Metropolis-Hastings Markov Chains in the High-Dimensional, Large Sample Size Regime

Galin Jones[1]

University of Minnesota

July 2024

[1] Joint work with Riddhiman Bhattacharya and Austin Brown

Toy Example

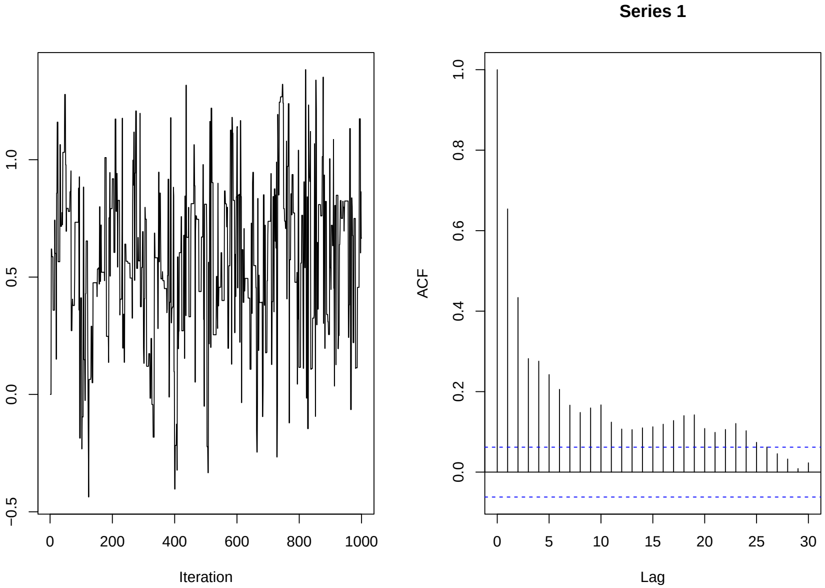
$$X_1, \ldots, X_n \overset{iid}{\sim} N(\theta, 1), \qquad \theta \sim N(0, 1)$$

$$\theta \mid x_1, \ldots, x_n \sim N\left(\frac{n\bar{x}_n}{n+1}, \frac{1}{n+1}\right)$$

Use random walk Metropolis-Hastings with Normal proposal having variance h

n=2, h=1

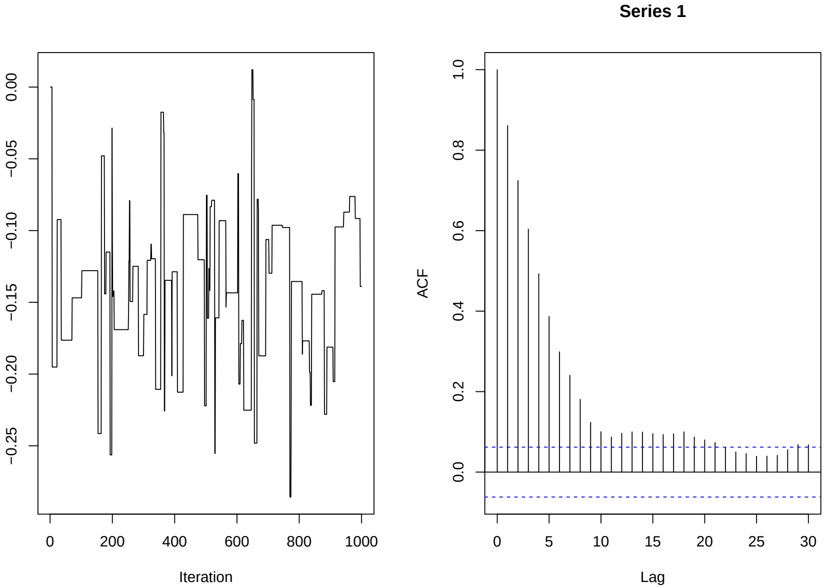
```
## [1] 0.365
```



Series 1
1.0
0.5
0.0
−0.5
0
200
400
600
800
1000
Iteration
ACF
1.0
0.8
0.6
0.4
0.2
0.0
0
5
10
15
20
25
30
Lag

n=20, h=1

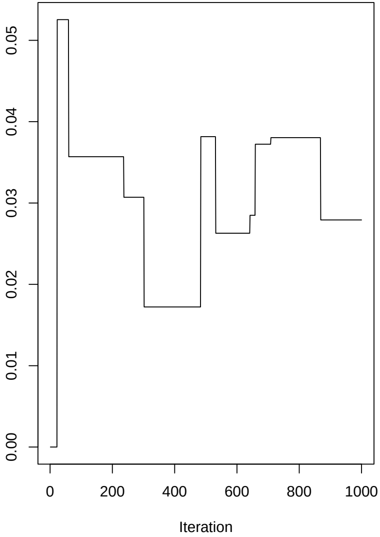
```
## [1] 0.08
```

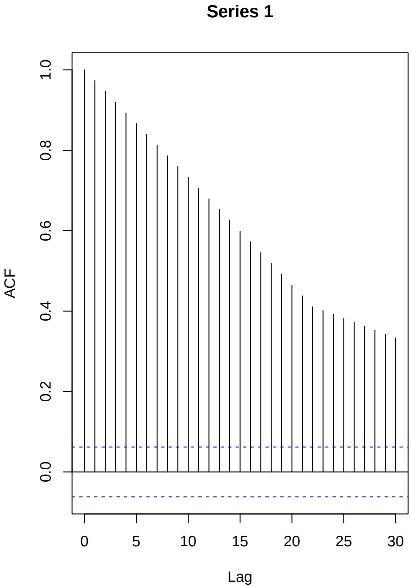


0.00
−0.05
−0.10
−0.15
−0.20
−0.25
0
200
400
600
800
1000
Iteration
Series 1
1.0
0.8
0.6
0.4
0.2
0.0
ACF
0
5
10
15
20
25
30
Lag

n=200, h=1

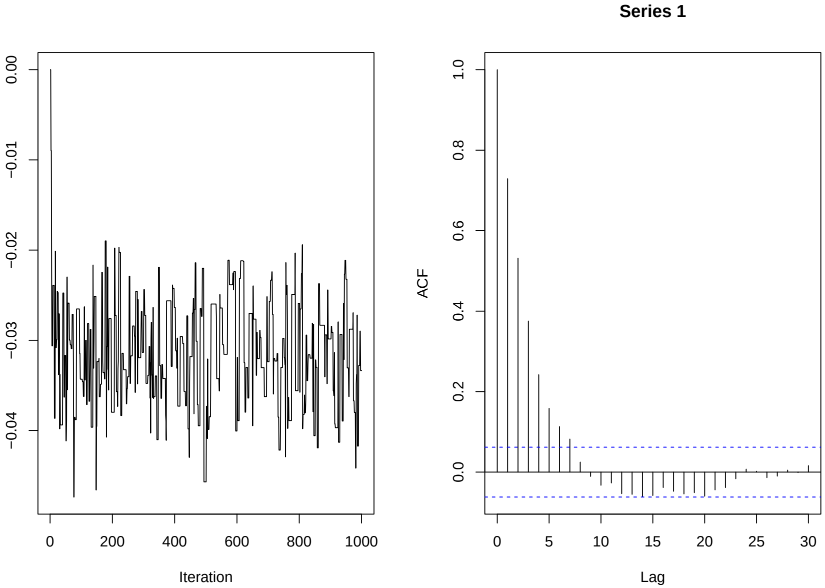
```
## [1] 0.01
```


0.00
0.01
0.02
0.03
0.04
0.05
0
200
400
600
800
1000
Iteration


Series 1
ACF
0.0
0.2
0.4
0.6
0.8
1.0
0
5
10
15
20
25
30
Lag

n=200, h=1/2000

```
## [1] 0.281
```



Take-Home Message

To avoid guaranteed failure of MH, the proposal scaling needs to account for the sample size and the dimension.

Can we identify when failure is guaranteed in general MH simulations so that we can avoid it?

Notation

The MH Markov kernel describes the dynamics of the Markov chain. Informally,

$$\Pr(X_{t+j} \in B \mid X_j = x) = P^t(x, B)$$

$$\|P^t(x, \cdot) - F(\cdot)\|_{TV} \to 0 \qquad t \to \infty$$

Question: When will this convergence take prohibitively long?

Lower Bound

Theorem If

$$A_h(x) = \int \left[\frac{f(x')q_h(x \mid x')}{f(x)q_h(x' \mid x)} \wedge 1 \right] q_h(x' \mid x) dx'.$$

then, for every x,

$$\|P^t(x, \cdot) - F(\cdot)\|_{TV} \geq [1 - A_h(x)]^t$$

Answer: Avoid $A_h(x) \approx 0$.

Gaussian Proposals

Suppose $\mu : \mathbb{R}^d \to \mathbb{R}^d$ and consider a proposal of the form

$$N_d(\mu(x), hC)$$

For example:

- Independence Sampler: $\mu(x) = c$
- Random Walk: $\mu(x) = x$
- MALA: $\mu(x) = x + h(\nabla \log x)/2$

Gaussian Proposals

Suppose $\mu : \mathbb{R}^d \to \mathbb{R}^d$ and consider a proposal of the form

$$N_d(\mu(x), hC)$$

then

$$A_h(x) \leq \frac{1}{f(x)(2\pi h)^{d/2} \det(C)^{1/2}}$$

Suggests

h must be small to avoid poor convergence properties

MH chains can have poor dimension dependence unless that scaling is chosen carefully:

$h \propto d^{-\delta}$ for some $\delta > 0$

Geometric Ergodicity

P is geometrically ergodic if there exists $\rho < 1$ such that

$$\|P^t(x, \cdot) - F(\cdot)\|_{TV} \leq M(x)\rho^t$$

Key sufficient condition for

- Markov chain Central Limit Theorem

$$\sqrt{m}(\bar{X}_m - E_F(X)) \to \mathrm{N}_p(0, \Sigma)$$

- Consistency of estimators of Σ such as batch means

Geometric Ergodicity

P is geometrically ergodic if there exists $\rho < 1$ such that

$$\|P^t(x, \cdot) - F(\cdot)\|_{TV} \leq M(x)\rho^t$$

Few results on constraining ρ for MH algorithms. That is, there are almost no results that identify constants such that

$$C_1 \leq \rho \leq C_2$$

Geometeric Ergodicity

P is geometrically ergodic if there exists $\rho < 1$ such that

$$\|P^t(x,\cdot) - F(\cdot)\|_{TV} \leq M(x)\rho^t$$

<u>Theorem</u> If P is geometrically ergodic, then

$$\rho \geq 1 - \inf_x A_h(x)$$

Another Toy Example

Let $b > 1$ and

$$g(x,y) = \frac{b}{\pi} e^{-(x^2+b^2y^2)} \quad \text{and} \quad h(x,y) = \frac{b}{\pi} e^{-(b^2x^2+y^2)}.$$

Set

$$f(x,y) = \frac{1}{2} g(x,y) + \frac{1}{2} h(x,y)$$

RWMH using a Gaussian proposal with scaling h is geometrically ergodic.

Our result says

$$\rho \geq 1 - \frac{1}{2bh}$$

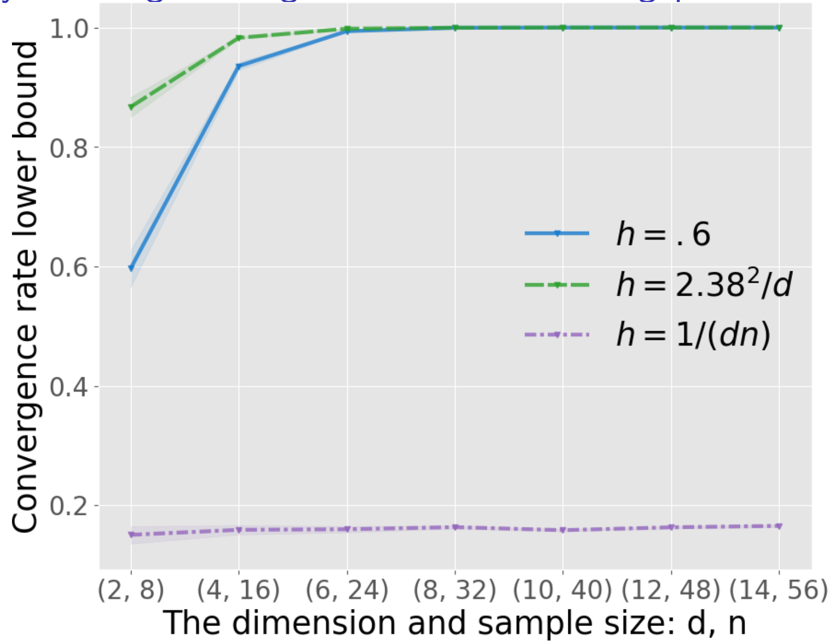
Bayesian Logistic Regression with Zellner's g-prior

$$\pi_n(\beta) \propto \prod_{i=1}^{n} s\left(\beta^T X_i\right)^{Y_i} \left(1 - s\left(\beta^T X_i\right)\right)^{1-Y_i} \exp\left(-\frac{1}{2g}\beta^T X^T X\beta\right).$$

<u>Proposition</u> Let β_n^* denote the point which maximizes π_n. If $n \to \infty$ in such a way that $d_n/n \to \gamma \in (0,1)$, then, with probability 1, for all sufficiently large n, the acceptance probability for RWMH satisfies

$$A(\beta_n^*) \leq \left(\frac{hn(1-\sqrt{\gamma})^2}{2g} + 1\right)^{-d_n/2}.$$

Bayesian Logistic Regression with Zellner's g-prior



Bayesian Logistic Regression with a flat prior

Suppose for $i = 1, \ldots, n$, (Y_i, X_i) are iid and

$$Y_i \mid X_i, \beta \overset{ind}{\sim} \text{Bern}\left(\left(1 + \exp\left(-\beta^T X_i\right)\right)^{-1}\right)$$

and $\nu(d\beta) = d\beta$.

<u>Theorem</u> If β_n^* maximizes the posterior and $d_n \leq n^\kappa$, $\kappa \in (0, 1)$, the posterior concentrates at β_n^*, under regularity conditions.

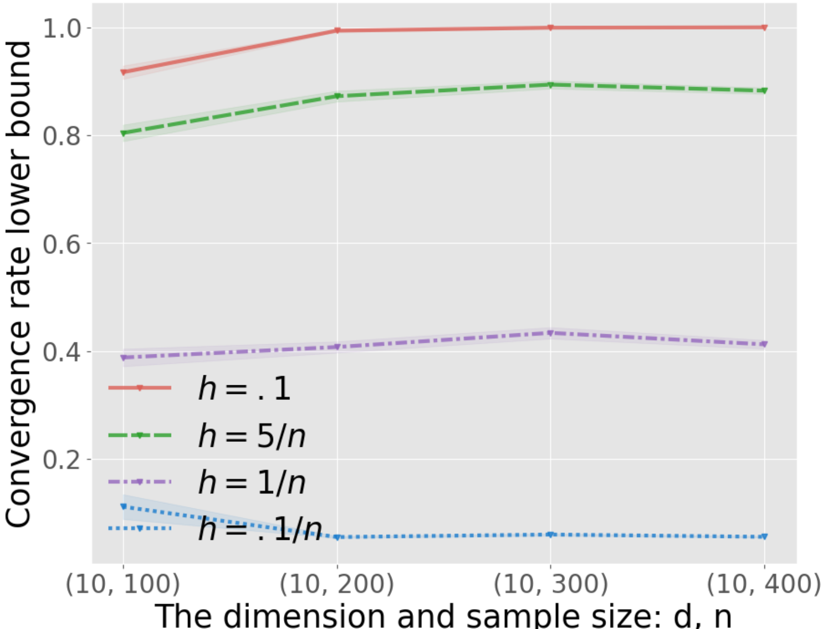
Bayesian Logistic Regression with a flat prior

If the proposal is $N_d(\mu(\beta), hC)$, then MH satisfies

$$A_h(\beta_n^*) \leq K \left(\frac{1}{nh}\right)^{d_n/2}$$

Take Home Message: Proposal scaling needs to depend carefully on both n and d.

Bayesian Logistic Regression with a flat prior



What's in the papers

Lower bounds:

Lower bounds in both total variation and Wasserstein distances

Comparison to conductance methods

General lower bounds under posterior concentration

Existence of a spectral gap is equivalent to a geometric rate of convergence in many Wasserstein distances

RWMH constraints

- Weaker conditions for geometric ergodicity of RWMH
- Explicit drift and minorization conditions
- Applications to a large class of Bayesian generalized linear models
- Lower bounds on ρ using spectral $(L^2(F))$ theory

The Papers

Brown and Jones (2024) Lower Bounds on the Rate of Convergence for Accept-Reject-Based Markov Chains in Wasserstein and Total Variation Distances, To appear in *Bernoulli*

Bhattacharya and Jones (2024) Explicit Constraints on the Geometric Rate of Convergence of Random Walk Metropolis-Hastings Algorithms, To appear in *Bernoulli*

Upper Bounds

$$(PV)(x) = \int V(y)P(x,dy) \le \lambda V(x) + L$$

and

$$P(x,\cdot) \ge \epsilon G(\cdot) \qquad \{V(x) \le d\} \quad \text{with} \quad d > 2L/(1-\lambda)$$

Theorem (Rosenthal, JASA, 1995) The Markov chain is geometrically ergodic and

$$M(x) \le 1 + \frac{L}{1-\lambda} + V(x)$$

and

$$\rho \le \max\left\{(1-\eta)^r, \alpha^{-(1-r)}c^r\right\}$$

with

$$\alpha^{-1} = \frac{1+2L+\lambda d}{1+d} \qquad \text{and} \qquad c = 1 + 2(\lambda + L)$$

Bounds for Random Walk MH

Consider RWMH on $\mathbb{R}^d$, then there are explicit drift and minorization conditions using the drift function

$$V(x) = \frac{1}{\sqrt{f(x)}}$$

Key assumption:

$$\limsup_{\|x\|\to\infty} \left\langle \frac{x}{\|x\|}, \frac{\nabla f(x)}{\|\nabla f(x)\|} \right\rangle < 0.$$

and hence

$$1 - \inf_x A_h(x) \leq \rho \leq \max\left\{(1-\eta)^r, \alpha^{-(1-r)} c^r\right\}$$