Note on Rejection sampling and exact sampling with the Metropolised Independence Sampler

Iain Murray
Gatsby Computational Neuroscience Unit
University College London, London WC1N 3AR, UK
http://www.gatsby.ucl.ac.uk/
i.murray@gatsby.ucl.ac.uk

V0.2, 5 August 2004, based on a tea-time talk 13 January 2004.

1 Introduction

This short note shows a close relationship between standard rejection sampling and exact sampling by coupling from the past applied to a Metropolised independence sampler. Little background is assumed, but [1] provides a clear review of all required material. I now know that this idea, first presented as a ten-minute tea-time talk, is probably a duplicate of an unavailable work [3], and is closely related to a paper by Jun S. Liu [2], who provides a *much* more detailed analysis. Perhaps this exposition will be of interest to some readers.

2 Rejection sampling

Rejection sampling [4] is a method to draw independent samples from a probability distribution $P(x) = P^*(x)/Z_P$. We may not know the normalising constant Z_P, but we assume that we can evaluate $P^*(x)$ at any position x we choose. It does not matter here if the function $P(x)$ gives probabilities for discrete x or describes a probability density function over continuous x.

Firstly we choose a distribution $Q(x) = Q^*(x)/Z_Q$ from which we can easily draw independent samples and evaluate $Q^*(x)$ at any x. We then find a constant c for which $cQ^*(x) \geq P^*(x)\ \forall x$. We try and choose c as small as we can, but this step may not be easy. Define the smallest possible choice (possibly unknown) to be c_{opt}. That we can tractably find a valid finite c at all is an assumption of this method. This setup is illustrated in figure 1.

Now we draw samples from our tractable distribution $Q(x)$ instead of our target distribution $P(x)$. At each location $x_i \sim Q(x)$ we evaluate $cQ^*(x_i)$ and draw a random height $h_i \sim \text{uniform}[0, cQ^*(x_i)]$, giving a random position drawn uniformly from under the curve $cQ^*(x)$. Then if $h_i < P^*(x_i)$, we accept x_i as a sample from $P(x)$, otherwise we reject the sample, throwing away both x_i and h_i. We end up with points drawn

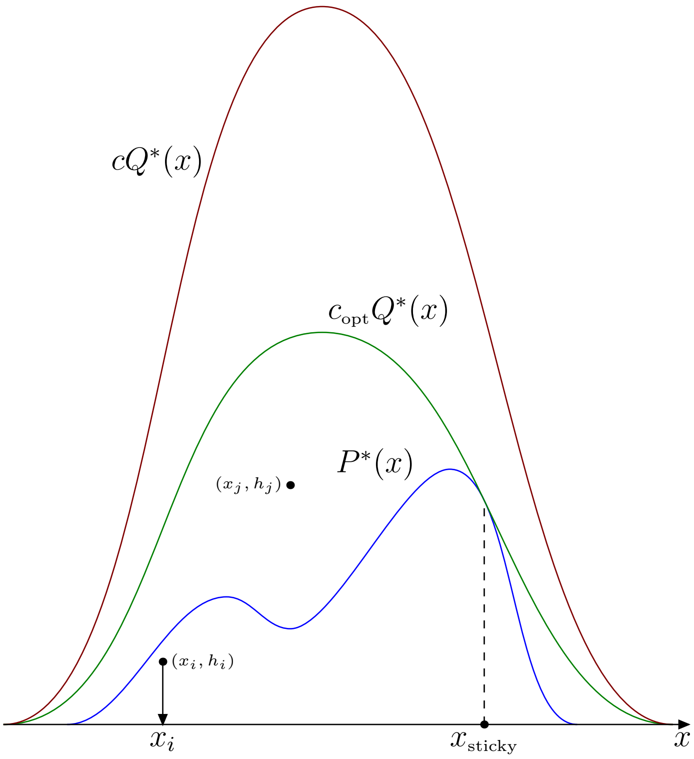

Figure 1: Rejection sampling draws independent samples from $P(x) \propto P^*(x)$ by taking the x values of points drawn uniformly from underneath the curve $P^*(x)$. This is achieved by drawing samples uniformly from underneath the curve $cQ^*(x) \propto Q(x)$ and ignoring any points lying above $P^*(x)$. If points (x_i, h_i) and (x_j, h_j) were drawn, x_i would be added to the list of samples from $P(x)$, while h_i, x_j and h_j are discarded. The method is most efficient for the $c = c_{\text{opt}}$ that minimises the area in which points are rejected. The method breaks if $c < c_{\text{opt}}$, as not all of the area under $P^*(x)$ would be sampled. The role of x_{sticky} is discussed in section 4.

uniformly from the area under the curve $P^*(x)$, the accepted values are therefore samples from $P(x)$. Figure 1 shows two points drawn from under $cQ^*(x)$ leading to one sample from $P(x)$.

The probability that a point is accepted is simply the ratio of the areas underneath $P^*(x)$ and $cQ^*(x)$, as a function of c this probability is:

$$p_{\text{accept}}(c) = \frac{Z_P}{cZ_Q}. \tag{1}$$

Thus the probability distribution over the number of samples from Q required for a single sample from P will be a geometric distribution with mean cZ_Q/Z_P.

3 Metropolised independence sampler

The Metropolis-Hastings method [5] is a Markov chain Monte Carlo (MCMC) method. That is, it sets up a biased random walk forming a Markov chain with a unique equilibrium distribution equal to the target density $P(x)$. On its own, this method only allows us to draw correlated samples, not independent samples as in the previous section.

We must start our Markov chain at some arbitrary position x_0. The procedure for constructing positions at future times, $x_1, x_2, \ldots x_T$, is:

```
1. for $t = 1 \ldots T$
2.     Propose $x' \sim Q(x; x_{t-1})$
3.     Compute $a = \frac{P^*(x')Q^*(x_{t-1};x')}{P^*(x_{t-1})Q^*(x';x_{t-1}})$
4.     Draw $r \sim$ uniform$[0,1]$. If $r < a$ set $x_t = x'$ otherwise set $x_t = x_{t-1}$.
5. end for
```

Ideally x_0 would be drawn from $P(x)$, as otherwise not only are $\{x_t\}$ correlated, they are also biased by the initial condition. However, one of the assumptions above is that direct sampling from $P(x)$ is not easy.

The Metropolised independence sampler is the special case $Q(x; x_{t-1}) \equiv Q(x)$; the proposal distribution is *independent* of the position of the Markov chain. Remember that the samples are not independent, they are still from a Markov chain.

It seems that the two methods above have very different properties. This note explores their similarities.

4 Sampling by coupling from the past

There were two unsolved problems in the previous section: $x_1 \ldots x_T$ are correlated and biased by x_0. The bias can be reduced by running the Metropolis-Hastings iterations many times before starting to record the samples. This "burn-in" period is designed to forget the initial condition so that the samples are approximately uncorrelated from x_0. Similarly inter-sample correlations can be limited by taking many steps of the Markov chain between recording each sample. Unfortunately knowing how many steps to take is a difficult problem.

Exact sampling by coupling from the past [6] is a clever practical method for running Markov chains to draw independent samples exactly (with no bias) from a target distribution.

We assume that many Metropolised independence sampler Markov chains starting from all possible x were started at time $t = -\infty$. Also each chain used the same supply of random numbers to draw x' and r at each step. Any chains that accepted in stage 4. at time $-\tau$ became identical for all later times $t > -\tau$. Those chains moved to the same $x_{-\tau+1}$ and made the same proposals and acceptances from that time onwards. At any time there was a small but finite probability that all chains coalesced; after infinite time at $t = 0$ all of the chains must be on top of one another. Therefore all chains end at a single x_0, totally independent of any initial condition. Our task is simply to identify x_0 without performing the infinite amount of computation implied by the above description.

A key observation is that the acceptance ratio for the Metropolis independence sampler factors into

$$a = \left(\frac{P^*(x')}{Q^*(x')}\right)\left(\frac{Q^*(x_{t-1})}{P^*(x_{t-1})}\right) \propto \left(\frac{Q^*(x_{t-1})}{P^*(x_{t-1})}\right). \tag{2}$$

The first ratio is a constant for all chains, as they all propose the same new point x'. This means that the chain with the smallest acceptance probability, $\min(1, a)$, is always at the position minimising $\frac{Q^*(x)}{P^*(x)}$, independent of the proposed position x'. Chains in this position do not like to move, so we call this location, which may be a set of points for all that follows, x_{sticky} (see figure 1).

Amazingly we can find x_0 by looking back no further than a time $t = -\tau_c$, when a chain with $x_{-\tau_c-1} = x_{\text{sticky}}$ accepted a proposal. As a chain at x_{sticky} has the smallest possible value of a, chains at any other position will also accept the proposal in step 4. We now know that all the coupled chains that started at $t = -\infty$ satisfy $x_{-\tau_c} = x'$. As discussed above the chains will now all follow the same path and x_0 can be identified by following a single Markov chain from $t = -\tau_c$ to $t = 0$.

A simple algorithm to find τ_c is to step back in time, $t = 0, -1, -2, \ldots$ sampling and storing proposals $x'(t)$ and acceptance random numbers $r(t)$. At each time we also evaluate and store $Q^*(x'(t))$ and $P^*(x'(t))$. We stop at $t = -\tau_c$ when all chains would accept the proposal $x'(-\tau)$, which we check by computing if x_{sticky}

would accept. We can then follow the Markov chain starting at position $x'(-\tau)$ and time $t = -\tau + 1$ until time $t = 0$ to identify x_0.

Each step of the algorithm above has the same joint distribution over the proposed point x', and the event $\mathcal{C}$, that a chain at x_{sticky} (and therefore all chains) would accept this proposal:

$$\begin{aligned} P(x', \mathcal{C}) &= P(\mathcal{C}|x') \times Q(x') \\ &= \frac{Q^*(x_{\text{sticky}})}{P^*(x_{\text{sticky}})} \frac{P^*(x')}{Q^*(x')} \times Q(x') \\ &= \frac{1}{c_{\text{opt}}} P^*(x') \frac{1}{Z_Q}. \end{aligned} \tag{3}$$

We can marginalise this expression to obtain the probability of a coalescence event:

$$\begin{aligned} P(\mathcal{C}) &= \frac{1}{c_{\text{opt}} Z_Q} \int \mathrm{d}x' P^*(x') = \frac{Z_P}{c_{\text{opt}} Z_Q} \\ &= p_{\text{accept}}(c_{\text{opt}}). \end{aligned} \tag{4}$$

Thus the probability distribution over τ_c will be a geometric distribution with mean $c_{\text{opt}} Z_Q / Z_P$. The same distribution as found over the number of steps required in rejection sampling for $c = c_{\text{opt}}$ (equation 1).

This section assumed we knew x_{sticky}, which amounts to knowing c_{opt}. In fact if we only knew a suboptimal (but valid) c we could still run the above algorithm: using the bound $\frac{Q^*(x_{\text{sticky}})}{P^*(x_{\text{sticky}})} \geq 1/c$. This would allow us to check that x_{sticky} had accepted a proposal, without actually knowing its location. This enables us to identify some coalescence events, although no longer guarantees finding the first. Again the distribution over the number of iterations required for exact sampling is the same as rejection sampling for the same choice of c.

Moreover, this method and standard rejection sampling need the same number of function evaluations and random numbers, although the algorithm described for coupling from the past requires more memory. Sampling τ_c directly from its geometric distribution and then conditioning proposals on $t = -\tau_c$ being the last detectable coalescence event before $t = 0$ would remove the extra storage requirements, making the two methods even more comparable.

5 Discussion

The author found it surprising that two traditionally very different methods, rejection sampling and Markov chain Monte Carlo, could be made to have the same probability distributions over the number of function evaluations and random numbers required for drawing perfect independent samples. The performance of both exact sampling methods hinges on the ratio of P^* and Q^*, or our knowledge of this ratio, at the single point x_{sticky}. It seems likely that the unavailable preprint [3] mentioned in the exact sampling bibliography [7], made these observations before.

There are two main reasons that we would not use the coupling from the past algorithm described here. Firstly it is just more cumbersome to think about than simple rejection sampling. Secondly, we may not really want exact samples. Note that the Metropolised independence sampler itself, section 3, does not require any consideration of x_{sticky} or valid c. However, the analysis of section 4 indicates that its samples are typically only correlated on length scales of $\approx c_{\text{opt}} Z_Q / Z_P$, regardless of whether we know it or not. While the exact samplers throw away all of the computation leading up to an independent sample, it is perfectly acceptable to use all of the correlated samples obtained from a MCMC sampler when approximating an expectation[1]. Doing so will provide unbiased estimators with a smaller variance than using the smaller number of independent samples obtained from rejection sampling in the same time. While previously unknown to me, this comparison of rejection sampling and Metropolis-Hastings is not new and was made quantitatively with a detailed eigen-analysis and consideration of forward coupling times in [2].

Acknowledgements

Thanks to Zoubin Ghahramani for useful comments.

References

[1] David J.C. MacKay. *Information Theory, Inference, and Learning Algorithms.* Cambridge University Press, 2003. available from http://www.inference.phy.cam.ac.uk/mackay/itila/.

[2] Jun S. Liu. Metropolized independent sampling. *Statistics and Computing*, 6:113–119, 1996.

[3] Haiyan Cai. A note on an exact sampling algorithm and Metropolis Markov chains, 1997. Preprint.

[4] John von Neumann. *Various techniques used in connection with random digits, in John von Neumann, Collected Works*, volume V. Oxford, 1963.

[5] W.K. Hastings. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57(1):97–109, April 1970.

[6] James G. Propp and David B. Wilson. Exact sampling with coupled Markov chains and applications to statistical mechanics. *Random Structures and Algorithms*, 9(1&2):223–252, 1996.

[7] David B. Wilson. Annotated bibliography of perfectly random sampling with Markov chains, 1998–. `http://dbwilson.com/exact/`.

[1] Provided x_0 was drawn exactly, or a suitable burn-in period was discarded.