Adaptive MCMC for Everyone

Jeffrey S. Rosenthal, University of Toronto

jeff@math.toronto.edu

http://probability.ca/jeff/

Estimation from sampling: Monte Carlo

In applications, we often have a complicated, high-dimensional density function $\pi : \mathcal{X} \to [0, \infty)$, for some $\mathcal{X} \subseteq \mathbf{R}^d$ (d large).

(e.g. Bayesian posterior distribution)

Want to compute expected values like:

$$\mathbf{E}_\pi(h) \;:=\; \int_{\mathcal{X}} h(x)\, \pi(x)\, dx\,.$$

If π is complicated, can't use calculus or numerical integration. Instead, can try to sample from π, i.e. generate on a computer

$$X_1, X_2, \ldots, X_M \sim \pi \quad (i.i.d.)\,,$$

then estimate by e.g.

$$\mathbf{E}_\pi(h) \;\approx\; \frac{1}{M} \sum_{i=1}^{M} h(X_i)\,.$$

Good. But how to sample? Often infeasible! Instead . . .

Markov Chain Monte Carlo (MCMC)

Define an ergodic Markov chain (random process) $X_0, X_1, X_2, \ldots$, which converges in distribution to $\pi(\cdot)$. Extremely popular!

Then for "large enough" n, $\mathcal{L}(X_n) \approx \pi(\cdot)$, so

$$\mathbf{E}_\pi(h) \;\approx\; \frac{1}{M} \sum_{i=n+1}^{n+M} h(X_i)\,, \quad \text{etc.}$$

For example ("Metropolis Algorithm"): Given X_{n-1}:

- Propose a new state $Y_n \sim Q(X_{n-1}, \cdot)$, e.g. $Y_n \sim N(X_{n-1}, \Sigma_p)$.
- Let $\alpha = \min\left[1, \; \frac{\pi(Y_n)}{\pi(X_{n-1})}\right]$.
- With probability α, accept the proposal (set $X_n = Y_n$).
- Else, with prob. $1 - \alpha$, reject the proposal (set $X_n = X_{n-1}$).

FACT: α is chosen just right so this Markov chain is reversible with respect to $\pi(\cdot)$, so $\pi(\cdot)$ stationary, and $X_n \to \pi(\cdot)$. [Javascript] (3/8)

Optimising MCMC?

e.g. Metropolis: what is optimal proposal $Q(X_{n-1}, \cdot)$? [Javascript]

There is various theory which specifies the optimal proposal distribution, in terms of properties of π.

(Mostly proven using diffusion limits: Roberts-Gelman-Gilks 1997; Roberts-R. 1998, 2001; Bédard-R. 2007; Atchadé-Roberts-R. 2011; Yang-Roberts-R. 2020; ...)

Great, except we might not know enough about π to use it.

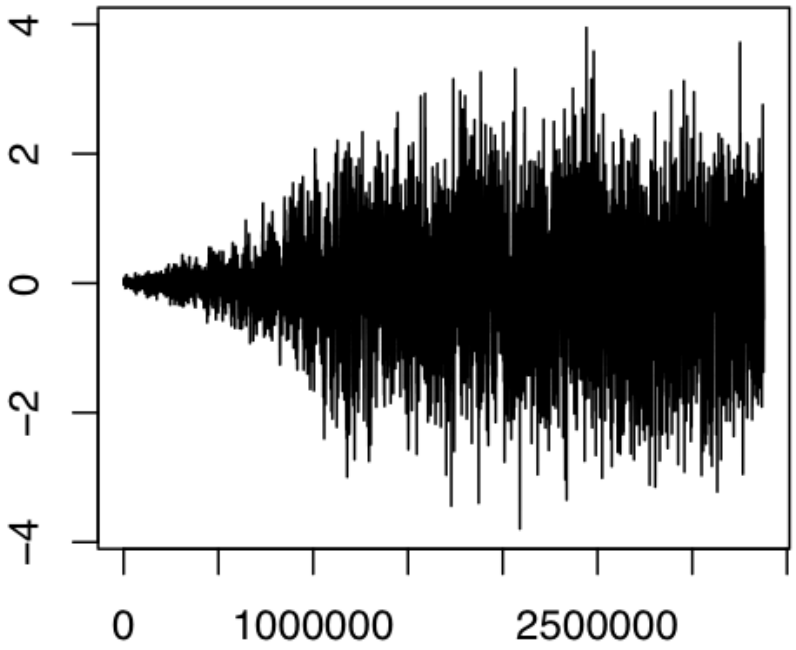
"Chicken and egg". So, let the computer decide, on the fly!

At iteration n, use Markov chain P_{Γ_n}, where $\{P_\gamma\}_{\gamma \in \mathcal{Y}}$ are each valid MCMC, and then $\Gamma_n \in \mathcal{Y}$ chosen according to some adaptive rules (depending on chain's history, etc.). [Javascript]

Can this help us to find better Markov chains? Yes!

(4/8)

Example: High-Dimensional Adaptive Metropolis



In dimension 200, takes over a million iterations, then finally learns a good proposal distribution and starts mixing well.

Great . . . but is it Ergodic?

No longer Markovian, so maybe not always! [Javascript]

Theorem [Roberts and R., J.A.P. 2007]. Ergodic if it satisfies:

(a) [Diminishing Adaptation] Adapt less and less as the algorithm proceeds. Formally, $\sup_{x \in \mathcal{X}} \|P_{\Gamma_{n+1}}(x, \cdot) - P_{\Gamma_n}(x, \cdot)\| \to 0$ in prob.

(b) [Containment] Times to stationary are bounded in probability as $n \to \infty$. Formally, $\forall \epsilon > 0$, $\{M_\epsilon(X_n, \Gamma_n)\}_{n=0}^{\infty}$ is tight, where $M_\epsilon(x, \gamma) = \inf\{n \geq 1 : \|P_\gamma^n(x, \cdot) - \pi(\cdot)\| < \epsilon\}$ is the time to converge to within ϵ of stationarity.

Enough to prove ergodicity of Adaptive Metropolis, componentwise versions, LLN, etc. Good!

Here (a) can always be made to hold, since adaption is user controlled. But (b) is always a challenge to verify. Instead . . .

Verifying Containment: "For Everyone"

- We proved general theorems about stability of "adversarial" Markov chains under various conditions (Craiu, Gray, Latuszynski, Madras, Roberts, and R., A.A.P. 2015).

- Then we applied them to adaptive MCMC, to get a list of directly-verifiable conditions which guarantee Containment:
  $\Rightarrow$ Never move more than some (big) distance D.
  $\Rightarrow$ Outside (big) rectangle K, use fixed kernel (no adapting).
  $\Rightarrow$ The transition or proposal kernels have continuous densities wrt Lebesgue measure. (or piecewise continuous: Yang & R. 2015)
  $\Rightarrow$ The fixed kernel is bounded above, and below on compact regions for jumps $\leq \delta$, by constants times Lebesgue measure. (Easily verified under continuity assumptions.)

- Can directly ensure these conditions in practice. So, this can be used by applied MCMC users. "Adaptive MCMC for everyone!"

Summary

- MCMC is extremely popular for estimating expectations.
- Adaptive MCMC tries to "learn" how to sample better. Good.
- Works well for high-dimentional Adaptive Metropolis, etc.
- But must be done carefully, or it will destroy stationarity. Bad.
- To converge to $\pi(\cdot)$, suffices to have each P_γ be valid, plus (a) Diminishing Adaptation (important), and (b) Containment (technical condition, usually satisfied, but hard to verify). Good.
- New "adversarial" conditions more easily verify Containment.
- Hopefully can use adaption on many other examples – try it!

All my papers, applets, software: probability.ca/jeff