Available online at www.sciencedirect.com
SCIENCE DIRECT®
Statistics & Probability Letters 72 (2005) 277–283
STATISTICS & PROBABILITY LETTERS
www.elsevier.com/locate/stapro

Another look at rejection sampling through importance sampling

Yuguo Chen*,[1]

Institute of Statistics and Decision Sciences, Duke University, 216 Old Chemistry building, Durham, NC 27708, USA

Received 7 December 2004
Available online 19 February 2005

Abstract

We show that rejection sampling is inferior to the importance sampling algorithm in terms of the χ^2 distance between the proposal distribution and the target distribution. Similar conclusions are drawn for comparing rejection control with importance sampling.
© 2005 Elsevier B.V. All rights reserved.

Keywords: χ^2 distance; Effective sample size; Importance sampling; Rejection control; Rejection sampling

1. Introduction

Many scientific and statistical problems involve estimating the mean of a function $h(x)$ under distribution $\pi(x)$, i.e.,

$$\mu = \int h(x)\pi(x)\,\mathrm{d}x. \tag{1}$$

In many cases, especially when $\pi(x)$ is in high-dimensional space, directly generating samples from $\pi(x)$ is not possible. The Markov chain Monte Carlo methods circumvent the difficulty by generating dependent samples from a Markov chain whose equilibrium distribution is $\pi(x)$. Two

*Fax: +1 919 684 8594.
E-mail address: yuguo@stat.duke.edu.
[1]This research was partly supported by the National Science Foundation Grant DMS-0203762.

0167-7152/$ - see front matter © 2005 Elsevier B.V. All rights reserved.
doi:10.1016/j.spl.2005.01.002

other popular methods based on generating independent samples are rejection sampling (von Neumann, 1951) and importance sampling (Marshall, 1956).

Rejection sampling requires that there is a density function $g(x)$ and a constant M, such that the "envelope property"

$$\pi(x) \leqslant Mg(x) \tag{2}$$

holds on the support of $\pi(x)$. For convenience, we call a density function $g(x)$ satisfying (2) an envelope function. The rejection sampling algorithm can be described as follows.

Rejection sampling algorithm

(a) Draw a sample X from $g(x)$ and a sample U from a uniform distribution on $[0, 1]$;
(b) Accept X if $U \leqslant \pi(X)/Mg(X)$, and reject X otherwise.

Suppose we run the above algorithm N times, and the accepted samples are $X_{i_1}, \ldots, X_{i_k}$, then we can estimate μ by

$$\hat{\mu}^{\mathrm{RS}} = \frac{h(X_{i_1}) + \cdots + h(X_{i_k})}{k}, \tag{3}$$

because $X_{i_1}, \ldots, X_{i_k}$ are iid samples from $\pi(x)$.

Importance sampling, on the other hand, draws iid samples $X_1, \ldots, X_N$ from a proposal distribution $q(x)$, whose support contains the support of $\pi(x)$, and estimates μ by a weighted average

$$\hat{\mu}^{IS} = \frac{w_1 h(X_1) + \cdots + w_N h(X_N)}{w_1 + \cdots + w_N}, \tag{4}$$

where $w_i = \pi(X_i)/q(X_i)$, $i = 1, \ldots, N$, are the importance weights. Estimates (3) and (4) are preferred to unbiased estimates $M[h(X_{i_1}) + \cdots + h(X_{i_k})]/N$ and $[w_1 h(X_1) + \cdots + w_N h(X_N)]/N$ because in practice we may only know $\pi(x)$ up to a normalizing constant.

We point out in Section 2 that rejection sampling is an importance sampling algorithm with a particular choice of the proposal distribution. In Section 3 we show that importance sampling is a better choice than rejection sampling if the proposal distribution of importance sampling is chosen to be the same as the envelope function of rejection sampling. Section 4 compares rejection control with importance sampling. Section 5 provides concluding remarks.

2. Rejection sampling: a special importance sampling algorithm

We can look at rejection sampling from a different perspective by putting it in the general framework of importance sampling. Denote the support of the target distribution $\pi(x)$ as Ω and assume that density function $g(x)$ satisfies the envelope property (2). We can define a new

distribution $\pi^*(x,y)$ on the space $\Omega^* = \Omega \times [0,1]$ as

$$\pi^*(x,y) = \begin{cases} Mg(x) & \text{for } x \in \Omega,\ y \in \left[0, \dfrac{\pi(x)}{Mg(x)}\right], \\ 0 & \text{otherwise.} \end{cases} \tag{5}$$

Note that

$$\begin{aligned} \int_{\Omega^*} \pi^*(x,y)\,\mathrm{d}y\,\mathrm{d}x &= \int_{\Omega}\int_0^{\pi(x)/Mg(x)} Mg(x)\,\mathrm{d}y\,\mathrm{d}x = \int_{\Omega} \frac{\pi(x)}{Mg(x)}\, Mg(x)\,\mathrm{d}x \\ &= \int_{\Omega} \pi(x)\,\mathrm{d}x = 1. \end{aligned} \tag{6}$$

The argument in (6) also implies that for the joint distribution $\pi^*(x,y)$, the marginal distribution of X is still $\pi(x)$. This immediately leads to the new expression of the mean μ.

$$\mu = \int_{\Omega} h(x)\pi(x)\,\mathrm{d}x = \int_{\Omega^*} h(x)\pi^*(x,y)\,\mathrm{d}y\,\mathrm{d}x. \tag{7}$$

If we want to estimate μ by importance sampling when $\pi^*(x,y)$ is treated as the target distribution, a natural choice of the proposal distribution is

$$g^*(x,y) = g(x) \quad \text{for } (x,y) \in \Omega^*. \tag{8}$$

We can draw iid samples $(X_1, Y_1), \ldots, (X_N, Y_N)$ from $g^*(x,y)$ and estimate μ by

$$\hat{\mu}^* = \frac{w_1^* h(X_1) + \cdots + w_N^* h(X_N)}{w_1^* + \cdots + w_N^*}, \tag{9}$$

where the importance weights are

$$w_i^* = \frac{\pi^*(X_i, Y_i)}{g^*(X_i, Y_i)} = \begin{cases} M & \text{for } X_i \in \Omega,\ Y_i \in \left[0, \dfrac{\pi(X_i)}{Mg(X_i)}\right], \\ 0 & \text{otherwise,} \end{cases} \tag{10}$$

for $i = 1, \ldots, N$. The sample (X_i, Y_i) can be realized by drawing X_i from $g(x)$ and Y_i from a uniform distribution on $[0,1]$. Suppose $(X_{i_1}, Y_{i_1}), \ldots, (X_{i_k}, Y_{i_k})$ are the samples with nonzero weight, then the estimate $\hat{\mu}^*$ becomes

$$\hat{\mu}^* = \frac{Mh(X_{i_1}) + \cdots + Mh(X_{i_k})}{kM} = \frac{h(X_{i_1}) + \cdots + h(X_{i_k})}{k}. \tag{11}$$

This importance sampling algorithm and the rejection sampling algorithm in Section 1 are equivalent: the sampling procedure for (X, U) and (X, Y) are the same; the acceptance rule is equivalent to the weight assignment (10); and the final estimates (3) and (11) are also the same. Therefore, rejection sampling algorithm on the space Ω can be viewed as an importance sampling

procedure on the augmented space $\Omega^* = \Omega \times [0, 1]$. This provides a unified way to look at the two popular sampling procedures.

3. Comparison of rejection sampling and importance sampling

The envelope function $g(x)$ in the rejection sampling algorithm can also serve as the proposal distribution for importance sampling, because the envelope property (2) guarantees that the support of g contains Ω, the support of $\pi(x)$. We refer to the importance sampling algorithm with target distribution $\pi(x)$ and proposal distribution $g(x)$ as IS-1. In Section 2, we showed that rejection sampling with envelope function $g(x)$ is equivalent to importance sampling with proposal distribution $g^*(x, y)$ (defined in (8)) and target distribution $\pi^*(x, y)$ (defined in (5)). We refer to this importance sampling algorithm as IS-2. Both IS-1 and IS-2 are estimating μ, which can be interpreted as the mean of $h(x)$ with respect to the underlying distribution $\pi(x)$ or $\pi^*(x, y)$ (see the two equivalent expressions of μ in (7)). Since the standard deviations of the two estimates based on IS-1 and IS-2 in general depend on the function $h(x)$, here we consider a "function-free" criterion: the χ^2 distance between the proposal distribution and the target distribution.

The χ^2 distance between two distributions $p(x)$ and $q(x)$ is defined as

$$\chi^2(p, q) = \int \frac{[p(x) - q(x)]^2}{q(x)}\,\mathrm{d}x = \mathrm{var}_q\left[\frac{p(X)}{q(X)}\right]. \tag{12}$$

When $p(x)$ is the proposal distribution and $q(x)$ is the target distribution, the χ^2 distance between p and q is the same as the square of the *coefficient of variation* (cv^2) of the importance weight $w(x) = p(x)/q(x)$, which is defined as $\mathrm{var}_q w(X)$. Kong et al. (1994) propose to use the *effective sample size* (ESS) to measure the overall efficiency of an importance sampling algorithm:

$$\mathrm{ESS} = \frac{N}{1 + \mathrm{cv}^2} = \frac{N}{1 + \chi^2(p, q)}.$$

Heuristically, the ESS measures how many iid samples are equivalent to the N weighted samples. The smaller the χ^2 distance between the proposal distribution and the target distribution, the closer the two distributions are and the larger the ESS is. That is why the χ^2 distance is of particular interest to measure the efficiency of importance sampling algorithms. The following theorem compares the χ^2 distance between the proposal distribution and the target distribution for IS-1 and IS-2.

Theorem 1. *The χ^2 distance between the proposal distribution and the target distribution for IS-1 is less than or equal to that for IS-2, i.e.,*

$$\mathrm{var}_g\left[\frac{\pi(X)}{g(X)}\right] \leqslant \mathrm{var}_{g^*}\left[\frac{\pi^*(X, Y)}{g^*(X, Y)}\right]. \tag{13}$$

Proof. From (10), we have

$$\begin{aligned} 1 + \mathrm{var}_{g^*}\left[\frac{\pi^*(X, Y)}{g^*(X, Y)}\right] &= E_{g^*}\left[\frac{\pi^*(X, Y)}{g^*(X, Y)}\right]^2 = \int_\Omega \int_0^{\pi(x)/Mg(x)} M^2 g(x)\,\mathrm{d}y\,\mathrm{d}x \\ &= \int_\Omega M\pi(x)\,\mathrm{d}x = M. \end{aligned} \tag{14}$$

Since $\pi(x)/g(x) \leqslant M$,

$$1 + \mathrm{var}_g\left[\frac{\pi(X)}{g(X)}\right] = \int_\Omega \frac{\pi^2(x)}{g(x)}\,\mathrm{d}x \leqslant \int_\Omega M\pi(x)\,\mathrm{d}x = M. \tag{15}$$

The theorem follows immediately from (14) and (15). □

Theorem 1 is a special case of the general marginalization principle (i.e., Rao-Blackwellization) discussed in Liu (2001, p. 37). Notice that the marginal distributions of X for $\pi^*(x, y)$ and $g^*(x, y)$ are $\pi(x)$ and $g(x)$, respectively. Therefore conducting importance sampling on the marginal distribution $\pi(x)$ is more efficient than on the joint distribution $\pi^*(x, y)$. Another direct conclusion from Theorem 1 is that based on the same number of samples, IS-1 has a larger effective sample size than IS-2. A discrete version of this conclusion is given in Liu (1996). Other comparisons of rejection sampling and importance sampling can be found in Robert and Casella (1999, pp. 92–96).

If $g(x)$ is an envelope function for rejection sampling, then we can use it as the proposal distribution for importance sampling. Theorem 1 implies that the resulting importance sampling algorithm IS-1 tends to be more efficient than rejection sampling with envelope function $g(x)$, which is equivalent to IS-2, because the variance of the importance weights is smaller for IS-1. IS-2 also requires sampling both X and Y, while IS-1 only need samples of X.

4. Rejection control as an importance sampling scheme

Rejection sampling requires a density function satisfying the envelope property (2). When such an envelope function is difficult to find, and we only have a proposal distribution $q(x)$ which may not satisfy the envelope property, a rejection control algorithm is proposed by Liu (2001, p. 44) to adjust the samples from $q(x)$. Rejection control is potentially useful if the evaluation of $h(x)$ is expensive. It has also been used to improve the efficiency of sequential importance sampling algorithms (Liu et al., 1998). Suppose we have samples $X_1, \ldots, X_N$ from $q(x)$. For any $c > 0$, the rejection control algorithm can be implemented as follows.

Rejection control

(a) Draw U_i from a uniform distribution on $[0, 1]$, and accept X_i if $U_i \leqslant r_i \equiv \min\{1, \pi(X_i)/cq(X_i)\}$, $i = 1, \ldots, N$.
(b) If X_i is accepted, update its weight from $\pi(X_i)/q(X_i)$ to $\tilde{c}\pi(X_i)/r_i q(X_i)$, where

$$\tilde{c} = \int_\Omega \min\left\{1, \frac{\pi(x)}{cq(x)}\right\} q(x)\,\mathrm{d}x.$$

In the algorithm, $\tilde{c}$ is a normalizing constant which can be ignored when estimating μ.

Following the same argument as in Section 2, it is easy to show that the above rejection control algorithm is equivalent to an importance sampling algorithm with

target distribution

$$\tilde{\pi}(x,y)=\begin{cases} cq(x)\max\left\{1,\dfrac{\pi(x)}{cq(x)}\right\} & \text{for } x\in\Omega,\ y\in[0,\min\{1,\pi(x)/cq(x)\}],\\ 0 & \text{otherwise.}\end{cases} \tag{16}$$

and proposal distribution

$$\tilde{q}(x,y)=q(x),\quad \text{for } (x,y)\in\Omega^*. \tag{17}$$

We refer to this importance sampling algorithm as IS-4, and refer to the original importance sampling algorithm with target distribution $\pi(x)$ and proposal distribution $q(x)$ as IS-3.

Liu et al. (1998) show that the accepted samples from rejection control follows a distribution closer to $\pi(x)$ in terms of the χ^2 distance. However, if all samples from rejection control are taken into account, including both accepted samples and rejected samples, then rejection control is equivalent to IS-4, and the following theorem shows that the χ^2 distance between the proposal distribution and the target distribution for IS-4 is greater than or equal to that for IS-3.

Theorem 2. *Assume* $\mathrm{var}_{\tilde{q}}[\tilde{\pi}(X,Y)/\tilde{q}(X,Y)]<\infty$. *Then the* χ^2 *distance between the proposal distribution and the target distribution for IS-3 is less than or equal to that for IS-4, i.e.,*

$$\mathrm{var}_q\left[\frac{\pi(X)}{q(X)}\right]\leqslant\mathrm{var}_{\tilde{q}}\left[\frac{\tilde{\pi}(X,Y)}{\tilde{q}(X,Y)}\right]. \tag{18}$$

Proof. Because

$$\begin{aligned}1+\mathrm{var}_{\tilde{q}}\left[\frac{\tilde{\pi}(X,Y)}{\tilde{q}(X,Y)}\right] &= \int_\Omega\int_0^{\min\{1,\pi(x)/cq(x)\}}\frac{[cq(x)\max\{1,\pi(x)/cq(x)\}]^2}{q(x)}\,\mathrm{d}x\\ &= \int_\Omega\frac{[cq(x)\max\{1,\pi(x)/cq(x)\}]cq(x)\pi(x)/cq(x)}{q(x)}\,\mathrm{d}x\\ &= \int_\Omega\frac{\max\{cq(x),\pi(x)\}\pi(x)}{q(x)}\,\mathrm{d}x\\ &\geqslant \int_\Omega\frac{\pi(x)^2}{q(x)}\,\mathrm{d}x=1+\mathrm{var}_q\left[\frac{\pi(X)}{q(X)}\right].\end{aligned}$$

The theorem is thus proved. □

5. Discussion

Rejection sampling can be viewed as a special importance sampling algorithm. We show that a density function g satisfying the envelope property can be used as a proposal distribution for importance sampling, and the resulting importance sampling algorithm tends to be more efficient than rejection sampling with envelope function g. The derivations in previous sections still hold if the target distribution $\pi(x)$ is only known up to a normalizing constant.

The envelope function g is just one potential choice of the proposal distribution for importance sampling. More efficient proposal distributions that do not have to be envelope functions may be designed. On the other hand, the envelope property is a necessary condition for rejection

sampling. In this sense, importance sampling is more flexible and more promising than rejection sampling for the goal of estimating integral (1). However, if having iid samples from the target distribution is the goal, instead of estimating μ, then rejection sampling is needed.

References

Kong, A., Liu, J.S., Wong, W.H., 1994. Sequential imputations and Bayesian missing data problems. J. Amer. Statist. Assoc. 89, 278–288.

Liu, J.S., 1996. Metropolized independent sampling with comparisons to rejection sampling and importance sampling. Statist. Comput. 6, 113–119.

Liu, J.S., 2001. Monte Carlo Strategies for Scientific Computing. Springer, New York.

Liu, J.S., Chen, R., Wong, W.H., 1998. Rejection control and sequential importance sampling. J. Amer. Statist. Assoc. 93, 1022–1031.

Marshall, A.W., 1956. The use of multi-stage sampling schemes in Monte Carlo computations. In: Meyer, M.A. (Ed.), Symposium on Monte Carlo Methods. Wiley, New York, pp. 123–140.

Robert, C.P., Casella, G., 1999. Monte Carlo Statistical Methods. Springer, New York.

von Neumann, J., 1951. Various techniques used in connection with random digits. Nat. Bureau Stand. Appl. Math. Ser. 12, 36–38.