Convergence Rate Analysis of Markov Chain Monte Carlo Based on Coarse Ricci Curvature and Its Improved Variant

数理情報学専攻 48-226201 Sho Adachi (足立 勝)
指導教員 Professor Fumiyasu Komaki （駒木 文保 教授）

1 Introduction

The quantification of convergence rates of MCMC is a critical issue. There are two principal metrics in this literature: the total variation distance and the Wasserstein distance. While the convergence rate in the total variation distance has been intensively studied (e.g., [2]), there is much room for developing the convergence rate analysis in the Wasserstein distance, particularly for the Metropolis-Hastings algorithm, which is one of the most widely used classes of MCMC including random-walk Metropolis (RWM), Gibbs sampler, Metropolis Adjusted Langevin Algorithm (MALA), and Hamiltonian Monte Carlo (HMC).

In this thesis, we consider target distributions defined on $\mathbb{R}$ and analyze the convergence rates of various MCMC algorithms (mainly Metropolis-Hastings) in 1-Wasserstein distance by proposing a new quantity. The proposal can be understood as an improved variant of the coarse Ricci curvature, which is a representative quantity for deriving convergence rates in the Wasserstein distance [1].

2 Preliminaries

We let $(\mathcal{X}, d)$ denote a Polish space and $\mathcal{B}(\mathcal{X})$ be the Borel σ-algebra over $\mathcal{X}$. We first state the definition of the 1-Wasserstein distance.

Definition 2.1. *For probability distributions ν_1 and ν_2 on $\mathcal{X}$, the 1-Wasserstein distance between them, $W_1(\nu_1, \nu_2)$, is defined as*

$$W_1(\nu_1,\nu_2) := \inf_{\xi\in\Pi(\nu_1,\nu_2)} \int_{(x,y)\in\mathcal{X}\times\mathcal{X}} d(x,y)\xi(\mathrm{d}x,\mathrm{d}y),$$

where $\Pi(\nu_1, \nu_2)$ denotes the set of couplings of ν_1 and ν_2.

In the special case where $\mathcal{X} = \mathbb{R}$ and d is the Euclidean distance, we can express the distance explicitly using cumulative distribution functions.

Theorem 2.2. *Let ν_1 and ν_2 be probability distributions on $\mathbb{R}$. Then, the following holds:*

$$W_1(\nu_1,\nu_2) = \int_{\mathbb{R}} \left| \int_{-\infty}^{x} \mathrm{d}(\nu_1-\nu_2) \right| \mathrm{d}x.$$

This property of 1-Wasserstein distance plays significantly important role in this thesis.

As previously noted, the coarse Ricci curvature proposed by [1] can quantify the convergence rate w.r.t. W_1. Its definition is as follows.

Definition 2.3. *Let $x, y \in \mathcal{X}$ be two distinct points. The coarse Ricci curvature of a transition kernel $\{m_x\}_{x\in\mathcal{X}}$ along (xy), $\kappa(x, y)$, is defined as*

$$\kappa(x,y) := 1 - \frac{W_1(m_x,m_y)}{d(x,y)}.$$

The coarse Ricci curvature is related to the convergence rate as the following proposition states (See [1, Corollary 21] for its proof):

Proposition 2.4. *For a transition kernel $\{m_x\}_{x\in\mathcal{X}}$, if $\kappa := \inf_{(x,y)\in\mathcal{X}\times\mathcal{X}} \kappa(x, y) > 0$ holds, then $\{m_x\}_{x\in\mathcal{X}}$ has a unique stationary distribution. Moreover, the convergence rate of $\{m_x\}_{x\in\mathcal{X}}$ is $O((1 - \kappa)^n)$ for any initial distribution.*

3 The proposed variant

Let $\{m_x\}_{x\in\mathbb{R}}$ be the transition kernel of a Markov chain on $\mathbb{R}$. In addition, for each $x \in \mathbb{R}$, we let F_x denote the cumulative distribution function of m_x, i.e., $F_x(z) := \int_{-\infty}^{z} m_x(s)\mathrm{d}s$ for $z \in \mathbb{R}$. We introduce the following quantity for Markov chains on $\mathbb{R}$:

$$W(x) := \int_{-\infty}^{\infty} \left| \frac{\partial F_x}{\partial x}(z) \right| \mathrm{d}z \tag{1}$$

This is the proposed variant of the coarse Ricci curvature. The following theorem asserts that if $\sup_{x\in\mathbb{R}} W(x) < 1$, then $\sup_{x\in\mathbb{R}} W(x)$ directly determines the convergence rate of the Markov chain.

Theorem 3.1. *Suppose that $F_x(z)$ is differentiable w.r.t. x for each z. In addition, assume that $\lim_{y\to\infty} F_y(z)$ is a constant which is independent of z. If $\omega := \sup_{x\in\mathbb{R}} W(x) <$*

1, then the convergence rate of the Markov chain with respect to 1-Wasserstein distance is given by $O(\omega^n)$.

Next, we state a result which relates our proposed quantity $W(x)$ to the coarse Ricci curvature.

Theorem 3.2. *If there exists some* $g_x : \mathbb{R} \to \mathbb{R}$ *such that* $\int_{\mathbb{R}} g_x(z)\mathrm{d}z < \infty$ *and* $\frac{|F_{x+\epsilon}(z) - F_x(z)|}{\epsilon} \le g_x(z)$ *for all* $z \in \mathbb{R}$, *then* $\lim_{\epsilon \to 0}(1 - \kappa(x, x+\epsilon)) = W(x)$ *holds.*

Under the assumption in Theorem 3.2, $\sup_{x \in \mathbb{R}} W(x) \le 1 - \inf_{(x,y) \in \mathbb{R}^2} \kappa(x,y)$ holds, and thus the proposal is ensured to derive tighter convergence rates than the coarse Ricci curvature.

4 Examples of the proposed quantity

As notation, we let $\overset{\mathrm{Met}}{W}(x)$ denote the proposed quantity of a Markov chain with the Metropolis test.

4.1 Example 1. (random walk Metropolis)

We give an example where the proposed quantity (1) can derive a convergence rate of the random walk Metropolis (RWM) while the coarse Ricci curvature fails. We define the target distribution π as $\mathcal{N}(0, \sigma^{*2})$ and set the proposal distribution as $m_x = \mathcal{N}(x, \sigma^2)$. In addition, we put an assumption that $\sigma < \sigma^*$, which facilitates our analysis of the quantity (1).

4.1.1 Analysis of $\omega := \sup_{x \in \mathbb{R}} \overset{\mathrm{Met}}{W}(x)$

In this example, we can evaluate $\overset{\mathrm{Met}}{W}(x)$ analytically and Figure 1 is its plot. The fact $\omega < 1$ and Theorem 3.1

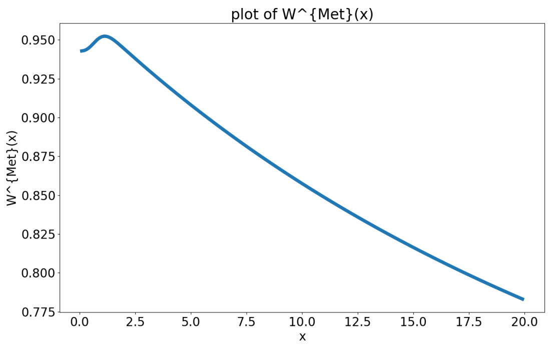

図 1. plot of $\overset{\mathrm{Met}}{W}(x)$ $(x > 0)$

indicates that RWM achieves the exponential convergence.

4.1.2 Analysis of the coarse Ricci curvature

For the coarse Ricci curvature, we can prove $\lim_{x \to \infty}(1 - \kappa(x, -x)) = 0$. As Figure 2 indicates, this fact can be confirmed through numerical experiments too. As a result, we can not apply Proposition 2.4.

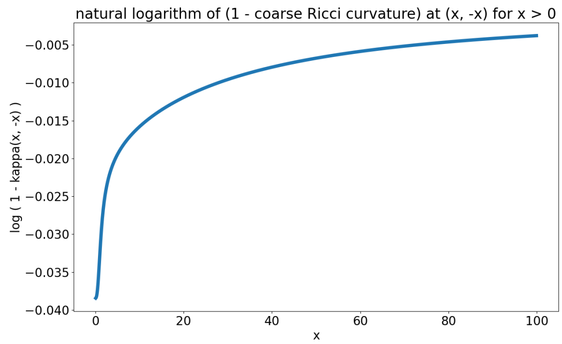

図 2. plot of $1 - \kappa(x, -x)$

4.2 Example 2. (MALA and HMC)

We define the target distribution π as $\pi = \mathcal{N}(0,1)$. In HMC, two parameters are necessary for discretizing Hamilton's equation by LeapFrog integration: the time step width ϵ and the number of LeapFrog integrations N. Figure 3 compares values of the proposed quantity among HMC with different N (Here, ϵN is fixed). We highlight that HMC with $N = 1$ is equivalent to MALA. Figure 3 shows that HMC attains faster convergence than MALA since $\sup_x \overset{\mathrm{Met}}{W}(x)$ of HMC $(N > 1)$ is smaller than that of MALA $(N = 1)$.

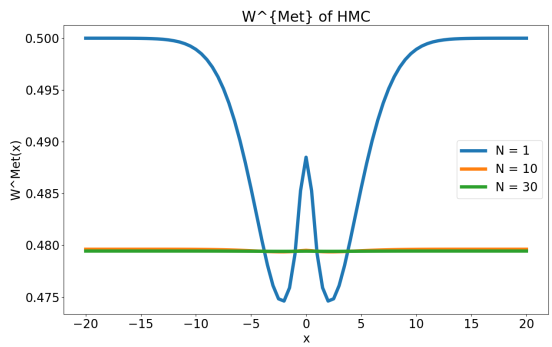

図 3. comparison of $\overset{\mathrm{Met}}{W}(x)$ of HMC with different N

Other examples and the extension of the proposed quantity from $\mathbb{R}$ to other sample spaces will be discussed in the presentation.

参考文献

[1] Y. Ollivier, Ricci curvature of Markov chains on metric spaces, *Journal of Functional Analysis*, 256(3), pp. 810–864, 2009.

[2] J. S. Rosenthal, Minorization conditions and convergence rates for Markov chain Monte Carlo, *Journal of the American Statistical Association*, 90, pp. 558–566, 1995.