Importance Splitting for Rare Event Estimation

Michael Amrein and Hans R. Künsch

Seminar für Statistik, ETH Zurich

UNC Chapel Hill, February 22, 2010

Overview

The problem

Computing the probability of rare events for Markov models arises in many applications, e.g.
• Buffer overflow in queueing models
• Transitions between different modes of the stationary distribution in stochastic kinetic models.

Typical probabilities are in the range $10^{-6} - 10^{-20}$. Naive Monte Carlo produces estimates equal to zero even with long simulations. Want the relative mean square error to be bounded.

Importance sampling can be used, but constructing good proposal distributions where weights have small variance is difficult.

Importance splitting

Importance splitting (Garvels and Kroese 1998, Glasserman et al. 1999, Garvels 2000, L'Ecuyer et al. 2006) is an attractive alternative, based on "divide and conquer".

Write the event of interest E as the last member in a decreasing sequence of m events

$$D_m = E \subset D_{m-1} \subset \ldots \subset D_0 = \Omega.$$

Then

$$P(E) = \prod_{k=1}^{m} p_k \text{ where } p_k := P(D_k \mid D_{k-1}) = \frac{P(D_k)}{P(D_{k-1})}$$

and we estimate p_k for each k.

Rare events in Markov models

- (X_t) Markov process on E (discrete or continuous time).
- Starting point x_0 fixed (unless stated otherwise).
- A, B two disjoint subsets of E, $x_0 \notin B$, often $x_0 \in A$.
- $\tau =$ first hitting time of B, $\xi =$ first time X_t (re)enters A.
- Want to know $\gamma := P(\tau < \xi)$

Other choices of ξ are possible, e.g. a fixed time t_0, but make notation more complicated.

Defining the sets D_k

- Choose an importance function $\Phi : E \to \mathbb{R}$ with $A = \{x; \Phi(x) \leq 0\}$, $B = \{x; \Phi(x) \geq 1\}$.
- Choose m and levels $l_0 = \Phi(x_0) < l_1 < \ldots < l_m = 1$.
- Let $\tau_k =$ first hitting time of $\{x; \Phi(x) \geq l_k\}$ and $D_k = \{\tau_k < \xi\}$.
- For later use, define also $\mu_k =$ law of X_{τ_k} given D_k

In particular, $\tau_0 = 0$, μ_0 is the point mass in x_0 and $\tau_m = \tau$.

The recursive algorithm

- Assume we have a sample of size N_k from μ_k.
- Simulate independent chains starting at each point of this sample until $\min(\tau_{k+1}, \xi)$ and let R_{k+1} be the number of chains where $\tau_{k+1} < \xi$.
- If $R_{k+1} = 0$, set $\widehat{\gamma} = 0$ and stop.
- Otherwise

$$\widehat{p_{k+1}} = \frac{R_{k+1}}{N_k}$$

and inflate the sample of the R_{k+1} values of $X_{\tau_{k+1}}$ to a sample of size N_{k+1}. (Ties will disappear in the next step of the recursion)

Sample inflation strategies

- Fixed splitting: $N_{k+1} = c_{k+1} R_{k+1}$ with c_{k+1} given.
- Fixed effort, random inflation: N_{k+1} given, inflation by sampling with replacement
- Fixed effort, balanced inflation: N_{k+1} given, take each value $[N_{k+1}/R_{k+1}]$ times plus a sample without replacement
- Fixed number of successes (our proposal): $R_{k+1} \geq 2$ given, sample with replacement at the level k until this is achieved. Need to use

$$\widehat{p_{k+1}} = \frac{R_{k+1} - 1}{N_k - 1}$$

(the UMVU estimator for the negative binomial)

Choice of the importance function

Bad choices of Φ make importance splitting fail! Garvels et al. (2002) propose

$$\Phi(x) = g(P(\tau < \xi | X_0 = x))$$

with g monotone. If we cannot compute γ, we cannot compute this Φ either.

If E is discrete and (X_t) time homogeneous, we suggest as approximation of the above
$\Phi(x) =$ probability of the most likely path from x to B without entering A.
This can be computed by Dijkstra's algorithm. Choice of g equivalent to choice of levels (to be discussed later).

Unbiasedness

Is $\widehat{p}_k$ unbiased for p_k ?
Is $\widehat{\gamma}$ unbiased for γ ?

The answer is no for the first question, and yes for the second (for all versions above).

Some “Proofs” of unbiasedness of $\widehat{\gamma}$ have appeared which claim conditional unbiasedness of $\widehat{p_k}$ given the history up to level $k-1$.

Correct proofs for some versions are in Del Moral and Garnier (2005) and Dean and Dupuis (2009). We give a more direct proof for all cases.

Unbiasedness

Is $\widehat{p}_k$ unbiased for p_k ?
Is $\widehat{\gamma}$ unbiased for γ ?

The answer is no for the first question, and yes for the second (for all versions above).

Some “Proofs” of unbiasedness of $\widehat{\gamma}$ have appeared which claim conditional unbiasedness of $\widehat{p_k}$ given the history up to level $k-1$.

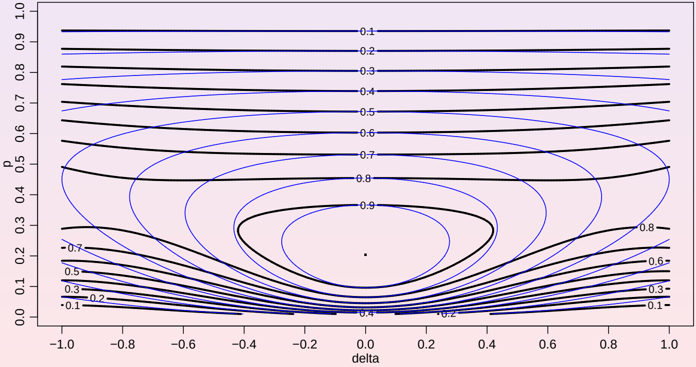
Correct proofs for some versions are in Del Moral and Garnier (2005) and Dean and Dupuis (2009). We give a more direct proof for all cases.

Advantages of fixed number of successes

- Fixed number of successes controls precision instead of effort.
- Fixed number of successes avoids the problem of returning estimates with the value zero.
- For a single step, fixed number of successes with $R_k = N_k p_k$ is slightly less efficient than fixed effort.
- For several steps, fixed effort with constant N_i is worse than fixed number of successes with constant R_i if the p_i are different.

Efficiencies for 2 steps

Efficiency for fixed effort (thin blue line) and fixed number of successes (thick black line) for $m = 2, p_1 = p^{1+\delta}, p_2 = p^{1-\delta}$.



Analysis assuming conditional independence

Analytic results in general situations are difficult to obtain. We make the simplifying assumption

$$P(D_{k+1}|D_k, X_{\tau_k} = x) = P(D_{k+1}|D_k) = p_k \quad (a.s.)$$

Although this is restrictive, it holds if

$$\Phi(x) = P(\tau < \xi | X_0 = x) \text{ and } \Phi(X_{\tau_k}) = l_k$$

("the process cannot jump over the levels").

We study how to choose the number of levels m, the probabilities p_k (subject to $\prod_{k=1}^{m} p_k = \gamma$) and the sample sizes N_k or R_k, respectively, to minimze workload W for a given relative mean square error q.

Main results

If workload is independent of m and k, then the optimal solution as $q \to 0$ is

$$m = -0.6275..\log(\gamma), \quad p_k \equiv 0.2032...$$

(the exact values can be written with the solution of the equation $\exp(1/c) = 2c/(2c-1)$). Furthermore

$$N_k \equiv n \sim 2.46 \frac{-\log(\gamma)}{q}, \quad R_k \equiv r \sim \frac{-\log(\gamma)}{2q}.$$

Can also express the optimal R_k's (or N_k's) for given m, q and $p_1, \ldots, p_m$.

Implications for the algorithm

Everything unknown at the beginning $\rightsquigarrow$ Two-stage procedure:

- In the first stage, take many levels and $R_k \equiv 20$.
- If the algorithm does not complete in reasonable time, find a better importance function (or buy a faster computer).
- Otherwise, we have initial guesses of γ and p_k. If some of the p_k seem close to zero or one, delete or introduce new levels.
- In a second stage, take $R_k \equiv r$ according to the formulae for optimal choice (where q is now the only tuning constant).
- Alternatively, take $\sqrt{r}$ replicates of the second stage with $\sqrt{r}$ number of successes. Average the results and compute a confidence interval by bootstrapping on the log scale.

Overflow in the G/G/1 queue

- Weibull inter arrival and service times with shape parameters k_a, k_s. Scale parameters λ, μ fixed.
- $X_t =$ (number of customers, remaining time until next arrival, remaining service time).
- $x_0 = (0, 0, 0)$, $A = \{x_1 = 0\}$, $B = \{x_1 = 104\}$.
- $\Phi(x) = x_1$.

Results

100 repetitions of the two-stage procedure above. Nominally $\mathrm{Var}(\widehat{\gamma}) = (0.1 \cdot \gamma)^2$, Workload $W(\widehat{\gamma})$ = generated random numbers.

k	0.75	1.00	1.25
$r \approx$	1290	2060	2910
		$r_k = r$	
$\langle\widehat{\gamma}\rangle$	$1.68 \cdot 10^{-6}$	$2.22 \cdot 10^{-10}$	$4.40 \cdot 10^{-15}$
$\widehat{\mathrm{Var}}(\widehat{\gamma})/\langle\widehat{\gamma}\rangle^2$	0.0134	0.0102	0.0116
$\langle W(\widehat{\gamma})\rangle$	$7.067 \cdot 10^6$	$2.254 \cdot 10^7$	$5.304 \cdot 10^7$
		$r_k = r^{1/2}$ with averaging	
$\langle\widehat{\gamma}\rangle$	$1.66 \cdot 10^{-6}$	$2.21 \cdot 10^{-10}$	$4.41 \cdot 10^{-15}$
$\widehat{\mathrm{Var}}(\widehat{\gamma})/\langle\widehat{\gamma}\rangle^2$	0.0197	0.0124	0.0127
$\langle W(\widehat{\gamma})\rangle$	$7.122 \cdot 10^6$	$2.259 \cdot 10^7$	$5.318 \cdot 10^7$

Overflow in the Jackson tandem queue

Two queues in series, arrival rate $\lambda = 1$, mean service times $\rho_i = 1/\mu_i$, $\rho_1 = 1/2$.
$X_t = (X_{1,t}, X_{2,t})$ = number of customers at both queues, $x_0 = (0, 0)$.
$A = \{x_0\}$, $B = \{x : x_2 \geq 30\}$. For $\rho_2 < \rho_1$, B is a rare event.
Can compute γ numerically.

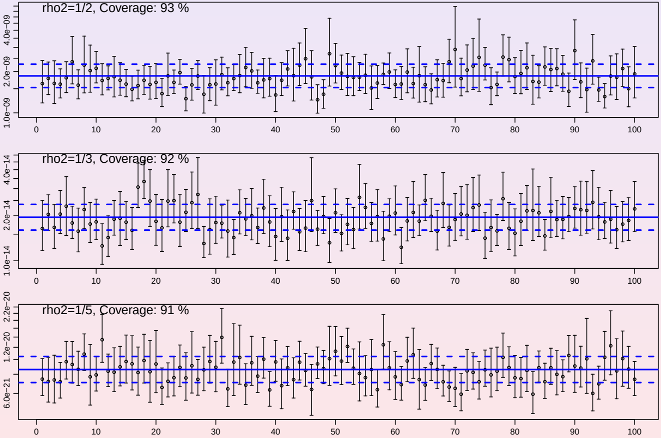
Naive importance function $\Phi(x) = x_2$, fails. Want $\Phi(x_1, x_2) < \Phi(x_1 + 1, x_2)$ for small x_1, because more customers at the first queue make B more likely. Our proposal provides this.

Results

100 repetitions of the two-stage procedure above with the averaged estimator. Nominally $\mathrm{Var}(\widehat{\gamma}) = (0.1 \cdot \gamma)^2$, Workload $W(\widehat{\gamma}) =$ generated random numbers.

ρ	1/2	1/3	1/5
γ	$1.86 \cdot 10^{-9}$	$1.94 \cdot 10^{-14}$	$8.59 \cdot 10^{-21}$
$\langle \widehat{\gamma} \rangle$	$1.85 \cdot 10^{-9}$	$1.92 \cdot 10^{-14}$	$8.58 \cdot 10^{-21}$
$\widehat{\mathrm{Var}}(\widehat{\gamma})/\gamma^2$	0.0321	0.0387	0.0340
$\langle W(\widehat{\gamma}) \rangle$	$3.704 \cdot 10^6$	$7.495 \cdot 10^6$	$2.183 \cdot 10^7$

100 Confidence intervals for overflow probability



A general framework (Feynman-Kac)

Want to sample successively from a sequence of target distribution μ_n on spaces F_n which are connected in the following way:

$$\mu_n(dz_n) = \frac{1}{M_n} g_n(z_n) \int_{F_{n-1}} \mu_{n-1}(dz_{n-1}) K_n(z_{n-1}, dz_n)$$

where M_n is a normalising constant, $g_n \geq 0$ and K_n is a transition kernel from F_{n-1} to F_n.

Look for weighted samples $(\zeta_{i,n}, \lambda_{i,n}; i = 1, 2, \ldots, N)$ that are constructed recursively for computational efficiency.

Filtering and smoothing in state space models

(Z_t) unobserved Markov process, observations Y_t conditionally independent of $(Z_s, Y_s; s \neq t)$ given Z_t.

$\mu_n =$ conditional distribution of Z_n given $Y_1, \ldots, Y_n$ (filtering), is an instance of the general framework if we put $g_n =$ conditional density of Y_n given Z_n, $K_n =$ transition kernel of Z_n given Z_{n-1}.

For $\mu_n =$ conditional distribution of $(Z_0, Z_1, \ldots, Z_n)$ given $(Y_1, \ldots, Y_n)$ (smoothing), g_n is as above and $K_n(z_{n-1}, .)$ is a point mass at z_{n-1} for components $0, 1, \ldots, n-1$ times the transition kernel of Z_n for the last component.

Importance splitting

Let $\Delta \notin E$ be an additional (absorbing) state and set

$$Z_n = X_{\tau_n} 1_{[\tau_n < \xi]} + \Delta 1_{[\tau_n > \xi]}.$$

If $g_n(z) = 1_{[z \neq \Delta]}$, then μ_n is the distribution of X_{τ_n} conditioned on $\tau_n < \xi$.

In this example, the normalizing constant is $M_n = P(D_n | D_{n-1})$, and we want to estimate it. In filtering, M_n is the conditional density of Y_n given $Y_1, Y_2, \dots, Y_{n-1}$ which is also of interest.

The vanilla version of the particle filter

- Resample: $\zeta^*_{i,n} = \zeta_{j,n}$ with probability $\lambda_{j,n}$
- Propagate: generate $\zeta_{i,n+1} \sim K_{n+1}(\zeta^*_{i,n}, dz_{n+1})$ independently
- Reweight: $\lambda_{i,n+1} \propto g_{n+1}(\zeta_{i,n+1})$
- Estimate:

$$\widehat{M}_{n+1} = \frac{1}{N} \sum_{i=1}^{N} g_{n+1}(\zeta_{i,n+1})$$

Resampling need not be i.i.d.. It suffices that the expected number of times $\zeta_{j,n}$ is chosen equals $N\lambda_{j,n}$.

More sophisticated versions

For more balanced weights, anticipate the effect of g_{n+1} at the resampling and propagation steps and adjust the weights:

- Choose a distribution ν for resampling, set $J_i = j$ with probability ν_j and $\zeta^*_{i,n} = \zeta_{J_i,n}$.
- Choose a kernel L for propagation and generate $\zeta_{i,n+1} \sim L(\zeta^*_{i,n}, dz_{n+1})$ independently,
- Reweight:

$$\lambda_{i,n+1} \propto \frac{\lambda_{J_i,n}}{\nu_{J_i}} \frac{dK_{n+1}(\zeta^*_{i,n}, \zeta_{i,n+1})}{dL(\zeta^*_{i,n}, \zeta_{i,n+1})} g_{n+1}(\zeta_{i,n+1}).$$

Implications from the connection

- Use asymptotic results (as number of particles $\to \infty$) for particle filtering to obtain corresponding results for importance splitting.
- In particular, functionals of the path from x to B given $\tau < \xi$ can also be estimated (but need large N).
- Particle filtering algorithms show how to combine importance sampling and importance splitting (need to consider the whole path on $[\tau_{n-1}, \tau_n]$ in order to compute the weights).

Implications, ctd.

- Unbiasedness of estimated normalizing constants is important also for particle filtering (used by Andrieu et al., JRSSB 2010). Notice that it implies that estimated log likelihood is biased.
- Working with random sample sizes to achieve a given effective sample size in the next step could be useful in particle filtering also (less extreme that accept/reject).

Thank you for your attention !