# Adaptive Metropolis and Gibbs Samplers

Jeffrey S. Rosenthal, University of Toronto

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

---

G.O. Roberts and J.S. Rosenthal, Coupling and Ergodicity of Adaptive MCMC. J. Appl. Prob. **44** (2007), 458–475.

G.O. Roberts and J.S. Rosenthal, Examples of Adaptive MCMC. J. Comp. Graph. Stat. **18** (2009), 349–367.

Y. Bai, G.O. Roberts, and J.S. Rosenthal, On the Containment Condition for Adaptive Markov Chain Monte Carlo Algorithms. Adv. Appl. Stat. **21(1)** (2011), 1–54.

S. Richardson, L. Bottolo, and J.S. Rosenthal, Bayesian models for sparse regression analysis of high dimensional data. *Bayesian Statistics 9* (Valencia 2010).

K. Latuszynski, G.O. Roberts, and J.S. Rosenthal, Adaptive Gibbs samplers and related MCMC methods. Ann. Appl. Prob., to appear (accepted Sept 2011).

# Markov Chain Monte Carlo (MCMC)

Have a complicated, high-dimensional target distribution $\pi(\cdot)$.

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

Then for "large enough" $n$, $\mathcal{L}(X_n) \approx \pi(\cdot)$, so $X_n$, $X_{n+1}$, $\ldots$ are approximate samples from $\pi(\cdot)$, and e.g.

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

Extremely popular: Bayesian inference, computer science, statistical physics, finance, ...

How to find the good chains among the bad ones?

# Ex.: Random-Walk Metropolis Algorithm (1953)

This algorithm defines the chain $X_0, X_1, X_2, \ldots$ as follows.

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 $\alpha$, accept the proposal (set $X_n = Y_n$).
- Else, with prob. $1 - \alpha$, reject the proposal (set $X_n = X_{n-1}$).

But what is a smart choice of proposal covariance $\Sigma_p$?

Even if $\Sigma_p = \sigma I$, how large should $\sigma$ be?

Important – can vary from efficient to infeasible!

# Adaptive MCMC

Suppose have a family $\{P_\gamma\}_{\gamma \in \mathcal{Y}}$ of possible Markov chains, each with stationary distribution $\pi(\cdot)$. How to choose among them?

Trial and error? No, let the computer decide!

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

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: High-Dimensional Adaptive Metropolis

Dim $d = 100$, with target $\pi(\cdot)$ having target covariance $\Sigma_t$.

Here $\Sigma_t$ is $100 \times 100$ (i.e., 5,050 distinct entries).

Known (Roberts-Gelman-Gilks 1997, Roberts-R. 2001, Bédard 2006) that "optimal" Gaussian RWM proposal is $N(x,\ (2.38)^2\, d^{-1}\, \Sigma_t)$.

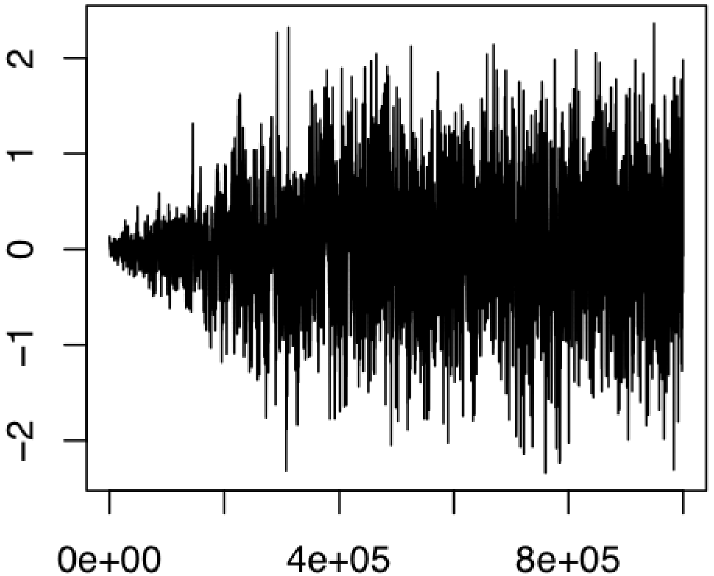
But usually $\Sigma_t$ unknown. Instead use empirical estimate, $\Sigma_n$. Let

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

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

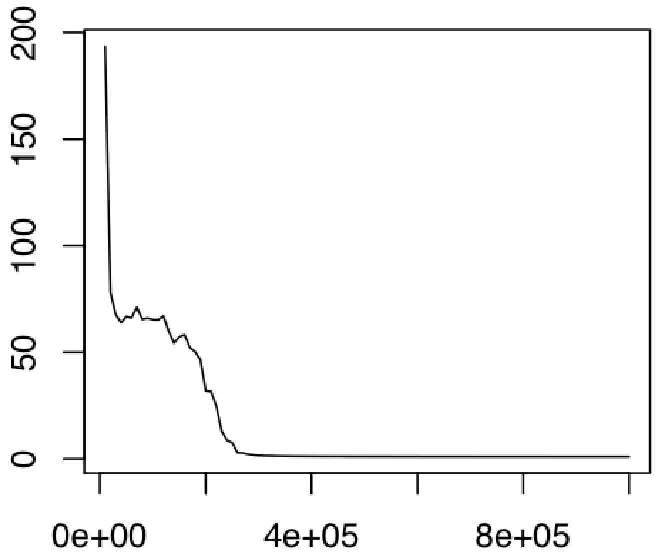
Let's try it …

# High-Dimensional Adaptive Metropolis (cont'd)



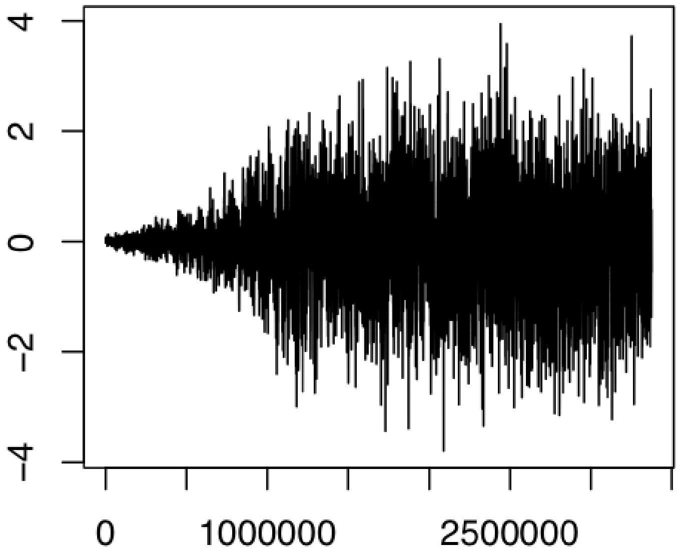

Plot of first coord. Takes about 300,000 iterations, then "finds" good proposal covariance and starts mixing well.

# 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 Adaptation




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

## Example: Adaptive Metropolis-within-Gibbs

Propose increment $N(0,\, e^{2\,ls_i})$ for $i^{\text{th}}$ coordinate, leaving the other other coordinates fixed; then repeat for different $i$.

Choice of $ls_i$??

Known that acceptance rate 0.44 is approximately optimal for one-dimensional Metropolis proposals. So:

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

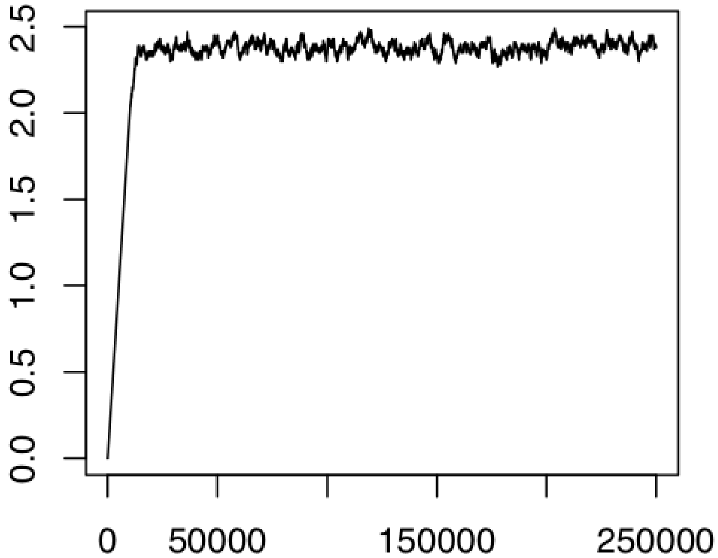
Adapt each $ls_i$, in batches, to seek 0.44 acceptance rate:

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

Let's try it ...

# Adaptive Metropolis-within-Gibbs (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 seems to find "good" values for the $ls_i$ values.

# Metropolis-within-Gibbs: Comparisons

| Variable | $J_i$ | Algorithm | $ls_i$ | ACT | Avr Sq Dist |
|---|---|---|---|---|---|
| $\theta_1$ | 5 | Adaptive | 2.4 | 2.59 | 14.932 |
| $\theta_1$ | 5 | Fixed | 0 | 31.69 | 0.863 |
| $\theta_2$ | 50 | Adaptive | 1.2 | 2.72 | 1.508 |
| $\theta_2$ | 50 | Fixed | 0 | 7.33 | 0.581 |
| $\theta_3$ | 500 | Adaptive | 0.1 | 2.72 | 0.150 |
| $\theta_3$ | 500 | Fixed | 0 | 2.67 | 0.147 |

The Adaptive algorithm mixes much more efficiently than the Fixed algorithm, with smaller integrated autocorrelation time (good) and larger average squared jumping distance (good).

And coordinates (e.g. $\theta_3$) that started good, stay good.

## Great . . . but is it Ergodic?

So, adaptive MCMC seems to work well in practice.

But will it be ergodic, i.e. converge to $\pi(\cdot)$?

Ordinary MCMC algorithms, i.e. with fixed choice of $\gamma$, are automatically ergodic by standard Markov chain theory (since they're irreducible and aperiodic and leave $\pi(\cdot)$ 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 $\gamma$ is such that $P_\gamma$ 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 $\mathcal{X}_0$ (see e.g. www.probability.ca/adaptjava).

# Ergodicity of Adaptive MCMC

Formally, suppose $\{P_\gamma\}_{\gamma\in\mathcal{Y}}$ is a family of Markov chains, with $\pi(\cdot)$ stationary for each $P_\gamma$, and adaption algorithm is defined by:

$$\mathbf{P}[X_{n+1} \in A \mid X_n = x, \Gamma_n = \gamma, \mathcal{G}_{n-1}] \; = \; P_\gamma(x, A)\,.$$

WANT: Simple conditions guaranteeing $\|\mathcal{L}(X_n) - \pi(\cdot)\| \to 0$,

where $\|\mathcal{L}(X_n) - \pi(\cdot)\| \equiv \sup_{A\subseteq\mathcal{X}} |\mathbf{P}(X_n \in A) - \pi(A)|$.

Many recent results, by many smart people, e.g.:

Finnish: Haario, Saksman, Tamminen, Vihola, . . .

French: Andrieu, Moulines, Robert, Fort, Atchadé, . . .

Australian: Kohn, Giordani, Nott, . . .

# 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. $\lim_{n\to\infty} \|\mathcal{L}(X_n) - \pi(\cdot)\| = 0$, 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 <u>made</u> 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 $\mathcal{Y}$ <u>finite</u>, or <u>compact</u>, 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.)

# Implications of Theorem

Adaptive Metropolis algorithm:

- Empirical estimates satisfy Diminishing Adaptation.

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

- So, Adaptive Metropolis is ergodic under such conditions.

Adaptive Metropolis-within-Gibbs algorithm:

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

- Satisfies Containment under boundedness or tail conditions.

- Hence, is also ergodic under such conditions.

Good!

## Choosing Which Coordinates to Update When

S. Richardson (statistical geneticist):

Successfully ran adaptive Metropolis-within-Gibbs algorithm on genetic data with thousands of coordinates (Turro, Bochkina, Hein, and Richardson, BMC Bioinformatics 2007). 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?

# Adapting the Gibbs Sampler Coordinate Weights

Consider "adaptive random-scan Gibbs samplers" (or "adaptive random-scan Metropolis-within-Gibbs algorithms"):

- At iteration $n$, choose coordinate $i$ with probability $\alpha_{n,i}$.

- Then, update coordinate $i$, either by proposing a move and then accepting/rejecting it (Metropolis-within-Gibbs), or by replacing its current value by a draw from its full conditional distribution (Gibbs Sampler).

- Allow the random-scan coordinate weights, $\{\alpha_{n,i}\}$, to be adapted, depending on the chain's history (e.g. gradually lower $\alpha_{n,i}$ if coordinate $i$ seems to change less often).

What conditions ensure ergodicity?

# Ergodicity of Adaptively Weighted Gibbs Samplers?

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.

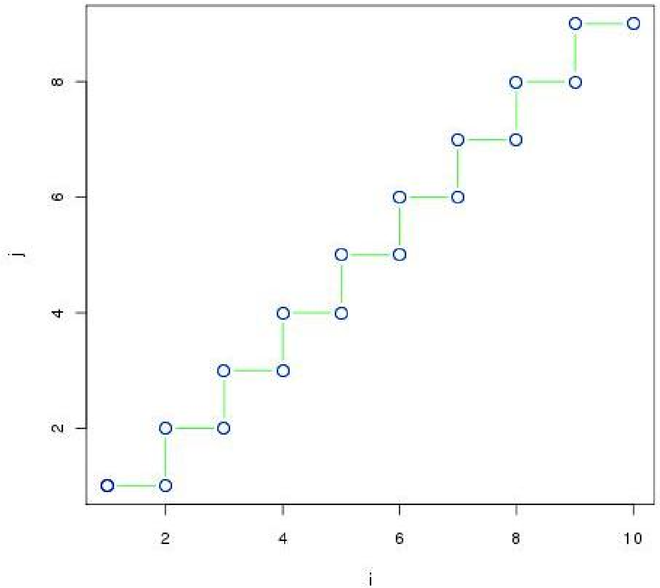
Really??

Proof seemed questionable ... but was result true?

Counter-example! (K. Latuszyński and R., 2009)

As follows ...

$\mathcal{X} = \{(i,j) \in \mathbf{N} \times \mathbf{N} : i = j \text{ or } i = j+1\}$ ("Stairway to Heaven").




Target $\pi(i,j) = C/j^2$, with adaptive coordinate weights given by:

$$\alpha_{n,1} = \begin{cases} (1/2) + \epsilon_n \,, & X_{n,1} = X_{n,2} \\ (1/2) - \epsilon_n \,, & X_{n,1} = X_{n,2} + 1 \end{cases}$$

and $\alpha_{n,2} = 1 - \alpha_{n,1}$, where $\epsilon_n \searrow 0$ sufficiently slowly.

Summary: $\mathcal{X} = \{(i,j) \in \mathbf{N} \times \mathbf{N} : i = j \text{ or } i = j+1\}$, and

$$\alpha_{n,1} \;=\; \begin{cases} (1/2) + \epsilon_n\,, & X_{n,1} = X_{n,2} \\ (1/2) - \epsilon_n\,, & X_{n,1} = X_{n,2} + 1 \end{cases}$$

and $\alpha_{n,2} = 1 - \alpha_{n,1}$, where $\epsilon_n \searrow 0$ sufficiently slowly.

Clearly $\alpha_{n,i} \to 1/2 =: \alpha_i^*$. And, the Gibbs sampler with fixed weights $(1/2,\, 1/2)$ is indeed ergodic (easy: usual MCMC).

So, the conditions of the previous "theorem" are satisfied.

However, the extra $\epsilon_n$ provides just enough outward "kick" that $\mathbf{P}(X_n \to \infty) > 0$, i.e. chain is <u>transient</u> and does not converge. Contradiction! "Theorem" is false!

So, we had better be smarter than that ...

# Ergodicity with Adaptive Coordinate Weights

We proved (Latuszynski, Roberts, and R., Ann. Appl. Prob., to appear) 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 $P_\gamma$, i.e. there is $V : \mathcal{X} \to [1, \infty)$ with

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

For the above counter-example, (i) fails because of the infinite tails, and (ii) fails because of the slight outward kick.

But if careful about continuity, boundedness, etc., then can guarantee ergodicity in many cases, including for high-dimensional genetics data (Richardson, Bottolo, R., Valencia 2010).

# Summary

Adaptive MCMC tries to "learn" how to sample better. Good.

Works well in examples like Adaptive Metropolis ($200 \times 200$ covariance matrix) and Metropolis-within-Gibbs (503 dimensions).

But must be done carefully, or it will destroy stationarity. Bad.

To converge to $\pi(\cdot)$, suffices to have stationarity of each $P_\gamma$, plus (a) Diminishing Adaptation (important), and (b) Containment (technical condition, usually satisfied). Good.

For Gibbs and Metropolis-within-Gibbs samplers, can also adapt the coordinate weights $\alpha_{n,i}$, but <u>only</u> if the target distribution satisfies certain uniformity or tail conditions. Good.

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