposition 2 gives a similar result for a broader range of truncation rates under more specific conditions. When translated into the context of truncated importance sampling, these propositions give rise to Theorems 1 and 2 of Section 4.

Proposition 1. *Let $\{Z_n, n = 1, 2, \ldots\}$ be a sequence of independent, identically distributed random variables with finite expectation, $E[Z_k] = \mu$. Let τ_n be a sequence for which $\tau_n n^{-1/2}$ is bounded and $\lim_{n\to\infty} \tau_n = \infty$. Define $Y_{nk} = (Z_k \wedge \tau_n) \vee -\tau_n$, $\mu_n = E[Y_{nk}]$ and $\sigma_n^2 = \mathrm{Var}(Y_{nk})$. Then, for all α,*

$$\lim_{n\to\infty} P\Big[\frac{1}{\sigma_n\sqrt{n}}\sum_{k=1}^{n}(Y_{nk} - \mu_n) < \alpha\Big] = \int_{-\infty}^{\alpha} \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz. \tag{14}$$

Proof. We make use of necessary and sufficient conditions for a Gaussian limit of sums in a triangular array of independent random variables provided by Gnedenko and Kolmogorov (1954, Section 26, Theorem 2). In the present context, the conditions are that, for all $\epsilon > 0$,

$$\lim_{n\to\infty} nP\Big[\frac{|Y_{nk} - \mu_n|}{\sigma_n\sqrt{n}} < \epsilon\Big] = 0 \tag{15}$$

$$\lim_{n\to\infty} \mathrm{Var}\Big(\frac{Y_{nk}}{\sigma_n} I\Big\{\frac{|Y_{nk} - \mu_n|}{\sigma_n\sqrt{n}} < \epsilon\Big\}\Big) = 1 \tag{16}$$

where $I\{A\}$ is an indicator random variable, taking value 1 if A occurs and 0 otherwise.

If $\mathrm{Var}(Z_k) = \infty$ then $\lim_{n\to\infty} \sigma_n = \infty$. Since there is a C with $|Y_{nk}| \le Cn^{1/2}$, we have, for all sufficiently large n, $|Y_{nk} - \mu_n|/(\sigma_n\sqrt{n}) < \epsilon$. This implies (15), and (16) then follows from the definition of σ_n. Alternatively, consider the case $\mathrm{Var}(X_k) = \sigma^2 < \infty$ with $\lim_{n\to\infty} \sigma_n^2 = \sigma^2$. In this case, (14) describes a standard central limit result with rate $\sqrt{n}$. To check (15), notice that $\mathrm{Var}(Z_k) < \infty$ implies that $\lim_{x\to\infty} x^{-2}P[|Z_k| > x] = 0$. Dominated convergence gives $\lim_{n\to\infty} \mathrm{Var}[Y_{nk}I\{|Y_{nk} - \mu_n| < \epsilon\sigma_n\sqrt{n}\}] = \sigma^2$ from which (16) follows. □

Theorem 1 of Section 4 follows from Proposition 1 by setting $Z_i = h(X_i)w(X_i)$ (for H_n'') and $Z_i = w(X_i)$ (for H_n').

Proposition 2. *Adopt the notation of Proposition 1 but replace the bound on $\tau_n n^{-1/2}$ with a requirement that $\lim_{n\to\infty} \tau_n n^{-1/(1+\alpha)} = 0$ for some $0 < \alpha < 1$. If Z_k has a density $f(z)$ with tail behavior $f(|z|) \sim |z|^{-(2+\alpha)}$ for $0 < \alpha < 1$ then the Gaussian limit in (14) holds.*

Proof. From Section 3, we know $\sigma_n^2 \sim \tau_n^{1-\alpha}$. Conditions (15) and (16) hold immediately if $\lim_{n\to\infty} \tau_n/(\sigma_n\sqrt{n}) = 0$, which is equivalent to $\lim_{n\to\infty} \tau_n n^{-1/(1+\alpha)} = 0$. □

Theorem 2 of Section 4 follows from Proposition 2 by setting $Z_i = w(X_i)$.

B Self-normalized weights

In Monte Carlo Markov chain methodology, the importance weights are only calculated up to an unknown constant, $\tilde{w}_i = cw_i$. This motivates the analysis of the truncated self-normalized estimator

$$\tilde{H}_n = \frac{(1/n)\sum_{i=1}^n h(X_i)(\tilde{w}_i \wedge \tau_n^N)}{(1/n)\sum_{i=1}^n (\tilde{w}_i \wedge \tau_n^D)}. \tag{17}$$

Denote the numerator and denominator of (17) as N_n and D_n respectively. N_n and D_n both have the form of truncated importance sampling estimates with respective truncation thresholds τ_n^N and τ_n^D. Write $E[N_n] = cH + a_n$ and $E[D_n] = c + b_n$. If N_n and D_n have a joint Gaussian limit distribution, as in Appendix A, then the Delta method (Rao, 1973, Section 6a.2) gives a Gaussian limit for $\tilde{H}_n$. Specifically, $\tilde{H}_n$ is then asymptotically normal with mean and variance given by

$$\mu_n = H + [a_n - Hb_n]/c \tag{18}$$

$$\sigma_n^2 = [\mathrm{Var}(N_n) + H^2\mathrm{Var}(D_n) - 2H\mathrm{Cov}(N_n, D_n)]/c^2 \tag{19}$$

meaning that

$$\lim_{n\to\infty} P\Big[\frac{\tilde{H}_n - \mu_n}{\sigma_n} < \alpha\Big] = \int_{-\infty}^{\alpha} \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz.$$

The conditions for applying the Delta method here are $\lim_{n\to\infty} a_n = 0$, $\lim_{n\to\infty} \sigma_n = 0$ and $b_n = O(\sigma_n)$. The bias and variance in (18) and (19) have similar forms to the corresponding calculations for the basic truncated sampling estimator H_n'. This suggests that similar considerations apply to bias/variance tradeoffs for self-normalized estimators as for the un-normalized estimators. In particular, truncation should be applied to both the numerator and denominator of (17). In principle, the Delta method may allow the interpretation of (18) and (19) as limiting moments, regardless of whether N_n and D_n have a Gaussian limit (Oehlert, 1992). However, bounding the moments of the Taylor series remainder is technically more awkward than considering convergence in distribution.

The unknown constant c does not affect the analysis of asymptotic rates of truncation. From a practical point of view, the unbiased risk estimator of Section 5 is scale-invariant so may be

applied to generate thresholds τ_n^N and τ_n^D for the numerator and denominator of $\tilde{H}_n$ without the need to know c. Alternatively, a preliminary estimate of c could be used to calibrate the truncation threshold. For example, setting $\hat{c}_n = (1/n)\sum_{i=1}^n \tilde{w}_i$, one could then calculate $\tilde{H}_n$ with $\tau_n^N = \tau_n^D = \hat{c}_n n^{1/2}$.

References

Aït-Sahalia, Y. (2002). Maximum likelihood estimation of discretely sampled diffusions: A closed-form approximation approach. *Econometrica*, 70(1):223–262.

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

Bartlett, M. S. (1960). *Stochastic Population Models in Ecology and Epidemiology*. Wiley, New York.

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

Beskos, A., Papaspiliopoulos, O., Roberts, G., and Fearnhead, P. (2006). Exact and computationally efficient likelihood-based inference for discretely observed diffusion processes. *Journal of the Royal Statistical Society, Ser. B*, 68:333–382.

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.

Bjørnstad, O. N. and Grenfell, B. T. (2001). Noisy clockwork: Time series analysis of population fluctuations in animals. *Science*, 293:638–643.

Cappé, O., Moulines, E., and Rydén, T. (2005). *Inference in Hidden Markov Models*. Springer, New York.

Diekmann, O. and Heesterbeek, J. A. P. (2000). *Mathematical epidemiology of infectious diseases*. Wiley, New York.

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

Durham, G. B. and Gallant, A. R. (2002). Numerical techniques for maximum likelihood estimation of continuous-time diffusion processes. *Journal of Business and Economic Statistics*, 20(3):297–338.

Elerian, O., Chib, S., and Shephard, N. (2001). Likelihood inference for discretely observed non-linear diffusions. *Econometrica*, 69:959–993.

Finkenstädt, B. F. and Grenfell, B. T. (2000). Time series modelling of childhood diseases: A dynamical systems approach. *Applied Statistics*, 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.

Gnedenko, B. V. and Kolmogorov, A. N. (1954). *Limit Distributions for Sums of Independent Random Variables*. Addison-Wesley, Cambridge, MA.

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.

Ionides, E. L. (2005). Maximum smoothed likelihood estimation. *Statistica Sinica*, 15:1003–1014.

Ionides, E. L., Bretó, C., and King, A. A. (2006). Inference for nonlinear dynamical systems. *Proceedings of the National Academy of Sciences of the USA*, 103:18438–18443.

Kermack, W. O. and McKendrick, A. G. (1927). A contribution to the mathematical theory of epidemics. *Proceedings of the Royal Society of London, Ser. A*, 115:700–721.

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. *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.

Liu, J. S. and Chen, R. (1998). Sequential Monte Carlo methods for dynamic systems. *Journal of the American Statistical Association*, 93:1032–1044.

Mallows, C. L. (1973). Some comments on C_p. *Technometrics*, 15:661–675.

Oehlert, G. W. (1992). A note on the delta method. *The American Statistician*, 46:27–29.

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

R Development Core Team (2006). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Rao, C. R. (1973). *Linear Statistical Inference and Its Applications*. Wiley, New York, 2nd 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. *Journal of the Royal Statistical Society, Ser. B*, 62:605–655.