A Repelling-Attracting Metropolis (RAM) Algorithm for Multimodality

Hyungsuk Tak[1], Xiao-Li Meng[2], and David A. van Dyk[3]

SAMSI[1], Harvard University[2] and Imperial College London[3]

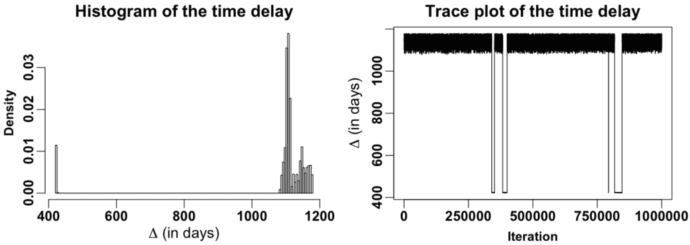
International CHASC Astro-Statistics Collaboration

2 Nov 2016

Motivation

Full posterior density function: $\pi(\Delta, \boldsymbol{\theta} \mid \text{Data})$

- Metropolis within Gibbs sampler (Tierney, 1994)
- A multimodal (marginal) posterior of Δ for Quasar Q0957+561



- Just 8 jumps out of a million iterations!
- MCMC estimate of the relative height of each mode is not reliable.
- Could we improve Metropolis' ability to jump between modes without losing its simple-to-implement characteristic?

Idea

There is a RAM on top of the mountain.
How would this RAM move to the top of the other mountain?

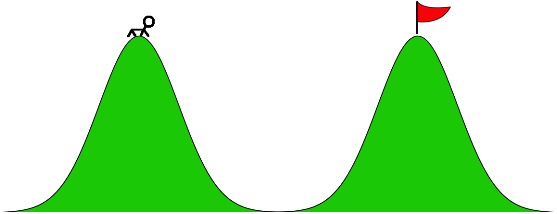


Image credit: www.launsteinimagery.com, www.cliparts.com, www.iconfinder.com

Idea (cont.)

1. Make a down-up movement in density to generate a proposal x''.

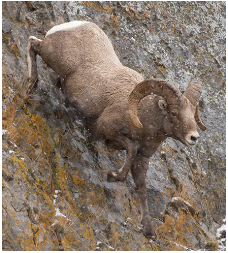
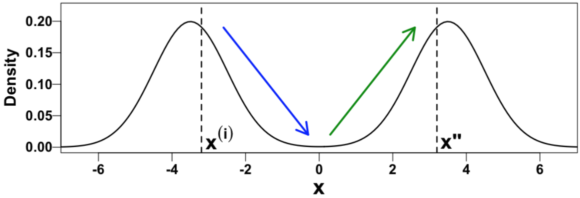


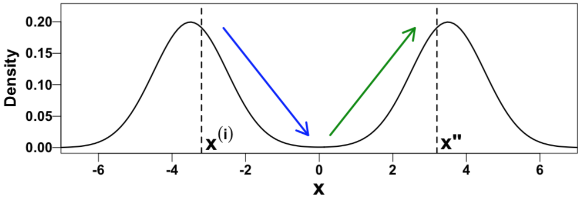

Image credit: http://www.bestofthetetons.com/, http://blog.showmenaturephotography.com/



2. Accept or reject x'' with probability $\min\left\{1,\ \frac{\pi(x'')q^{\mathrm{DU}}(x^{(i)}|x'')}{\pi(x^{(i)})q^{\mathrm{DU}}(x''|x^{(i)})}\right\}$.

Note: q^{DU} is a down-up (DU) proposal density.

RAM: Proposal



Two-step procedure

$x^{(i)}$: Current state $\searrow$ x': Intermediate proposal $\nearrow$ x'': Final proposal

1. (Downhill Metropolis) Generate $x' \sim \mathrm{N}(x^{(i)}, \sigma^2)$ and accept x' with probability $\alpha_\epsilon^{\mathrm{D}}(x' \mid x^{(i)}) = \min\left\{1,\ \frac{\pi(x^{(i)})+\epsilon}{\pi(x')+\epsilon}\right\}$. Repeat this step until one proposal is accepted (a forced Metropolis).

   Note: σ is RAM's tuning parameter, and $\epsilon = 10^{-308}$ prevents $0/0$.

2. (Uphill Metropolis) Generate $x'' \sim \mathrm{N}(x', \sigma^2)$ and accept x'' with probability $\alpha_\epsilon^{\mathrm{U}}(x'' \mid x') = \min\left\{1,\ \frac{\pi(x'')+\epsilon}{\pi(x')+\epsilon}\right\}$ Repeat this step until one proposal is accepted (a forced Metropolis).

RAM: Acceptance/Rejection

Accept x'' with a Metropolis-Hastings acceptance probability

$$\alpha^{\mathrm{DU}}(x'' \mid x^{(i)}) = \min\left\{1,\ \frac{\pi(x'')q^{\mathrm{DU}}(x^{(i)} \mid x'')}{\pi(x^{(i)})q^{\mathrm{DU}}(x'' \mid x^{(i)})}\right\}$$

$$= \min\left\{1,\ \frac{\pi(x'')\int \mathrm{N}(x \mid x^{(i)}, \sigma^2)\alpha_\epsilon^{\mathrm{D}}(x \mid x^{(i)})dx}{\pi(x^{(i)})\int \mathrm{N}(x \mid x'', \sigma^2)\alpha_\epsilon^{\mathrm{U}}(x \mid x'')dx}\right\}.$$

Is there a way to avoid calculating this ratio of intractable integrals?

If we explore an expanded space with a correct marginal $\pi(x)$, then there can be a way to cancel this intractable ratio (Møller et al., 2006).

RAM: Auxiliary variable approach

An auxiliary variable z with $\pi^{\mathrm{C}}(z \mid x)$ well-defined.

- Joint target density: $\pi^{\mathrm{J}}(z, x) = \pi(x)\pi^{\mathrm{C}}(z \mid x) = \pi(x)\mathrm{N}(z \mid x, \sigma^2)$
- Joint proposal density:

$$q^{\mathrm{J}}(z'', x'' \mid z^{(i)}, x^{(i)}) = q_1(x'' \mid z^{(i)}, x^{(i)})q_2(z'' \mid x'', z^{(i)}, x^{(i)})$$
$$= q^{\mathrm{DU}}(x'' \mid x^{(i)})q^{\mathrm{D}}(z'' \mid x'')$$

  Note: q^{D} is a forced downhill kernel density.

- Joint acceptance probability:

$$\alpha^{\mathrm{J}}(z'', x'' \mid z^{(i)}, x^{(i)}) = \min\left[1, \frac{\pi^{\mathrm{J}}(z'', x'')q^{\mathrm{J}}(z^{(i)}, x^{(i)} \mid z'', x'')}{\pi^{\mathrm{J}}(z^{(i)}, x^{(i)})q^{\mathrm{J}}(z'', x'' \mid z^{(i)}, x^{(i)})}\right]$$
$$= \min\left[1, \frac{\pi(x'')\min\{1, \frac{\pi(x^{(i)})+\epsilon}{\pi(z^{(i)})+\epsilon}\}}{\pi(x^{(i)})\min\{1, \frac{\pi(x'')+\epsilon}{\pi(z'')+\epsilon}\}}\right], \text{ nothing intractable here!}$$

RAM: Overall algorithm

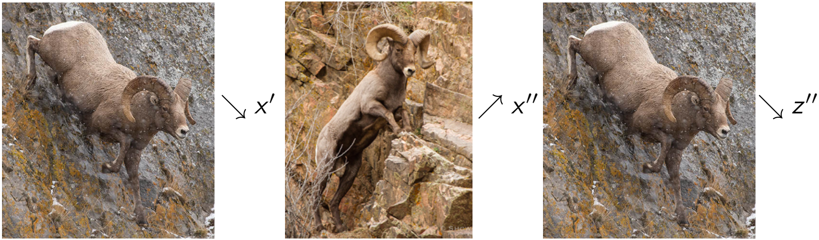


Image credit: http://www.bestofthetetons.com/, http://blog.showmenaturephotography.com/

A RAM is composed of four steps in each iteration.

Steps 1–3: Generating a joint proposal (z'', x'')

1. $(\searrow)$ Redraw $x' \sim \mathrm{N}(x^{(i)}, \sigma^2)$ until $u_1 \sim \mathrm{Unif}(0, 1) < \alpha_\epsilon^{\mathrm{D}}(x' \mid x^{(i)})$
2. $(\nearrow)$ Redraw $x'' \sim \mathrm{N}(x', \sigma^2)$ until $u_2 \sim \mathrm{Unif}(0, 1) < \alpha_\epsilon^{\mathrm{U}}(x'' \mid x')$
3. $(\searrow)$ Redraw $z'' \sim \mathrm{N}(x'', \sigma^2)$ until $u_3 \sim \mathrm{Unif}(0, 1) < \alpha_\epsilon^{\mathrm{D}}(z'' \mid x'')$

Step 4: Accept or reject the joint proposal (z'', x'')

4. Set $(z^{(i+1)}, x^{(i+1)}) = (z'', x'')$ if $u_4 < \alpha^{\mathrm{J}}(z'', x'' \mid z^{(i)}, x^{(i)})$, where $u_4 \sim \mathrm{Unif}(0, 1)$, and set $(z^{(i+1)}, x^{(i+1)}) = (z^{(i)}, x^{(i)})$ otherwise.

Example 1: Quasar Q0957+561 (Hainline et al, 2012)

A Metropolis within Gibbs sampler for $p(\Delta, \boldsymbol{\theta} \mid \mathrm{Data})$

$$\text{Step 1: Sample } \Delta^{(i)} \sim p(\Delta \mid \boldsymbol{\theta}^{(i-1)}, \mathrm{Data})$$
$$\text{Step 2: Sample } \boldsymbol{\theta}^{(i)} \sim p(\boldsymbol{\theta} \mid \Delta^{(i)}, \mathrm{Data})$$

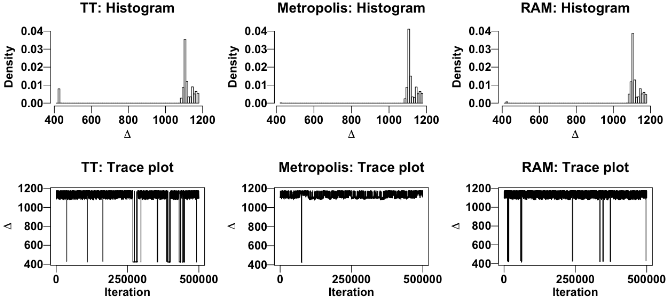
Implementation details:

- We use tempered transitions (TT) (Neal, 1996), Metropolis or RAM to draw Δ in Step 1.
- We run 10 chains each of length 100,000, discarding the first 50,000 as burn-in, with 10 initial values of Δ spread across the space $[-1100,\ 1100]$.
- We set a fairly large proposal scale ($\sigma = 500$).
- We consider two cases, the same number of iterations and the same amount of CPU time.

Example 1: Quasar Q0957+561 (cont.)

The same number of iterations

TT: Tempered Transitions



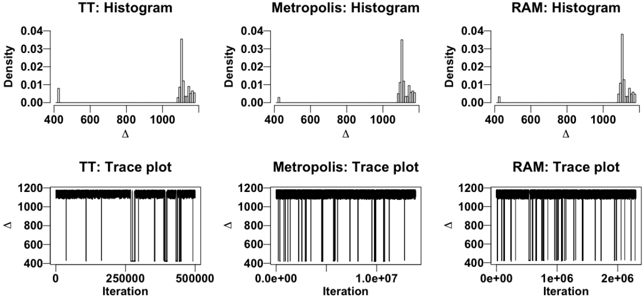
	# of iterations of each chain	Avg. CPU time (seconds)	# of chains that jump	Total # of jumps
TT	100,000	2,072	7	46
Metropolis	100,000	144	1	4
RAM	100,000	740	6	64

MCMC estimate of the relative height of the mode near 400 days?

Example 1: Quasar Q0957+561 (cont.)

The same amount of CPU time

TT: Tempered Transitions



	# of iterations of each chain	Avg. CPU time (seconds)	# of chains that jump	Total # of jumps
TT	100,000	2,072	7	46
Metropolis	1,436,150	2,072	10	126
RAM	279,872	2,071	10	228

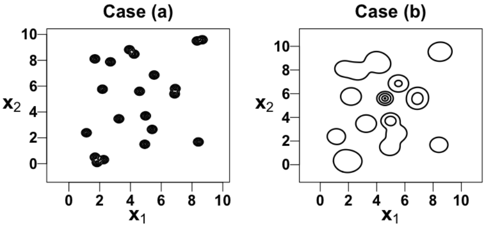
More reliable MCMC estimate of the relative height of the mode!

Example 2: Mixture of 20 bivariate Normals

A mixture of 20 bivariate Gaussian distributions (Kou et al., 2006)

$$\pi(x) \propto \sum_{j=1}^{20} \frac{w_j}{2\pi\tau_j^2} \exp\left(-\frac{1}{2\tau_j^2}(x-\mu_j)^\top (x-\mu_j)\right),$$

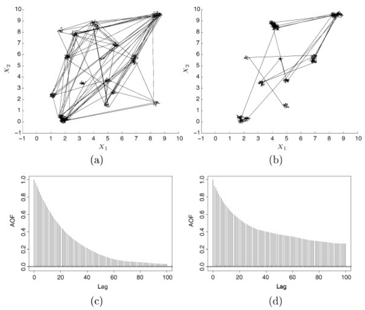
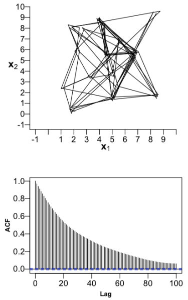
where $x = (x_1, x_2)^\top$. According to weights and variances, Kou et al. (2006) set up two different targets distributions.



Implementation details (the same configuration)

- We run 20 chains each of length 100,000, discarding the first 50,000.
- We do not consider the CPU time.
- Proposal scales: $\Sigma = 4^2 I_2$ for case (a) and $\Sigma = 3.5^2 I_2$ for case (b).

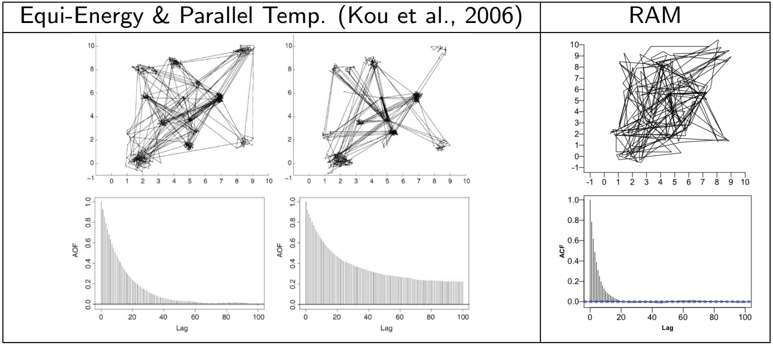
Example 2: Case(a) for EE vs PT vs RAM

Equi-Energy & Parallel Temp. (Kou et al., 2006)	RAM
	

	$E(X_1)$	$E(X_2)$	$E(X_1^2)$	$E(X_2^2)$
True value	4.478	4.905	25.605	33.920
RAM	4.4741 (0.094)	4.9016 (0.107)	25.6251 (0.943)	33.8972 (1.083)
EE	4.5019 (0.107)	4.9439 (0.139)	25.9241 (1.098)	34.4763 (1.373)
PT	4.4185 (0.170)	4.8790 (0.283)	24.9856 (1.713)	33.5966 (2.867)
MSE ratio (EE/RAM)	1.36	1.82	1.47	1.87
MSE ratio (PT/RAM)	3.67	7.05	3.73	7.09

RAM is better w.r.t. MSE, meaning that each chain visits 20 modes more consistently (more reliable estimates for relative heights!).

Example 2: Case (b) for EE vs PT vs RAM



	$E(X_1)$	$E(X_2)$	$E(X_1^2)$	$E(X_2^2)$
True value	4.688	5.030	25.558	31.378
RAM	4.687 (0.026)	5.035 (0.039)	25.662 (0.252)	31.532 (0.330)
EE	4.699 (0.072)	5.037 (0.086)	25.693 (0.739)	31.433 (0.839)
PT	4.709 (0.116)	5.001 (0.134)	25.813 (1.122)	31.105 (1.186)
MSE ratio (EE/RAM)	7.84	4.82	7.59	5.33
MSE ratio (PT/RAM)	20.53	12.16	17.81	11.17

RAM is much better w.r.t. MSE, meaning that each chain visits 20 modes more consistently (more reliable estimates for relative heights!).

Concluding remarks

Take-home messages

- Simple to implement (Please keep it in your MCMC toolbox).
- Always possible to replace Metropolis with RAM for multimodality.
- RAM also shows a better high-dimensional behavior compared to Metropolis (ArXiv 1601.05633)

Future directions

- Theoretical convergence rate
- Possibly many (and better) down-up schemes, e.g., anti-Langevin + Langevin (Christian Robert), negative temperature + positive temperature (Art Owen).
- A global optimizer based on the down-up idea (analog to annealing)

Xi'an's Og, "love-hate Metropolis algorithm" (hate-love?)

References

1. Hainline, L. J. et al. (2012) "A New Microlensing Event in the Doubly Imaged Quasar Q0957+ 561" *The Astrophysical Journal*, **744**, 2, 104.
2. Kou, S. C. et al. (2006) "Discussion Paper: Equi-Energy Sampler with Applications in Statistical Inference and Statistical Mechanics" *The Annals of Statistics*, **34**, 4, 1581–1619.
3. Møller, J., Pettitt, A. N., Reeves, R., Berthelsen, K. K. (2006) "An Efficient Markov Chain Monte Carlo Method for Distributions with Intractable Normalising Constants" *Biometrika*, **93**, 2, 451–458.
4. Neal, R. M. (1996) "Sampling From Multimodal Distributions Using Tempered Transitions" *Statistics and Computing*, **6**, 4, 353–366.
5. Tak, H., Meng, X.-L., van Dyk, D. A. (in progress), "A Repulsive-Attractive Metropolis Algorithm for Multimodality." arXiv 1601.05633
6. Tak, H., Mandel, K., van Dyk, D. A., Kashyap, V. L., Meng, X.-L., Siemiginowska, A. (in progress), "Bayesian Estimates of Astronomical Time Delays between Gravitationally Lensed Stochastic Light Curves." arXiv 1602.01462
7. Tierney, L. (2009) "Markov Chains for Exploring Posterior Distributions" *The Annals of Statistics*, **22**, 4, 1701–1728.