Lower Bounds on the Rate of Convergence for Accept-Reject-Based Markov Chains

Austin Brown [1] joint work with Galin L. Jones (University of Minnesota)

Postdoc at the University of Warwick, Coventry, United Kingdom

July 5, 2023

[1]austin.d.brown@warwick.ac.uk

Setting

We have a target distribution Π on $\mathbb{R}^d$ possibly depending on data of size n (e.g. Bayesian posteriors) with Lebesgue density π with support $\Theta \subseteq \mathbb{R}^d$.

We want to generate representative samples $\theta_1, \ldots, \theta_T$ from Π to approximate expectations (e.g. predictions and inference in Bayesian statistics)

$$\frac{1}{T}\sum_{t=1}^{T} g(\theta_t) \approx \int g(\theta)\pi(\theta)d\theta.$$

Bayesian Applications

Consider the Bayesian model

$$\theta \sim \pi_0$$
$$X_1, Y_1, \ldots, X_n, Y_n | \theta \sim p_{n,\theta}(\cdot)$$

Observe $x_1, y_1, \ldots, x_n, y_n$ and the Bayesian posterior density

$$\pi_n(\theta) \propto \pi_0(\theta) p_{n,\theta}(x_1, y_1, \ldots, x_n, y_n).$$

Often the density is intractable and the normalizing constant is difficult to approximate.

Example:

- Bayesian GLM's (i.e. logistic regression)

Difficulties in Sampling

- **Independent sampling methods are infeasible for modern complex target distributions.**
  - Inverting the distribution function / transformation methods (e.g Normal distributions)
  - Rejection sampling (Gamma distributions, etc)
  - Normalizing constant of π is often unknown (difficult for importance sampling)

MCMC approach: Simulate a Markov chain for sufficiently long with stationary distribution Π.

Main concern: How long do we need to simulate the Markov chain?

Accept-reject-based Markov chains

Generate $\theta_t | \theta_0 \sim P^t(\theta_0, \cdot)$ in discrete-time using a proposal $Q(\cdot, \cdot)$ by

$$\theta_t | \theta_{t-1} = \begin{cases} \theta'_t, & \text{if } u_t \leq a(\theta_{t-1}, \theta'_t) \\ \theta_{t-1}, & \text{else} \end{cases}$$

where $\theta'_t | \theta_{t-1} \sim Q(\theta_{t-1}, \cdot)$ and $u_t \sim \text{Unif}(0, 1)$.

Choose $a(\cdot, \cdot)$ in a way so Π is invariant (not necessarily reversible).

Examples: Metropolis-Hastings [Metropolis et al., 1953, Hastings, 1970, Tierney, 1998], Barker's Barker [1964], non-reversible Metropolis-Hastings Bierkens [2015]

Metropolis-Hastings

If $Q(\cdot, \cdot)$ with transition density $q(\cdot, \cdot)$, **Metropolis-Hastings:**

$$a(\theta_{t-1}, \theta_t') = \frac{\pi\left(\theta_t'\right) q\left(\theta_t', \theta_{t-1}\right)}{\pi\left(\theta_{t-1}\right) q\left(\theta_{t-1}, \theta_t'\right)} \wedge 1$$

Optimal in a Peskun sense [Tierney, 1998].

Figure: Arianna Rosenbluth, Nicholas Metropolis, Keith Hastings, and Luke Tierney

Drawbacks

- **Drawbacks:**
  - Requires choosing an acceptance function $a(\cdot, \cdot)$.
  - Requires **choosing** and **tuning** a proposal $Q(\cdot, \cdot)$.
    - Independent proposal
    - Random-walk proposals (RWM): $\theta'_t \sim N(\theta_{t-1}, hI_d)$
    - Discretize Langevin diffusions (MALA): $\theta'_t \sim N(\theta_{t-1} + h\nabla \log(\pi(\theta_{t-1}))/2, hI_d)$
  - Can be **unreliable** if the proposal is chosen poorly.

The Problem

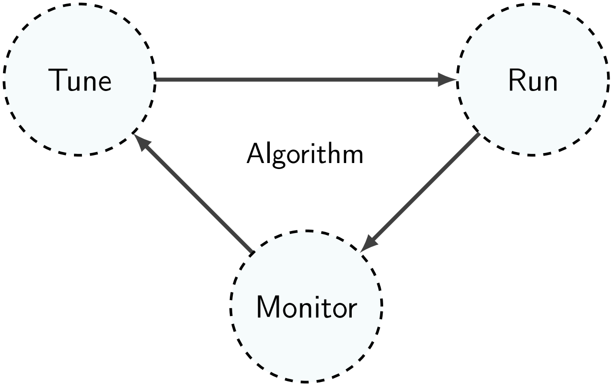
Problem: Practitioners often require tuning proposals by trial and error to avoid poor empirical performance in many applications.

We want to contribute to existing tools for choosing tuning parameters:

- Optimal scaling for RWM, MALA Roberts et al. [1997], Roberts and Rosenthal [1998]
- Adaptive algorithms Haario et al. [2001]
- Convergence analysis

Trial and Error

- **Trial and error:** Tune the algorithm, run the algorithm, monitor acceptances, and restart if acceptances are low.



Drawbacks:

- Computationally intensive and time consuming.

Convergence Rate Upper bounds

Recent interest in (geometric) convergence rate upper bounds for MCMC algorithms in terms of the problem size d, n [Belloni and Chernozhukov, 2009, Ekvall and Jones, 2021, Johndrow et al., 2019, Qin and Hobert, 2019, Rajaratnam and Sparks, 2015, Yang et al., 2016].

Also an interest in new coupling techniques in Wasserstein distances since they appear to scale better in high dimensions [Hairer et al., 2014, Qin and Hobert, 2019, 2021].

Difficulties for Convergence Rate Upper Bounds

Difficulties: Explicit convergence rates for Metropolis-Hastings in TV or Wasserstein is a challenging problem and the convergence rates are largely unknown [Jarner and Hansen, 2000, Hairer et al., 2014].

Some mixing time bounds and recent work in the area [Dwivedi et al., 2018, Andrieu et al., 2022].

Convergence Rate Lower bounds

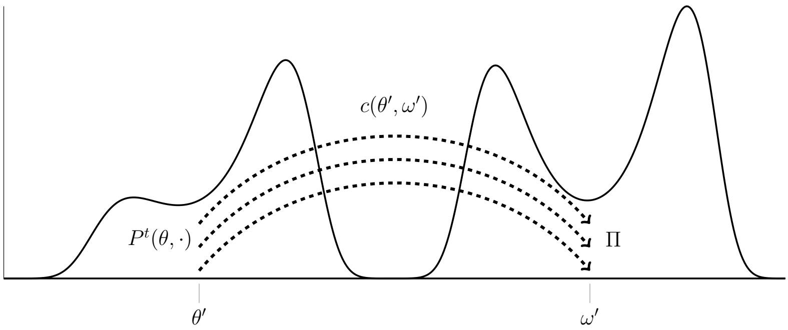
Use lower bounds on the convergence rates to aid practitioners in understanding which tuning parameters may cause the algorithms to *fail* produce a representative sample in an available number of iterations in terms of d, n.

Drawbacks: Does not tell you when the algorithm performs well.

Related literature: Exact rates for independence samplers [Wang, 2022, Brown and Jones, 2021]. Necessary conditions for geometric ergodicity [Roberts and Tweedie, 1996].

Geometric Ergodicity in TV and Wasserstein Distances

Intuition: Transportation Distances



Optimally transport all the mass from one probability distribution to the other with cost $c(\cdot, \cdot)$.

Examples: $c(\theta', \omega') = I_{\theta' \neq \omega'}$ and $c(\theta', \omega') = \|\theta' - \omega'\|$.

Transportation Distances

Let $\mathcal{C}\left(P^t(\theta,\cdot),\Pi\right)$ be the set of couplings. The Wasserstein distance is defined as

$$\mathcal{W}^p_{\|\cdot\|}\left(P^t(\theta,\cdot),\Pi\right) = \left(\inf_{\xi\in\mathcal{C}(P^t(\theta,\cdot),\Pi)} \int \left\|\theta' - \omega'\right\|^p d\xi(\theta',\omega')\right)^{1/p}$$

in comparison to

$$\left\|P^t(\theta,\cdot) - \Pi\right\|_{\mathsf{TV}} = \inf_{\xi\in\mathcal{C}(P^t(\theta,\cdot),\Pi)} \int I_{\theta'\neq\omega'} d\xi(\theta',\omega')$$

Figure: Leonid Kantorovich, Leonid Vaseršteĭn, Cédric Villani

Geometric Ergodicity

An accept-reject-based Markov chain is (ρ, M)-**geometrically ergodic** if for $\rho \in (0,1)$ and a function $M(\cdot)$, we have for every initialization $\theta \in \Theta$,

$$\left\| P^t(\theta, \cdot) - \Pi \right\|_{\mathsf{TV}} \leq M(\theta)\rho^t$$

and $(\|\cdot\|, p, \rho, M)$-geometrically ergodic if

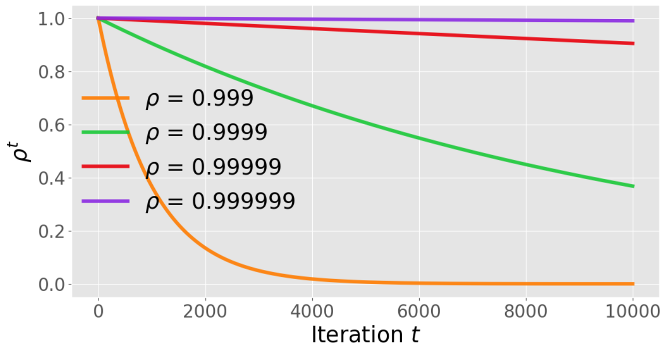
$$\mathcal{W}^p_{\|\cdot\|}(P^t(\theta, \cdot), \Pi) \leq M(\theta)\rho^t.$$

Motivation: Upper bounds on convergence rates in Wasserstein distances tend to scale better in large problem sizes [Hairer et al., 2014, Qin and Hobert, 2019].

Geometric Ergodicity Can be Slow to Converge

Convergence can be **slow** if $\rho \approx 1$.

- Generated samples are **not trustworthy**
- Suggests **unreliable** estimators from the Markov chain



Lower Bounds

Lower bounds on the TV Convergence Rate

Theorem (Theorem 1, 2 [Brown and Jones, 2022])

For any $\theta \in \Theta$

$$\left\| P^t(\theta, \cdot) - \Pi \right\|_{TV} \geq [1 - A(\theta)]^t$$

where $A(\theta) = \int a(\theta, \theta') Q(\theta, d\theta')$. *If* (ρ, M)*-geometrically ergodic so*

$$\left\| P^t(\theta, \cdot) - \Pi \right\|_{TV} \leq M(\theta) \rho^t,$$

then

$$1 - \inf_{\theta \in \Theta} A(\theta) \leq \rho$$

- Method independent (e.g. drift and minorization, coupling)
- Does not require reversiblility

Comparing Algorithms

If P is (ρ, M)-geometrically ergodic and $A(\theta) \leq A_{MH}(\theta)$ (Peskun ordered) where A_{MH} is the version for Metropolis-Hastings, then

$$1 - \inf_{\theta \in \Theta} A_{MH}(\theta) \leq \rho.$$

Lower Bounds for Wasserstein Distances

Theorem (Theorem 4, 5 [Brown and Jones, 2022])

If π is bounded, then there is a $C_{d,\pi} > 0$ so every $\theta \in \Theta$

$$\mathcal{W}^p_{\|\cdot\|}(P^t(\theta, \cdot), \Pi) \geq C_{d,\pi} \left[1 - A(\theta)\right]^{t\left(1+\frac{1}{d}\right)}.$$

If $(\|\cdot\|, p, \rho, M)$-geometrically ergodic so

$$\mathcal{W}^p_{\|\cdot\|}(P^t(\theta, \cdot), \Pi) \leq M(\theta)\rho^t,$$

then

$$1 - \inf_{\theta \in \Theta} A(\theta) \leq \rho^{\frac{d}{d+1}}.$$

- Similar to total variation in high dimensions

The Approach

- **Find problematic point:** The maximum of the target density θ^* can be a problematic for Metropolis-Hastings.
- **Study the computational complexity:** Study how $A(\theta^*) \to 0$ with the problem size d, n.
- **Use lower bounds:** Lower bounds in TV give $1 - \rho \leq A(\theta^*) \to 0$ with the problem size d, n.

Focus on M-H and focus on TV since Wasserstein will be similar.

Applications Under Concentration

Application: RWM for Log-concave targets

Consider $\pi \propto \exp(-f)$ and suppose M-H with RWM proposal $\theta'_t \sim N(\theta_{t-1}, hI_d)$ is (ρ, M)-geometrically ergodic.

Corollary (Corollary 1, 2 [Brown and Jones, 2022])

If $f(\cdot) - \frac{1}{2\xi}\left\|\cdot\right\|_2^2$ is convex on $\mathbb{R}^d$,

$$1 - \rho \leq \frac{1}{(h/\xi + 1)^{d/2}}.$$

- **Need to choose $h \propto \xi/d$ as $d \to \infty$ to avoid $\rho \to 1$**
- **Many examples:** Bayesian GLM with Gaussian priors

Example: Bayesian Logistic Regression with Zellner's g-prior

Consider the posterior Π_n with i.i.d. data $(Y_i, X_i)_i$

$$Y_i | X_i, \beta \sim \text{Bern}\left(\text{sigmoid}(X_i^T \beta)\right) \quad \beta \sim N_d\left(0, g\left(X^T X\right)^{-1}\right)$$

Assume $(X_{i,j})_{i,j}$ are i.i.d. random variables with zero mean, unit variance, and a finite fourth moment.

Suppose M-H is (ρ_n, M_n) geometrically ergodic each n with a RWM proposal $\theta_t' \sim N(\theta_{t-1}, hI_d)$.

Example: Bayesian Logistic Regression with Zellner's g-prior

Proposition (Proposition 5 [Brown and Jones, 2022])

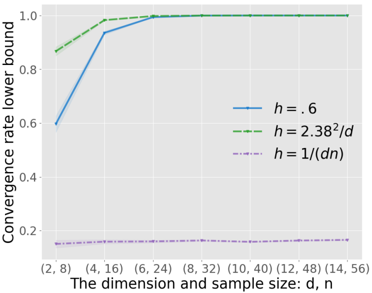
Suppose $n \to \infty$ *with* $d_n/n \to \gamma \in (0,1)$*. Then w.p. 1 and large* n*,*

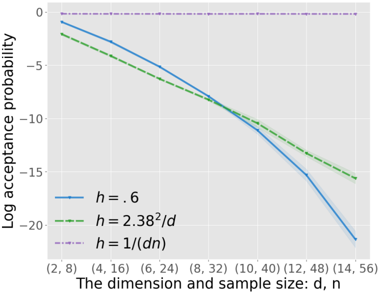
$$1 - \rho_n \leq \frac{1}{\left(\frac{hn(1-\sqrt{\gamma})^2}{2g} + 1\right)^{d_n/2}}.$$

Choose $h \propto 1/(dn)$ **or** $\lim_{n,d_n} \rho_n = 1$ **can be rapid!**

Numerical Simulation

- Generate repeatedly 50 times artificial data with increasing dimensions $d, n = 4d$: $(d, n) \in \{(2,8), ..., (14,56)\}$.
- Use optimization and Monte Carlo to estimate $A(\beta_n^*)$ and lower bounds.
- $h = .6, 2.38^2/d, 1/(dn)$.





More General Lower Bounds Under Concentration

Suppose conditions (roughly speaking) on π_n:

- local λ_0^{-1}-strongly convex condition
- strict maximum
- sufficient tail decay

Suppose M-H with general proposal $\theta_t' \sim N(\mu(\theta_{t-1}), hC)$ (i.e. RWM, MALA, Riemannian manifold MALA [Girolami and Calderhead, 2011]) is (ρ_n, M_n)-geometrically ergodic for each n.

Lower Bounds Under Concentration

Proposition (Proposition 6, 8, [Brown and Jones, 2022])

Under conditions on π_n and $d_n \leq n^\kappa$, $\kappa \in (0, 1)$, then for large (n, d_n),

$$1 - \rho_n \leq \left(\frac{\lambda_0}{nh} \right)^{d_n/2} \frac{2}{\det(C)^{1/2}}. \tag{1}$$

- **For a large class of proposals, if h, C do not depend carefully on n, then $\lim_{(n,d_n)\to\infty} \rho_n = 1$ rapidly!**
- **Similar bound holds for any bounded proposal.**

Application: Flat prior Bayesian logistic regression

Consider π_n for the model with i.i.d. data $(Y_i, X_i)_i$ with $\|X_i\|_2 \leq 1$ w.p. 1

$$Y_i | X_i, \beta \sim \text{Bern}\left(\text{sigmoid}(X_i^T \beta)\right) \qquad \beta \propto 1$$

Suppose M-H with RWM proposal $\theta_t' \sim N(\theta_{t-1}, h I_d)$ is (ρ_n, M_n)-geometrically ergodic.

Theorem (Theorem 3 [Brown and Jones, 2022])

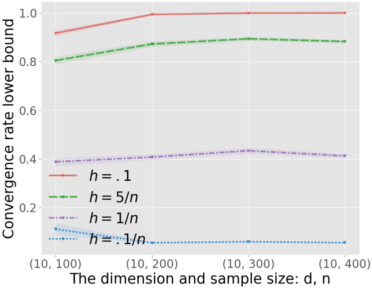
In fixed dimension d and under conditions so the target exists [Chen and Shao, 2000] and the MLE is consistent w.p. 1 and $X_i^T u \neq 0$ if $u \neq 0$. There is a $\lambda_0 > 0$ such that w.p. 1, for large n,

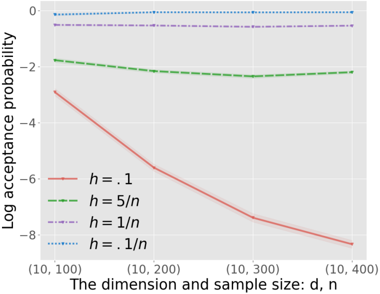
$$1 - \rho_n \leq 2 \left(\frac{\lambda_0}{nh} \right)^{d/2}.$$

Can choose $h \propto 1/n$ to avoid $\lim_n \rho_n = 1$.

Numerical Simulation

- Generate repeatedly 50 times artificial data with $(d, n) =\in \{(10, 100), \ldots, (10, 400)\}$.
- $h = .1, 5/n, 1/n, .1/n$





Comparison to Spectral Methods

Comparison to Spectral Methods

Proposition (Proposition, Proposition 8 [Brown and Jones, 2022])

If P is reversible and there is a $\rho \in (0, 1)$, for every μ with $d\mu/d\Pi \in L^2(\Pi)$, there is a $M_\mu > 0$ such that

$$\mathcal{W}^1_{\|\cdot\| \wedge 1}\left(\mu P^t, \Pi\right) \leq M_\mu \rho^t.$$

If $A(\cdot)$ is upper semicontinuous, then

$$1 - \inf_{\theta \in \Theta} A(\theta) \leq \rho.$$

Based on previous results [Hairer et al., 2014]

Summary

Choose tuning parameters carefully!

Manuscript is on arXiv (submitted to Annals of Statistics):
https://arxiv.org/abs/2212.05955

- Developed **similar general lower bounds** in both total variation and Wasserstein distances in terms of the acceptance probability.
- Studied applications in Bayesian logistic regression for choosing tuning parameters which scale to large problem sizes to avoid the convergence rate rapidly tending to 1.

References I

Christophe Andrieu, Anthony Lee, Sam Power, and Andi Q. Wang. Explicit convergence bounds for metropolis markov chains: isoperimetry, spectral gaps and profiles, 2022.

A. A. Barker. Monte Carlo calculations of the radial distribution functions for a proton-electron plasma. *Australian Journal of Physics*, 18:119–132, 1964.

Alexandre Belloni and Victor Chernozhukov. On the computational complexity of MCMC-based estimators in large samples. *The Annals of Statistics*, 37(4):2011–2055, 2009.

Joris Bierkens. Non-reversible Metropolis-Hastings. *Statistics and Computing*, 26(6):1213–1228, 2015.

Austin Brown and Galin L. Jones. Exact convergence analysis for Metropolis-Hastings independence samplers in Wasserstein distances. *preprint arXiv:2111.10406*, 2021.

References II

Austin Brown and Galin L. Jones. Lower bounds on the rate of convergence for accept-reject-based markov chains. *preprint arXiv*, 2022.

Ming-Hui Chen and Q. Shao. Propriety of posterior distribution for dichotomous quantal response models. *Proceedings of the American Mathematical Society*, 129(1):293 – 302, 2000.

Raaz Dwivedi, Yuansi Chen, Martin J Wainwright, and Bin Yu. Log-concave sampling: Metropolis-Hastings algorithms are fast! In *Proceedings of the 31st Conference On Learning Theory*, volume 75 of *Proceedings of Machine Learning Research*, pages 793–797, 2018.

Karl Oskar Ekvall and Galin L. Jones. Convergence analysis of a collapsed Gibbs sampler for Bayesian vector autoregressions. *Electronic Journal of Statistics*, 15:691 – 721, 2021.

References III

Mark Girolami and Ben Calderhead. Riemann manifold Langevin and Hamiltonian Monte Carlo methods. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 73(2): 123–214, 2011.

Heikki Haario, Eero Saksman, and Johanna Tamminen. An adaptive Metropolis algorithm. *Bernoulli*, 7(2):223 – 242, 2001.

Martin Hairer, Andrew M. Stuart, and Sebastian J. Vollmer. Spectral gaps for a Metropolis–Hastings algorithm in infinite dimensions. *The Annals of Applied Probability*, 24:2455–2490, 2014.

W. K. Hastings. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57(1), 1970.

Søren F. Jarner and Ernst Hansen. Geometric ergodicity of Metropolis algorithms. *Stochastic Processes and their Applications*, 85:341–361, 2000.

References IV

James E. Johndrow, Aaron Smith, Natesh Pillai, and David B. Dunson. MCMC for imbalanced categorical data. *Journal of the American Statistical Association*, 114(527):1394–1403, 2019.

N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller. Equations of state calculations by fast computing machine. *Journal of Chemical Physics*, 21, 1953.

Qian Qin and James P Hobert. Convergence complexity analysis of Albert and Chib's algorithm for Bayesian probit regression. *Annals of Statistics*, 47:2320–2347, 2019.

Qian Qin and James P. Hobert. On the limitations of single-step drift and minorization in Markov chain convergence analysis. *The Annals of Applied Probability*, 31(4):1633 – 1659, 2021.

References V

Bala Rajaratnam and Doug Sparks. MCMC-based inference in the era of big data: A fundamental analysis of the convergence complexity of high-dimensional chains. *preprint arXiv:1508.00947*, 2015.

G. O. Roberts, A. Gelman, and W. R. Gilks. Weak convergence and optimal scaling of random walk metropolis algorithms. *The Annals of Applied Probability*, 7(1):110–120, 1997.

Gareth O. Roberts and Jeffrey S. Rosenthal. Optimal scaling of discrete approximations to langevin diffusions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 60 (1):255–268, 1998.

Gareth O. Roberts and Richard L. Tweedie. Geometric convergence and central limit theorems for multidimensional Hastings and Metropolis algorithms. *Biometrika*, 83:95–110, 1996.

References VI

Luke Tierney. A note on Metropolis-Hastings kernels for general state spaces. *The Annals of Applied Probability*, 8:1–9, 1998.

Guanyang Wang. Exact convergence analysis of the independent Metropolis-Hastings algorithms. *Bernoulli*, 28(3):2012 – 2033, 2022.

Yun Yang, Martin J. Wainwright, and Michael I. Jordan. On the computational complexity of high-dimensional Bayesian variable selection. *Annals of Statistics*, 44:2497–2532, 2016.