Discussion: Adaptive MCMC For Everyone

Jeffrey S. Rosenthal
University of Toronto

jeff@math.toronto.edu
http://probability.ca/jeff/

(INFORMS Meeting, Nashville, November 14, 2016)

Introduction and Context

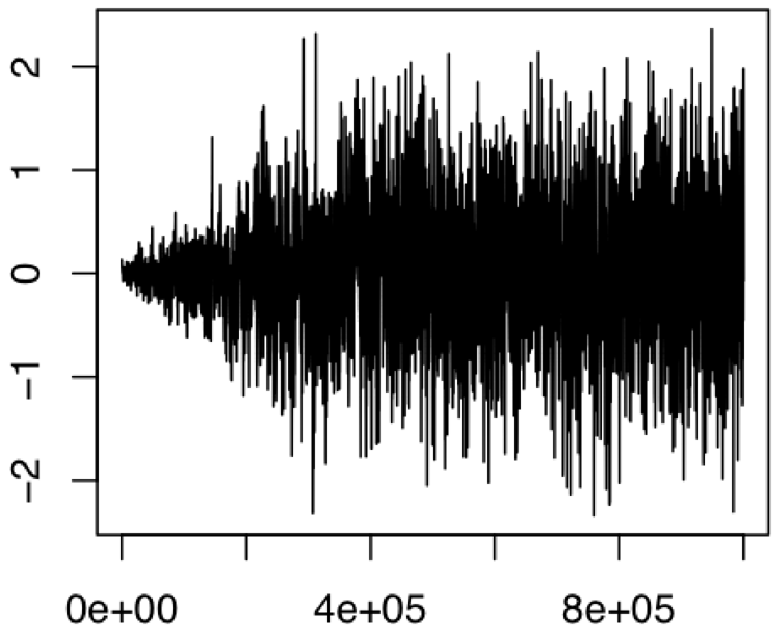
Recall:

- MCMC is really really really important.
- Some MCMC algorithms converge much faster than others.
- Can find optimality results from diffusion limits.
- e.g. Gaussian Random-Walk Metropolis: optimal choice has acceptance rate around 0.234 (how?), and proposal covariance $(2.38)^2 \, d^{-1} \, \Sigma_t$ where Σ_t is the target covariance (unknown).
- So, we have guidance about optimising MCMC in terms of acceptance rate, target covariance matrix Σ_t, etc.
- But we don't know what proposal will lead to a desired acceptance rate, nor how to compute Σ_t.
- What to do? Trial and error? (difficult, especially in high dimension) Or . . .

Adaptive MCMC

- Suppose have a family $\{P_\gamma\}_{\gamma \in \mathcal{Y}}$ of possible Markov chains, each with stationary distribution π.
- How to choose among them?
- Let the computer decide, on the fly!
- At iteration n, use Markov chain P_{Γ_n}, where $\Gamma_n \in \mathcal{Y}$ chosen according to some adaptive rules (depending on history, etc.).
- Simple example: [APPLET]
- e.g. Estimate true target covariance Σ_t by the empirical estimate, Σ_n, based on the observations so far $(X_1, X_2, \ldots, X_n)$.
- Can this help us to find better Markov chains? (Yes!)
- On the other hand, the Markov property, stationarity, etc. are all destroyed by using an adaptive scheme.
- Is the resulting algorithm still ergodic? (Sometimes!)

Example: 100-Dimensional Adaptive Metropolis



Plot of first coord. Takes about 300,000 iterations, then "finds" good proposal covariance and starts mixing well. Good!

- Similarly Adaptive Componentwise Metropolis, Gibbs, etc.

But What About the Theory?

- So, adaptive MCMC seems to work well in practice.

- But will it be ergodic, i.e. converge to π? (Converge at all ... never mind how quickly ...)

- Ordinary MCMC algorithms, with fixed choice γ, are automatically ergodic by standard Markov chain theory (since they're irreducible and aperiodic and leave π stationary). But adaptive algorithms are more subtle, since the Markov property and stationarity are destroyed by using an adaptive scheme.

- e.g. if the adaption of Γ_n is such that P_{Γ_n} usually moves slower when x is in a certain subset $\mathcal{X}_0 \subseteq \mathcal{X}$, then the algorithm will tend to spend much more than $\pi(\mathcal{X}_0)$ of the time inside $\mathcal{X}_0$, even if each update on its own preserves stationarity. [APPLET]

- Some previous results, but they require limiting / hard-to-verify conditions, like bounded state space, or existence of simultaneous geometric drift conditions, or Doeblin condition, or ...

- Need more general, easily-verified theorems ...

One Particular Convergence Theorem

- Theorem [Roberts and R., J.A.P. 2007]: Adaptive MCMC will converge, i.e. $\lim_{n\to\infty} \sup_{A\subseteq\mathcal{X}} \|\mathbf{P}(X_n \in A) - \pi(A)\| = 0$, if:

(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. [Can always be made to hold, since adaption is user controlled.]

(b) [Containment] Times to stationary from X_n, if fix $\gamma = \Gamma_n$, remain bounded in probability as $n \to \infty$. [Technical condition, to avoid "escape to infinity". Holds if e.g. $\mathcal{X}$ and $\mathcal{Y}$ finite, or compact, or ... And always seems to hold in practice.]

(Also guarantees WLLN for bounded functionals. Various other results about LLN / CLT under stronger assumptions.)

Good, but ... Containment condition is a pain.

Can we eliminate it?

What about that "Containment" Condition?

- Recall: adaptive MCMC is ergodic if it satisfied Diminishing Adaptation (easy: user-controlled) and Containment (technical).

- Is Containment just an annoying artifact of the proof? No!

- Theorem (Latuszynski and R., 2014): If an adaptive algorithm does not satisfy Containment, then for all $\epsilon > 0$,

$$\lim_{K\to\infty} \limsup_{n\to\infty} \mathbf{P}(M_\epsilon(X_n, \gamma_n) > K) \ > \ 0\,,$$

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.

That is, an adaptive algorithm without Containment will take arbitrarily large numbers of steps (K) to converge. Bad!

- Conclusion: Yay Containment!?!?

- But how to verify it??

Verifying Containment: "For Everyone"

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

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

- All my papers, applets, software: www.probability.ca