

How to Combine Fast Heuristic Markov Chain Monte Carlo with Slow Exact Sampling

Antar Bandyopadhyay

(Joint work with David J. Aldous)

University of California, Berkeley
Department of Statistics

Problem :

Let π be a given probability distribution on a set S . Given a function $g : S \rightarrow \mathbb{R}$, the problem is to estimate its mean $\bar{g} := \int_S g(s) \pi(ds)$.

Elementary Statistical Solution :

Generate n *i.i.d* samples from π and take the sample average of the g -values as an estimator for \bar{g} .

- Unbiased estimator,
- Can obtain a rigorous confidence interval for \bar{g} of length $O(n^{-1/2})$.

Drawback of This Solution :

Getting one exact sample from π may be prohibitively slow (in the sense of computational time).

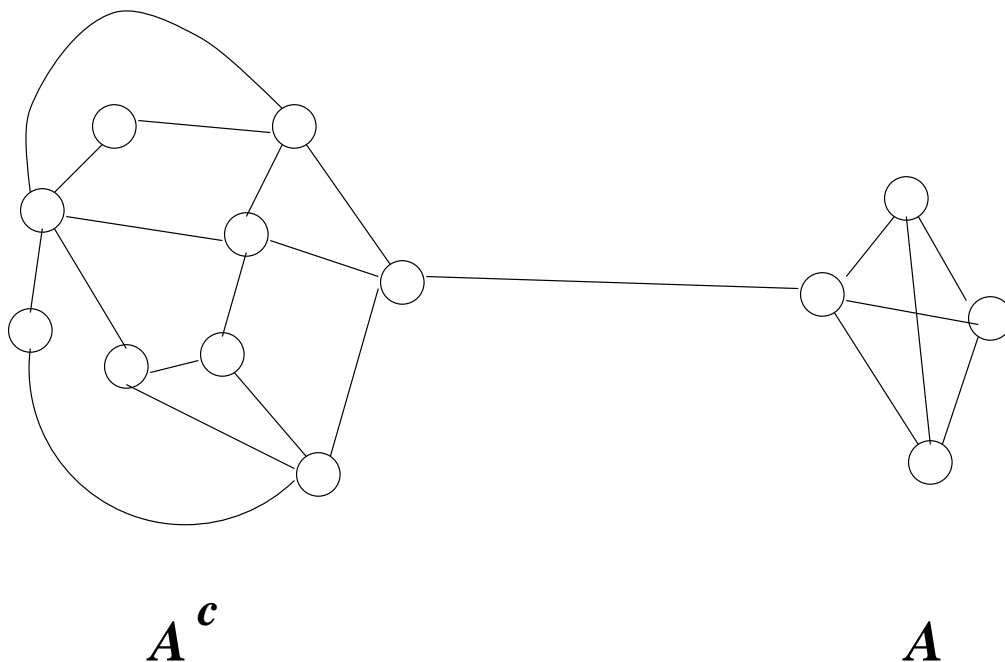
Markov Chain Monte Carlo Method :

Markov Chain Monte Carlo (MCMC) Method provides a solution for such a problem :

- Design a Markov chain with state-space S and stationary law π .
- Sample average of the g -values over a “*long*” run of the chain is a *heuristic* estimator of \bar{g} .

Problem with such Estimator :

In general one can not make the heuristic estimator rigorous, because one can not eliminate the possibility of having a “*bottle neck*”, that is to have a small part of the state-space say a sub-set A *almost* disconnected from the rest.



Typically rigorous estimates require an *a priori* bound on some notion of the chain's *mixing time* (e.g. the *relaxation time*).

Interface between Rigor and Heuristics :

Suppose we have a guess $\hat{\tau}$ for the mixing time of the chain. Suppose also that we have a scheme of generating exact samples from π . We define

$$\rho = \frac{\text{Cost of one exact sample}}{\text{Cost of } \hat{\tau}\text{-steps of the chain}}.$$

$\rho < 1$: One would just carry on with exact samples.

ρ is very large : One can not even get one exact sample, and so forced to rely only on MCMC.

$\rho \gg 1$, but not very large : This is the case of our interest. Here we can get some exact independent samples from π and other dependent samples by MCMC.

Example :

Generating samples from a d -dimensional density using say *acceptance-rejection method*, when d is moderately large.

In such a case we do the following which we call as our “*procedure*” :

- Use the exact sampler to get n independent samples from π . [Notice that using these samples one can construct a confidence interval for \bar{g} of length $O(n^{-1/2})$.]
- Use these n independent samples as initial states and generate n independent m -step realizations of the Markov chain.
- Take the overall average of the g -values as an estimator for \bar{g} .

Heuristics :

Since $\hat{\tau}$ is our guess for the *mixing time* so heuristically the “*effective sample size*” is $\left(n \times \frac{m}{\hat{\tau}}\right)$. So the heuristic estimate of error should be $O\left(\sqrt{\frac{\hat{\tau}}{nm}}\right)$.

Mathematical Assumptions :

1. We assume that the set S is finite.
2. We also assume that the function $g : S \rightarrow \mathbb{R}$ satisfies $0 \leq g(\cdot) \leq 1$.
3. The Markov chain is reversible, that is the transition matrix say $K = ((k_{ij}))$ satisfies

$$\pi_i k_{ij} = \pi_j k_{ji} \quad \forall i, j.$$

Definition of Relaxation Time :

Let $1 = \lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_s \geq -1$ be the eigen-values of the matrix K , then $\tau_2 := (1 - \lambda_2)^{-1}$ is called the *relaxation time* for the reversible K -chain.

Remarks :

We observe that there are two main “obstacles” to sharp estimation.

- (a) If g_2 be the eigen-vector for the eigen-value λ_2 , then for $g = g_2$ an easy computation shows that the variance is $O\left(\frac{\tau_2}{nm}\right)$; so we can not hope to have smaller estimation error.
- (b) Suppose we have a *bottle neck* and the sub-set A has $\pi(A) = O(1/n)$. In that case,
- It is not unlikely to have all the n exact samples in A^c and hence all the sub-sequent realizations of the chains are also in A^c .
 - So the contribution $\mathbf{E}[g(\cdot)I_A]$ to \bar{g} is invisible to our simulation.

Notice that under our second assumption this intrinsic error is bounded by $O(1/n)$.

So the best one could hope is to get a confidence interval of length of order

$$\max\left(\frac{1}{n}, \sqrt{\frac{\tau_2}{nm}}\right).$$

Theorem 1 Fix $n, m \geq 1$ and $0 < \alpha < 1$. Based on $2n$ independent exact samples from π and $2n$ independent m -step realizations of the K -chain, we can construct an interval I such that

$$\mathbf{P}(\bar{g} \notin I) \leq \alpha,$$

and

$$\begin{aligned} \mathbf{P} \left[\text{Length}(I) > \frac{2}{n} \left(\sqrt{\frac{2}{\alpha}} \log n + \log \frac{4}{\alpha} \right) \right] \\ \leq 5n \exp \left[- \min \left(\frac{1}{8}, \frac{m}{48n\tau_2} \right) \log^2 n \right]. \end{aligned}$$

The Upshot :

From the **Theorem** we observe that

- We have a procedure to get a confidence interval for \bar{g} which is always of correct level.
- Further the length of the interval depends on the data, but if we take $m = O(n\hat{\tau})$ and in case our guess $\hat{\tau} \leq \tau_2$, then with high probability we have length $O\left(\frac{\log n}{n}\right)$, which is a significant improvement on $O\left(\frac{1}{\sqrt{n}}\right)$ -length interval.

Construction of the Confidence Interval :

Key Steps:

S1 : Perform our “procedure” of simulating n realizations of m -steps of the Markov chain, starting from n independent exact samples. Output the overall average of g -values as \bar{g}^* , this is our initial guess for \bar{g} .

S2 : Perform the same “procedure” once more independently, and let A_i be the average of g -values over i^{th} m -step realizations of the chain.

S3 : Truncate A_i 's at $\bar{g}^* \pm \frac{\log n}{\sqrt{n}}$ to get \tilde{A}_i , for $1 \leq i \leq n$.

$$\text{Put } \tilde{A} := \frac{1}{n} \sum_{i=1}^n \tilde{A}_i.$$

S4 : Let $N_n = \sum_{i=1}^n I(A_i \neq \tilde{A}_i)$ be the number of truncations. Define $h : (\mathbb{N} \cup \{0\}) \times \mathbb{N} \times (0, 1) \rightarrow [0, \infty)$ as

$$h(z, n; \alpha) := \begin{cases} \frac{z}{n} + \frac{c_\alpha}{\sqrt{n}} & \text{if } z \neq 0 \\ \frac{d_\alpha}{n} & \text{if } z = 0 \end{cases},$$

where $c_\alpha := \frac{1}{\sqrt{2\alpha}}$ and $d_\alpha := \log \frac{2}{\alpha}$.

S5 : Report confidence interval for \bar{g} as

$$I := \tilde{A} \pm \left(\sqrt{\frac{2}{\alpha}} \frac{\log n}{n} + h(N_n, n; \alpha/2) \right).$$

Note :

If $N_n = 0$, that is there is no truncation done then we report a “short” interval of length $O(\frac{\log n}{n})$, otherwise we get a “long” interval of length $O(\frac{1}{\sqrt{n}})$.

Proof of the Theorem :

For proving the **Theorem** we need the following two results :

Proposition 1 For any $b > 0$

$$\mathbf{P} \left[\left| \tilde{A} - \bar{g} \right| > b \frac{\log n}{n} + h(N_n, n; \alpha) \right] \leq \frac{1}{b^2} + \alpha.$$

Proposition 2

$$\mathbf{P}(N_n > 0) \leq 5n \exp \left(- \min\left(\frac{1}{8}, \frac{m}{48n\tau_2}\right) \log^2 n \right).$$

Replacing α by $\frac{\alpha}{2}$ and setting $b = \sqrt{\frac{2}{\alpha}}$ in the first proposition, we get that the interval

$$I = \tilde{A} \pm \left(\sqrt{\frac{2}{\alpha}} \frac{\log n}{n} + h(N_n, n; \alpha/2) \right)$$

satisfies the first requirement that $\mathbf{P}(\bar{g} \notin I) \leq \alpha$.

Further notice that if $N_n = 0$ then

$$\text{Length}(I) = \frac{2}{n} \left(\sqrt{\frac{2}{\alpha}} \frac{\log n}{n} + d_{\alpha/2} \right).$$

So the proof then follows from the second proposition. ■

Proposition 1 For any $b > 0$

$$\mathbf{P} \left[\left| \tilde{A} - \bar{g} \right| > b \frac{\log n}{n} + h(N_n, n; \alpha) \right] \leq \frac{1}{b^2} + \alpha.$$

Proof :

Notice that after observing that the events $[|A_i - \bar{g}^*| \leq \frac{\log n}{\sqrt{n}}]$ happening n times out of n , that is $N_n = 0$, we can be confident that its probability is $1 - O(1/n)$.

Also the truncated variables take values in $\bar{g}^* \pm \frac{\log n}{\sqrt{n}}$, thus \tilde{A} has s.d. of order $\frac{1}{\sqrt{n}} \times \frac{\log n}{\sqrt{n}} = \frac{\log n}{n}$.



Proposition 2

$$\mathbf{P}(N_n > 0) \leq 5n \exp\left(-\min\left(\frac{1}{8}, \frac{m}{48n\tau_2}\right) \log^2 n\right).$$

Proof :

Notice that

$$\mathbf{P}(N_n > 0) \leq n\mathbf{P}\left(|A_1 - \bar{g}^*| > \frac{\log n}{\sqrt{n}}\right).$$

To complete the proof we use the following *large deviation* estimates :

- (Alon & Spencer, 1992)

$$\mathbf{P}\left(|\bar{g}^* - \bar{g}| > \frac{u}{\sqrt{n}}\right) \leq 2e^{-u^2/2} \quad \forall u > 0;$$

- (Lezaud, 1998)

$$\mathbf{P}(|A_1 - \bar{g}| > \lambda) \leq 3 \exp\left[-\frac{\lambda^2 m}{12\tau_2}\right] \quad \forall \lambda > 0.$$

■

References :

[1] D.J. Aldous and J.A. Fill. Reversible Markov Chains and Random Walks on Graphs. Book in preparation, 2001.

[2] N. Alon and J.H. Spencer. *The Probabilistic Method*. Wiley, 1992.

[3] P. Lezaud. Chernoff-type Bound for Finite Markov Chains. *Ann. appl. Probab.*, 8 : 849-867, 1998.

[4] D. Randall and A. Sinclair. Self-testing Algorithms for Self-avoiding Walks. *J. Math. Phys.*, 41 : 1570-1584, 2000.