Computational Complexity of Metropolis

- Combining this complexity result with the Metropolis weak convergence results immediately shows that:
  - The speeded-up processes U_t^d converge to π in $O(1)$ time.
- But U_t^d equals the original Metropolis algorithm's first coordinate process $X_{n,1}^d$, sped up by a factor of d.
- Hence, the original Metropolis algorithm's first coordinate process $X_{n,1}^d$ must converge to π in $O(d)$ time.
- Hence, the Metropolis algorithm converges (coordinatewise at least) in time $O(d)$. Right?
- One technicality: we need weak convergence from any starting point X_0, not from stationarity $X_0 \sim \pi$ . . . but that also holds if the powers of the target density f in the moment assumptions are increased slightly (from 8 and 4, to 12 and 6). Phew!
- Also, still requires unrealistic conditions on π . . . but we're working on that. Then have: convergence in $O(d)$ iterations!

How to Use the Optimality Information?

Recall: We have guidance about optimising MCMC in terms of acceptance rate, target covariance matrix Σ_t, etc.

In particular:

1. Want acceptance rate around 0.234.

2. Optimal Gaussian RWM proposal is $N\Big(x,\, (2.38)^2\, d^{-1}\, \Sigma_t\Big)$, where Σ_t is the covariance matrix of the target π.

Great, except . . . we don't know what proposal will lead to a desired acceptance rate. And, we don't know how to compute Σ_t.

So, what to do?

Trial and error? (difficult, especially in high dimension)

Or . . . let the computer decide, on the fly!

Adaptive MCMC

Suppose we have a family $\{P_\gamma\}_{\gamma \in \mathcal{Y}}$ of possible Markov chains, each with stationary distribution π.

Let the computer choose among them!

At iteration n, use Markov chain P_{Γ_n}, where $\Gamma_n \in \mathcal{Y}$ chosen according to some adaptive rules (depending on chain's history, etc.). [APPLET]

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!)

We begin with some simulation examples . . .

Example: High-Dimensional Adaptive Metropolis

Dim $d = 100$, with target π having target covariance Σ_t.
Here Σ_t is 100×100 (i.e., 5,050 distinct entries).

Here optimal Gaussian RWM proposal is $N\Big(x,\, (2.38)^2\, d^{-1}\, \Sigma_t\Big)$.

But usually Σ_t unknown. Instead use empirical estimate, Σ_n, based on the observations so far $(X_1, X_2, \ldots, X_n)$. Then let

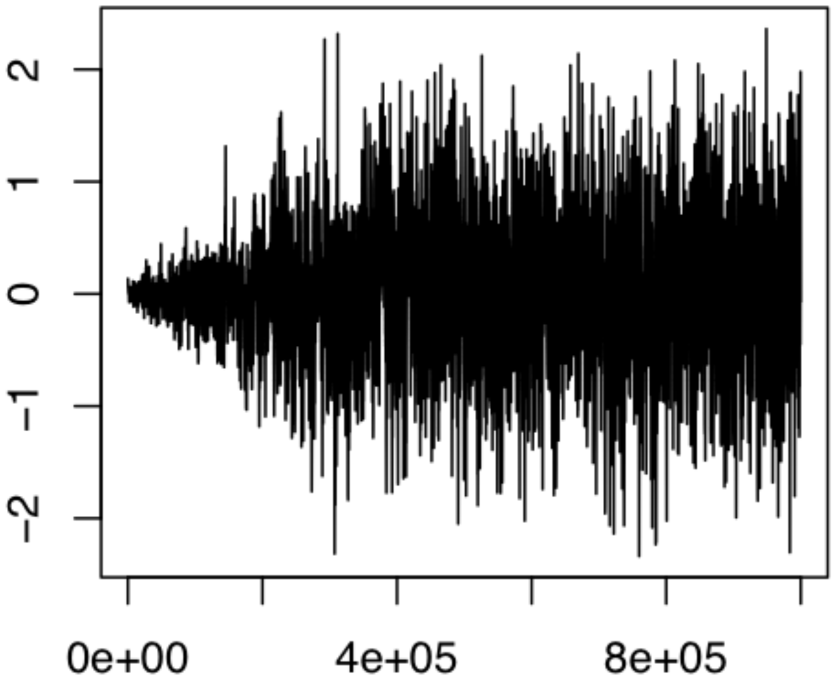
$$Q_n(x, \cdot) \;=\; (1-\beta)\, N\Big(x,\, (2.38)^2\, d^{-1}\, \Sigma_n\Big) + \beta\, N\Big(x,\, (0.1)^2\, d^{-1}\, I_d\Big)\,,$$

where e.g. $\beta = 0.05$.

(Slight variant of the algorithm of Haario et al., Bernoulli 2001.)

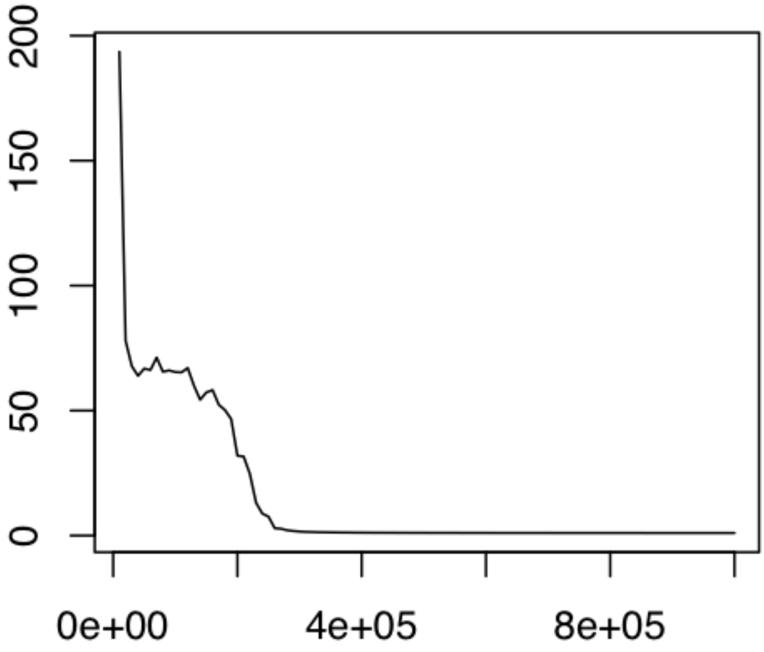
Let's try it . . .

High-Dimensional Adaptive Metropolis (cont'd)



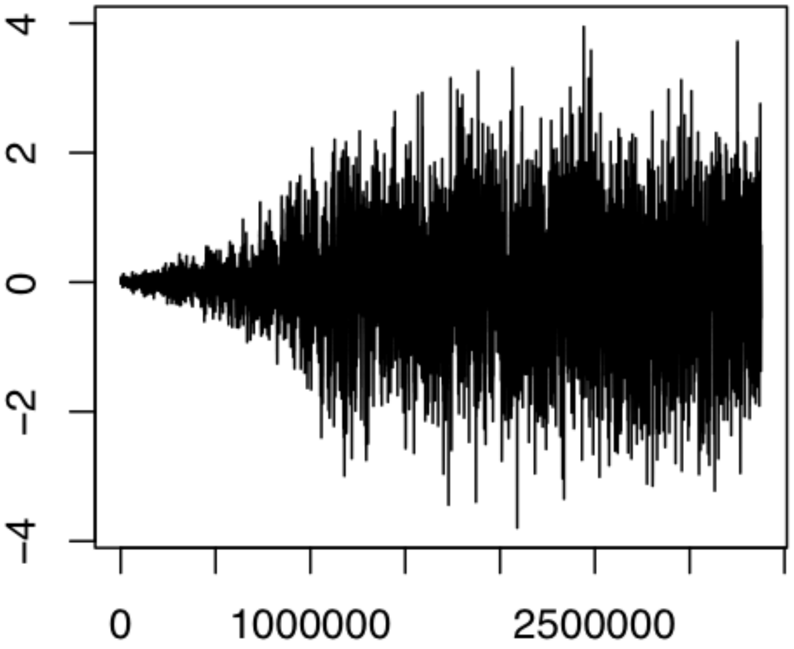
Plot of first coord. Takes about 300,000 iterations, then "finds" good proposal covariance and starts mixing well.

High-Dimensional Adaptive Metropolis (cont'd)



Plot of sub-optimality factor $b_n \equiv d\left(\sum_{i=1}^{d}\lambda_{in}^{-2} / (\sum_{i=1}^{d}\lambda_{in}^{-1})^2\right)$, where $\{\lambda_{in}\}$ eigenvals of $\Sigma_n^{1/2}\,\Sigma^{-1/2}$. Starts large, converges to 1.

Even Higher-Dimensional Adaptative Metropolis



In dimension 200, takes about 2,000,000 iterations, then finds good proposal covariance and starts mixing well.

Another Example: Componentwise Adaptive Metropolis

Propose new value $y_i \sim N(x_i,\, e^{2\, ls_i})$ for the i^{th} coordinate, leaving the other coordinates fixed; then repeat for different i.

Choice of scaling factor ls_i?? (i.e., "$\log(\sigma_i)$")

Recall: optimal one-dim acceptance rate is ≈ 0.44. So:

Start with $ls_i \equiv 0$ (say).

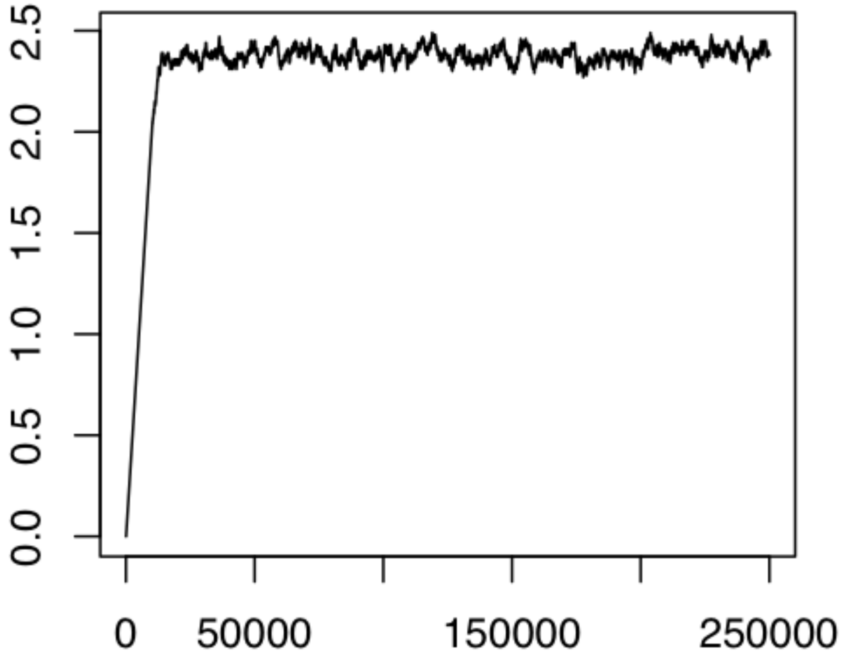
Adapt each ls_i, in batches, to seek 0.44 acceptance rate:

After the j^{th} batch of 100 (say) iterations, decrease each ls_i by $1/j$ if the acceptance rate of the i^{th} coordinate proposals is < 0.44, otherwise increase it by $1/j$.

Let's try it . . .

Adaptive Componentwise Metropolis (cont'd)

Test on Variance Components Model, with $K = 500$ (dim=503), J_i chosen with $5 \le J_i \le 500$, and simulated data $\{Y_{ij}\}$.



Adaption quickly finds "good" values for the ls_i values.

Great ... but is it Ergodic?

Adaptive MCMC seems to work well in practice.

But will it be ergodic, i.e. converge to π?

Ordinary MCMC algorithms, i.e. with fixed choice of 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 the adaptive scheme. [APPLET]

WANT: Simple conditions guaranteeing $\|\mathcal{L}(X_n) - \pi\| \to 0$, where $\|\mathcal{L}(X_n) - \pi\| \equiv \sup_{A \subseteq \mathcal{X}} |\mathbf{P}(X_n \in A) - \pi(A)|$.

(Alternative: Just do "finite adaptation" and diagnose when to stop, e.g. Yang & R., Comp. Stat. 2017; R package "atmcmc".)

One Simple Convergence Theorem

THEOREM [Roberts and R., J.A.P. 2007]: An adaptive scheme using $\{P_\gamma\}_{\gamma\in\mathcal{Y}}$ will converge, i.e. $\lim_{n\to\infty} \|\mathcal{L}(X_n) - \pi\| = 0$, 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 sub-exponential tails, or . . . (Bai, Roberts, and R., Adv. Appl. Stat. 2011). And always seems to hold in practice.]

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

Other results by: Haario, Saksman, Tamminen, Vihola; Andrieu, Moulines, Robert, Fort, Atchadé; Kohn, Giordani, Nott; . . .

Outline of Proof (one page only!)

Define a second chain $\{X'_n\}$, which begins like $\{X_n\}$, but which stops adapting after time N. ("coupling")

Containment says that the (ordinary MCMC) convergence times are bounded, so that for large enough M, we "probably" have $\mathcal{L}(X'_{N+M}) \approx \pi(\cdot)$, i.e. $\mathbf{P}(X'_{N+M} \in A) \approx \pi(A)$ for all A and N.

And, Diminishing Adaptation says that we adapt less and less, so that for large enough N (depending on M),

$$(X_N, X_{N+1}, \ldots, X_{N+M}) \;\approx\; (X'_N, X'_{N+1}, \ldots, X'_{N+M})\,.$$

Combining these, for large enough N and M, we "probably" have

$$\mathcal{L}(X_{N+M}) \;\approx\; \mathcal{L}(X'_{N+M}) \;\approx\; \pi(\cdot)\,, \quad \text{Q.E.D.}$$

Implications of Theorem

Adaptive Metropolis algorithm:

- Empirical estimates satisfy Diminishing Adaptation.

- And, Containment easily guaranteed if we assume π has bounded support (Haario et al., 2001), or sub-exponential tails (Bai, Roberts, and R., 2011).

- COR: Adaptive Metropolis is ergodic under these conditions.

Adaptive Componentwise Metropolis:

- Satisfies Diminishing Adaption, since adjustments $\pm 1/j \to 0$.

- Satisfies Containment under boundedness or tail conditions.

- COR: Ad. Comp. Metr. also ergodic under these conditions.

So, previous adaptive algorithms work (at least asymptotically).

Similar convergence results for: regional adaptation (Craiu, R., C. Yang, JASA 2009), and adaptive multiple-try Metropolis (J. Yang, Levi, Craiu, R., under revision). Good!

Choosing Which Coordinates to Update When

S. Richardson (statistical geneticist): Successfully ran adaptive Componentwise Metropolis algorithm on genetic data with thousands of coordinates. Good!

But many of the coordinates are binary, and usually do not change.

She asked: Do we need to visit every coordinate equally often, or can we gradually "learn" which ones usually don't change and downweight them? Good question – how to proceed?

Suppose at each iteration n, we choose to update coordinate i with probability $\alpha_{n,i}$, and then we update the random-scan coordinate weights $\{\alpha_{n,i}\}$ on the fly.

What conditions ensure ergodicity?

Seemed hard! Then we found a claim [J. Mult. Anal. **97** (2006), p. 2075]: Suffices that $\lim_{n\to\infty} \alpha_{n,i} = \alpha_i^*$, where the Gibbs sampler with fixed weights $\{\alpha_i^*\}$ is ergodic.

Really?? No, counter-example! (K. Latuszyński)

Ergodicity with Adaptive Coordinate Weights

So, we had to be smarter than that!

We proved (Latuszynski, Roberts, and R., Ann. Appl. Prob. 2013) that adaptively weighted samplers are ergodic if either:

(i) some choice of weights $\{\alpha_i^*\}$ make it uniformly ergodic, or

(ii) there is simultaneous inward drift for all the kernels P_γ, i.e. there is $V : \mathcal{X} \to [1, \infty)$ with

$$\limsup_{|x|\to\infty} \sup_{\gamma\in\mathcal{Y}} \frac{(P_\gamma V)(x)}{V(x)} < 1 .$$

Then, by being careful about continuity, boundedness, etc., can guarantee ergodicity in many cases, including for high-dimensional genetics data (Richardson, Bottolo, R., Valencia 2010). Good!

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., J.A.P. 2014): If an adaptive algorithm does not satisfy Containment, then it is "infinitely inefficient": that is, for all $\epsilon > 0$,

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

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. Bad!

Conclusion: Yay Containment!?!

But how to verify it??

A Method For Verifying Containment

(Craiu, Gray, Latuszynski, Madras, Roberts, and R., A.A.P. 2015)

- We first proved general theorems about stability of "adversarial" Markov chains under various conditions.

- Suppose a random process $\{X_n\}$ on $\mathcal{X}$ satisfies:

⇒ We always have $\mathrm{dist}(X_{n+1}, X_n) \leq D$, for some fixed (large) constant $D < \infty$.

⇒ Outside of some fixed (large) bounded subset $K \subseteq \mathcal{X}$, $\{X_n\}$ follows a fixed ergodic Markov transition kernel P.

(But within K, an adversary can make it do anything ...)

⇒ There is a fixed probability measure μ_* on $\mathcal{X}$ with $P(x, dz) \leq M\, \mu_*(dz)$, and $P^{n_0}(x, dz) \geq \epsilon\, \mu_*(dz)$, for $x \in K_{2D} \setminus K$.

THEOREM: Then $\{X_n\}$ is tight, i.e. the sequence $\{\mathrm{dist}(X_n, \mathbf{0})\}_{n=0}^{\infty}$ remains bounded in probability as $n \to \infty$.

Verifying Containment (cont'd)

- We then applied this to adaptive MCMC, to get a list of directly-verifiable conditions which guarantee Containment:

⇒ Never move more than some (big) distance D.

⇒ Outside (big) rectangle K, use fixed kernel (no adapting).

⇒ The transition or proposal kernels have continuous densities wrt Lebesgue measure. (or piecewise continuous: Yang & R. 2015)

⇒ 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 easily used by applied MCMC users.
- "Adaptive MCMC for everyone!"

See also the nice recent "AIR MCMC" approach of Chimisov, Latuszynski, and Roberts, arXiv 2018.

Summary

- MCMC is extremely popular for estimating expectations.
- Basic Markov chain theory establishes convergence.
- Quantitative convergence bounds can sometimes be obtained using coupling with minorisation (and drift) conditions.
- Rescaled MCMC sometimes converges to diffusion limits.
- MCMC can be optimised by maximising the speed.
- Metropolis (with special forms of π) has an explicit maximisation, corresponding to AR $= 0.234$.
- Best proposal covariance is proportional to the target π.
- Weak convergence implies computation complexity is $O(d)$.
- Working on extending the diffusion limits to more general target distributions.
- But how to use the optimality information?

Summary (cont'd)

- Adaptive MCMC tries to "learn" how to sample better. Good.
- Works well in examples like Adaptive Metropolis (200 × 200 covariance) and Componentwise Metropolis (503 dimensions).
- But must be done carefully, or it will destroy stationarity. Bad.
- To converge to π, suffices to have stationarity of each P_γ, plus (a) Diminishing Adaptation (important), and (b) Containment (technical condition, usually satisfied, necessary). Good.
- This can demonstrate convergence for adaptive Metropolis, coordinatewise adaptation, adaptive coordinate weights, etc.
- 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