Exact Convergence Analysis for Independence Samplers in Wasserstein Distances

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

Reseach Fellow at the University of Warwick, Coventry, United Kingdom

July 5, 2023

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

General setting

We have a **high-dimensional** target distribution Π on $\mathbb{R}^d$ possibly depending on **large data** of size n (e.g. Bayesian posteriors) with Lebesgue density $\pi > 0$ on $\Theta \subseteq \mathbb{R}^d$.

We want to generate representative samples $\theta_t, \ldots, \theta_{t+T-1}$ from Π to approximate expectations (e.g. predictions and inference in Bayesian statistics)

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

The Metropolis-Hastings independence sampler

With π known up to a normalizing constant, generates $\theta_t \sim P^t(\theta_0, \cdot)$ using a proposal density q by sampling $\theta_t | \theta_{t-1}$

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

where $\theta_t' \sim Q$ and $u_t \sim \text{Unif}(0,1)$.

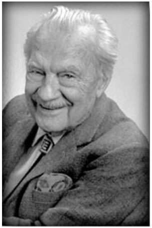



Figure: Arianna Rosenbluth, Nicholas Metropolis, Keith Hastings, and Luke Tierney

Convergence rates in TV

Let $\epsilon^* = \inf_\theta q(\theta)/\pi(\theta)$. The upper bound is known [Tierney, 1994]

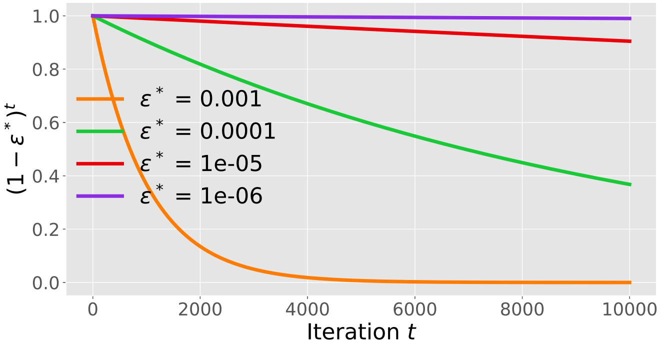
$$\sup_\theta \left\| P^t(\theta, \cdot) - \Pi \right\|_{\mathsf{TV}} \leq (1 - \epsilon^*)^t.$$

- The rate is exact rate in total variation [Wang, 2022] and the same for every initialization θ.
- Only studied a trivial example (exponential distribution).

Geometric Ergodicity Can be Slow to Converge

Convergence can be **slow** if $1 - \epsilon^* \approx 1$.

- Generated samples are **not trustworthy**
- Suggests **unreliable** estimators



Exact convergence rates in Wasserstein distances

Motivation

Wasserstein distances appear to scale better in high dimensions [Hairer et al., 2014, Qin and Hobert, 2021b,a].

Can we find a specific initialization θ and convergence rate $1 - \epsilon_\theta$ which scales in high dimensions / big data problems?

Mathematically:

$$\mathcal{W}_{\|\cdot\| \wedge 1}(P^t(\theta, \cdot), \Pi) \leq (1 - \epsilon_\theta)^t < (1 - \epsilon^*)^t$$

Exact Convergence in the Wasserstein distance $\mathcal{W}_\rho$

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

Let $\epsilon^* = \inf_\theta q(\theta)/\pi(\theta)$*. If* q *is l.s.c. and* π *is u.s.c.,* Θ *is sigma-compact, then*

$$
\begin{aligned}
&(1-\epsilon^*)^t \inf_\theta \int \|\omega - \theta\| \wedge 1 d\Pi(\omega) \\
&\leq \sup_\theta \mathcal{W}_{\|\cdot\| \wedge 1}(P^t(\theta, \cdot), \Pi) \\
&\leq (1-\epsilon^*)^t \sup_\theta \int \|\omega - \theta\| \wedge 1 d\Pi(\omega).
\end{aligned}
$$

- Holds for L_1-Wasserstein distances with lower semicontinuous metric $\rho(\cdot, \cdot) \leq 1$.

Exact Convergence in the Wasserstein distance

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

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

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

- Holds for general L_1-Wasserstein distances with lower semicontinuous metric $\rho(\cdot, \cdot)$.

Convergence rate at every initialization

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

Suppose π, q are locally $\|\cdot\|$-Lipschitz continuous and bounded on $\mathbb{R}^d$ and a point θ^* satisfies $\epsilon^* = q(\theta^*)/\pi(\theta^*)$. Then for any initialization $\theta \in \Theta$, the Wasserstein convergence rate is the same with

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

Applications

Bayesian generalized models

With a Gaussian prior $N(0, \alpha^{-1}C)$, ,consider

- Bayesian logistic and probit regression
- Bayesian negative-binomial regression
- Bayesian Poisson regression

Corollary (Corollary 1 Brown and Jones [2021])

Using a "centered Gaussian proposal" $Q \equiv N_d(\beta^*, \alpha^{-1}C)$ where β^* is the maximum of the posterior density,

$$\mathcal{W}_{\|\cdot\| \wedge 1}\left(P^t(\beta^*, \cdot), \Pi(\cdot|X, Y)\right) = M_0 \left(1 - \epsilon^*\right)^t.$$

where $M_0 = \int \|\beta - \beta^*\| \wedge 1 d\Pi\left(\beta|X, Y\right)$.

- Holds for general L_1-Wasserstein distances.

High dimensions and large data application

High-dimensional Bayesian logistic regression

Assume:

- $Y_i | X_i, \beta \sim$ Bernoulli $\left(S\left(\beta^T X_i\right)\right)$ and $X_i \sim N_d(0, \sigma^2 n^{-1} I_d)$.
- $tr(C) \to s_0$ as $d \to +\infty$.

Theorem (Theorem 4, Corollary 2 Brown and Jones [2021])

if $d, n \to +\infty$ with $d/n \to \gamma \in (0, \infty)$, then almost surely

$$\limsup_{d,n} \mathcal{W}_{\|\cdot\| \wedge 1}\left(P^t(\beta^*, \cdot), \Pi(\cdot | X, Y)\right) \leq M_0(1 - \exp(-a_0))^t$$

where $a_0 > 0$ is known and
$M_0 = \limsup_{d,n} \int \|\beta - \beta^*\| \wedge 1 d\Pi(\beta | X, Y)$.

- Generalizes under technical conditions on the likelihood.
- Holds for general L_1-Wasserstein distances.

Limitations

We observe if $d/n \to \gamma$ is large, the number of iterations needed to approximately converge may still increase rather rapidly!

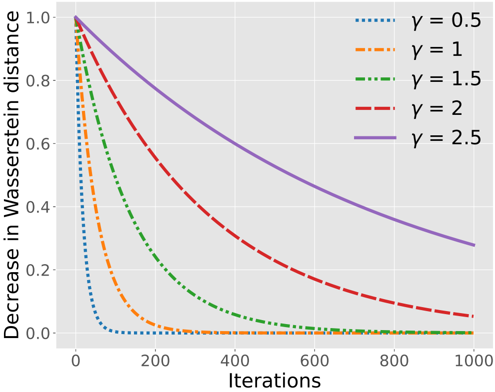


Figure: The limiting decrease in the Wasserstein distance using different values of γ, the limiting ratio of the dimension and sample size, versus the number of iterations.

Summary

- We showed the **exact convergence rate in Wasserstein distances** weaker than total variation matches the convergence rate in total variation for every initialization.
- We showed many non-trivial examples of exact convergence rates in Bayesian statistics.
- Despite this, we showed convergence rates can scale to large problem sizes using a **novel proposal** and **exact convergence analysis**
- First known Metropolis-Hastings algorithm to upper bound the convergence rate when d, n increasing together.

References I

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

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.

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, 2021a.

Qian Qin and James P Hobert. Wasserstein-based methods for convergence complexity analysis of MCMC with applications. To appear in *Annals of Applied Probability*, 2021b.

Luke Tierney. Markov chains for exploring posterior distributions. *The Annals of Statistics*, 22:1701–1728, 1994.

References II

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