Lower Bounds on the Rate of Convergence Metropolis-Hastings in Wasserstein Distances

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

Postdoc at the University of Toronto, Toronto, Canada

October 5, 2023

[1] ad.brown@utoronto.ca

Introduction

Setting

- **High-dimensional** target distribution Π on $\mathbb{R}^d$
- **Large data** of size n (e.g. Bayesian posteriors)
- Lebesgue density $\pi > 0$ on $\Theta \subseteq \mathbb{R}^d$

Simulate a Markov chain for sufficiently long until samples $\theta_t, \ldots, \theta_{t+T-1}$ are from Π (approximately) and

$$\frac{1}{T} \sum_{s=0}^{T-1} f(\theta_{t+s}) \approx \int f d\Pi.$$

Metropolis-Hastings

Generate $\theta_t|\theta_0 = \theta \sim P^t(\theta, \cdot)$ using a proposal $Q(\cdot, \cdot)$, $q(\cdot, \cdot)$ by

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

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



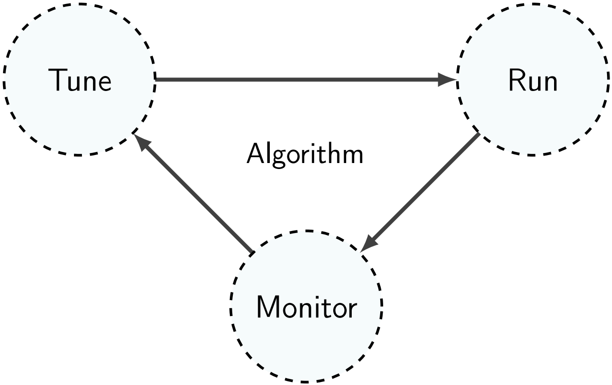
Figure: Arianna Rosenbluth, Nicholas Metropolis, Keith Hastings, and Luke Tierney

Drawbacks

- Requires **choosing** and **tuning** a proposal $Q(\cdot, \cdot)$.
  - Independent proposal
  - Proposals $\theta_t' \sim N(\mu_{h,C}(\theta_{t-1}), hC)$.
    - 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.

Trial and error

- **Problem:** Practitioners often require tuning proposals by trial and error to avoid poor empirical performance.



Drawbacks:

- Computationally intensive and time consuming.

Contribution

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
  - Challenging with limited result (Independence sampler, RWM [Andrieu et al., 2022, Belloni and Chernozhukov, 2009]).

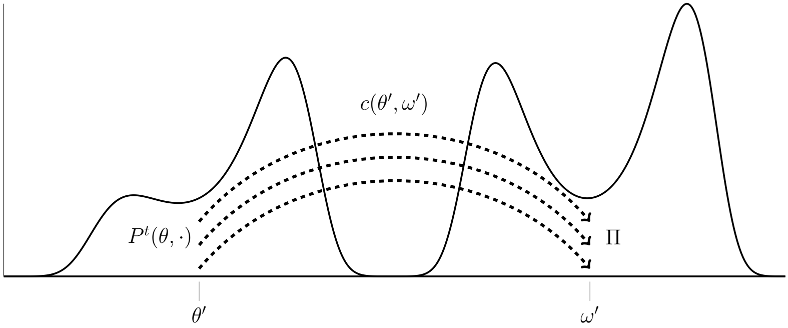
Convergence in Wasserstein Distances

Why use Wasserstein?

We are interested in large problems:

- TV tends to scale poorly to large problems
- Develop lower bounds in Wasserstein distances used for large problems

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 $c(\theta', \omega') = \|\theta' - \omega'\| \wedge 1$.

Transportation distances

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

$$\mathcal{W}_c\left(P^t(\theta,\cdot),\Pi\right) = \inf_{\xi\in\mathcal{C}(P^t(\theta,\cdot),\Pi)} \int c(\theta',\omega')d\xi(\theta',\omega')$$

Figure: Leonid Kantorovich, Leonid Vaseršteĭn, Cédric Villani

Examples of transportation distances

Standard definition:

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

Metrise strong convergence:

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

Metrise weak convergence:

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

Wasserstein geometric ergodicity

Metropolis-Hastings is **Wasserstein geometrically ergodic** if for every $\theta \in \Theta$,

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

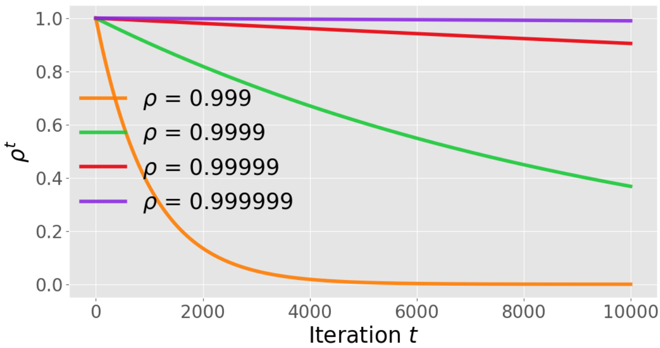
where

- $\rho \in (0, 1)$ (convergence rate)
- $M(\cdot)$ (cost of an imperfect initialisation)

Geometric Ergodicity Can be Slow to Converge

Convergence can be **slow** if the lower bound on ρ is bad i.e. $\rho \approx 1$.

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



Application of lower bounds

- **Lower bounds give a rate function:**
  $r($ { problem size } $,$ { tuning parameters } $)$:
- $1-\rho \leq r($ { problem size } $,$ { tuning parameters }$) \to 0$ with the problem size d, n.

Applications of lower bounds

Use lower bounds on the convergence rates to aid practitioners in understanding which tuning parameters may cause the algorithms to *fail* in practice.

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

Lower Bounds for the Independence Sampler

Convergence rates in TV

Let $\epsilon^* = \inf_\theta q(\theta)/\pi(\theta)$. For every $\theta \in \Theta$,

$$\mathcal{W}_{\|\cdot\|\wedge 1}\left(P^t(\theta,\cdot),\Pi\right) \leq \left\|P^t(\theta,\cdot) - \Pi\right\|_{\mathsf{TV}} \leq (1-\epsilon^*)^t.$$

- Under conditions, the rate is exact rate in total variation [Wang, 2022] and the same for every initialisation θ.

Exact convergence in the Wasserstein distance

Theorem (Proposition 1, Theorem 3, Brown and Jones [2021])

If the point θ^* satisfies $\epsilon^* = q(\theta^*)/\pi(\theta^*)$, then

$$\mathcal{W}_{\|\cdot\|\wedge 1}\left(P^t(\theta^*, \cdot), \Pi\right) = (1 - \epsilon^*)^t \int \|\omega - \theta^*\| \wedge 1 d\Pi(\omega).$$

If π, q are locally Lipschitz continuous and bounded, then for any $\theta \in \Theta$

$$\lim_{t \to \infty} \mathcal{W}_{\|\cdot\|\wedge 1}(P^t(\theta, \cdot), \Pi)^{1/t} = 1 - \epsilon^*.$$

Generalise?

- Acceptance probability for independence sampler:

$$A(\theta) = \mathbb{P}\,(\text{Accept from proposal at } \theta\ )$$

- Exact convergence rate for independence sampler: $1 - A(\theta^*)$
- Lower bound for **general Metropolis-Hastings**? Should be roughly

$$1 - \rho \leq A(\theta^*)$$

General Lower Bounds

Lower bounds on the TV convergence rate

Acceptance probability: $A(\theta) = \int \left[\frac{\pi(\theta')q(\theta',\theta)}{\pi(\theta)q(\theta,\theta')} \wedge 1\right] q(\theta,\theta')d\theta$.

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

For any $\theta \in \Theta$

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

If geometrically ergodic, then

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

- Method independent (e.g. drift and minorisation, coupling)

Lower bounds for Wasserstein distances

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

If π is bounded, then there is a $C_0 > 0$ so every $\theta \in \Theta$

$$\mathcal{W}_{\|\cdot\|}(P^t(\theta, \cdot), \Pi) \geq C_0 \left[1 - A(\theta)\right]^{t\left(1+\frac{1}{d}\right)}.$$

If Wasserstein geometrically ergodic, then

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

- Similar to total variation in high dimensions

Application of lower bounds

- **Use problematic point:** Maximum of π is problematic: $\pi(\theta^*)$ is large
- **Study the computational complexity:** Lower bounds give $1 - \rho^{\frac{d}{d+1}} \leq A(\theta^*) \to 0$ with the problem size d, n
- **Practical estimate:** $A(\theta^*)$ is simple to estimate with Monte Carlo in practice

Applications Under Concentration

Lower bounds under concentration

Use general proposal $\theta_t' \sim N(\mu_{h,C}(\theta_{t-1}), hC)$.

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

Under concentration conditions and Wasserstein geometrically ergodic, then for large (n, d_n)*,*

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

- $\lim_{(n,d_n)\to\infty} \rho_n = 1$ **rapidly if** C, h **do not depend carefully on** n**.**

Flat prior Bayesian logistic regression

Use RWM and consider i.i.d. data $(Y_i, X_i)_i$ and flat prior Bayesian logistic regression.

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

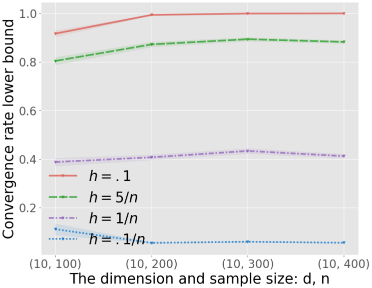
Under technical conditions and in fixed dimension d, w.p. 1, if Wasserstein geometrically ergodic,

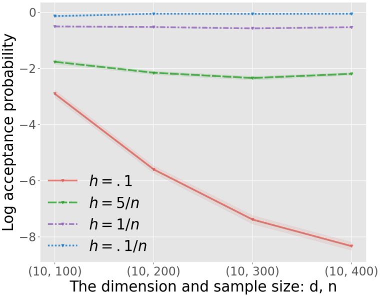
$$1 - \rho_n^{\frac{d}{d+1}} \lesssim \left(\frac{1}{nh}\right)^{d/2}.$$

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

Numerical simulation

- Use Monte Carlo to estimate lower bound
- 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

Let P be any Markov operator on a metric space (Ω, d) reversible with respect to Π.

Proposition (Proposition 8 [Brown and Jones, 2022], [Hairer et al., 2014])

For every $d\mu/d\Pi \in L^2(\Pi)$, there is a $\rho \in (0, 1)$

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

Weak convergence rate $\rho \iff$ TV convergence rate ρ

Spectral method lower bound

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

Initializing at μ and $A(\cdot)$ is upper semicontinuous, then

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

Summary

- Developed lower bounds for Metropolis-Hastings in Wasserstein distances for large problem sizes.
- **Practical applications to tuning Metropolis-Hastings.**
- More examples not presented here.

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.

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.

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

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

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

References II

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.

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.

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