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

(with Antar Bandyopadhyay).

See `www.stat.berkeley.edu/`

MCMC

Target prob. dist. π on space S.
Seek to sample from π.
Design a Markov transition matrix $P(x, y)$ to have π as its stationary distribution. Use the Markov chain X_t to get samples; naively, take large t_0 and regard

$$X_{t_0}, X_{2t_0}, X_{3t_0}, \ldots$$

as approx independent samples from π.

A fundamental question in both theory and applications is: how large must t_0 be? Much theory deals with definitions and bounding techniques for mixing times; useful mainly in highly-structured "prior model" settings. In real-data applications ("posterior models") one instead relies on simulation diagnostics to assess whether convergence has occurred.

We consider reversible $P(x,y)$:

$$\pi(x)P(x,y) = \pi(y)P(y,x) \quad \forall\ x,y.$$

Matrix P has eigenvalues $1 = \lambda_1 \geq \lambda_2 \geq \cdots$; call $\tau := (1-\lambda_2)^{-1}$ the relaxation time. On finite state-space, $\tau < \infty$ iff irreducible.

Can we eliminate theory by just estimating τ from simulations? One would like a general-purpose diagnostic algorithm which has the following property.

- At each step the algorithm chooses some already-seen state x' and gets a sample from $P(x',\cdot)$.
- Eventually stop; output estimate $\hat{\tau}$ of τ.

Want a property like

$$P(\tau > 10\hat{\tau}) < 1/10.$$

For trivial reasons, no such algorithm can exist.

Issue: Without theoretical bound on mixing time, can one make rigorous inference from MCMC?

In general, no. But what about making side assumptions?

Coupling from the past provides one route, but implementation requires analytic niceness of $P(x, y)$ (e.g. monotonicity) and human tailoring of algorithm to the particular S, P .

We will cheat in a different way. Suppose we have some just-feasible way to sample exactly from π (a **slow exact sampler**). Can we combine this with MCMC for which we lack rigorous mixing time bounds?

Conceptual setting

Target dist. π.

Reversible MC with stationary dist. π. Have a guess $\widehat{\tau}$ for its relaxation time τ, but no rigorous bound on τ.

Separately, have an "exact sampler" for π.

Which to use? Consider the ratio

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

$\boxed{\rho < 1}$: Just use exact samples.

$\boxed{\rho \text{ is very large}}$: Canot afford even one exact sample – forced to rely on MCMC and heuristics.

$\boxed{\rho \gg 1, \text{ but not very large}}$: 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.

Suppose our goal is to estimate $\bar{g} := \int_S g(s)\,\pi(ds)$ for some given function $g : S \to R$. From n exact samples, can give a confidence interval (CI) of length $O(1/\sqrt{n})$. Our result will say (roughly) that from n exact samples and $n^2\hat{\tau}$ MC steps we can get a 95% CI for $\bar{g}$ such that

- The CI is valid without any assumptions on the MC's mixing time
- if the unknown τ is indeed $\leq \hat{\tau}$, or if we're just lucky, then the length of the CI will be $O(\frac{\log n}{n})$.

Here $O(\cdot)$ is shorthand for explicit non-asymptotic bound. Note
(i) the latter accuracy would require about n^2 exact samples; our procedure is cheaper by the factor ρ.
(ii) If we knew $\hat{\tau} = \tau$ then a priori we would have a length $O(1/n)$ CI from the $n^2\hat{\tau}$ MC steps. In this sense, the result makes rigorous the CI one would write down based on heuristics.

Consider the following multiple-runs procedure.

- Use the exact sampler to get n independent samples from π.
- 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 $\bar{g}$.

Our algorithm uses a modification of this procedure. First, let's consider the obstacles to sharp estimation in this context. There are two different obstacles.

(a) If g_2 is the eigenvector for the eigenvalue λ_2, then for $g = g_2$ an easy computation shows that the variance is $O\left(\frac{\tau}{nm}\right)$; so we can not hope to have smaller estimation error.

(b) Suppose state-space almost disconnects into $\{A, A^c\}$ where $\pi(A) \approx 1/n$. In that case,

- It is not unlikely to have all the n exact samples in A^c and hence all the subsequent realizations of the chains are also in A^c.

- So the contribution $E[g(\cdot)I_A]$ to $\bar{g}$ is invisible to our simulation.

By assuming $0 \leq g(\cdot) \leq 1$, we make this contribution be $O(1/n)$.

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

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

Taking $m = n\tau$ makes this as small as possible; in practice take $m = n\hat{\tau}$.

$P(x, y)$ is π-reversible

$0 \leq g(\cdot) \leq 1$.

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 P-chain, we can construct an interval I such that*

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

and

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

Use with m of order $n\hat{\tau}$ to get the result stated informally earlier.

Can also do adaptive version – use

$$m = m_0, 2m_0, 4m_0, 8m_0, 16m_0, \ldots$$

until get short CI or exhaust computational resources.

Construction of the Confidence Interval :

S1 : Perform the "multiple-runs procedure", 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 (if needed) the A_i's at $\bar{g}^* \pm \frac{\log n}{\sqrt{n}}$ to get $\tilde{A}_i$, for $1 \leq i \leq n$. Put $\tilde{\tilde{A}} := \frac{1}{n}\sum_{i=1}^{n} \tilde{A}_i$.

S4 : Report confidence interval for $\bar{g}$ as

$$I := \tilde{\tilde{A}} \pm \left(\sqrt{\frac{2}{\alpha}} \frac{\log n}{n} + \frac{1}{n} \log \frac{2}{\alpha} \right)$$

if no truncation needed; report alternate (longer) CI else.

Why does this work?

(a) s.d. $(\tilde{A}_i) \leq \frac{\log n}{n}$, so
$\text{ave}_i \tilde{A}_i$ has s.d. $\leq \frac{1}{\sqrt{n}} \times \frac{\log n}{n} = \frac{\log n}{n}$.

(b) Observing the event "$A_i \notin \bar{g}^* \pm \frac{\log n}{n}$" occur 0 times out of n gives confidence its probability is $\leq c/n$. This bounds effect of truncation on $\bar{g}$, leads to validity of CI.

(c) Universal large deviation bounds for sample averages of stationary reversible MCs (Gillman 1993; Lezaud 1998) shows $\exists$ universal β such that for $\lambda > 0$

$$P(|A_i - \bar{g}| > \lambda n^{-1/2}) \leq 3 \exp(-\beta \lambda^2 m / n\tau).$$

For $m \approx n\tau$ and $\lambda \approx \log n$ this bound is $\ll 1/n$; truncations not needed.