Vol. 42 (2022)
No. 2

数 学 杂 志
J. of Math. (PRC)

SAMPLING WITH LANGEVIN DYNAMICS IN NON-CONVEX SETTING

HUANG Jian-hua

(*School of Mathematics and Statistics, Wuhan University, Wuhan 430072, China*)

Abstract: In this work, we investigate the problem of sampling with Langevin diffusion in non-convex setting. Compared with [6], we extend the result under 2-Wassertain distance, and we find a better explicit positive constant c in the exponential convergency.

Keywords: Langevin diffusion; sampling; Wasserstein distance

2010 MR Subject Classification: 60H10; 65C30; 65C05

Document code: A **Article ID:** 0255-7797(2022)02-0121-08

1 Introduction

Langevin diffusion processes are often used to solve the problem of sampling from a given probability distribution π, which has density with respect to the Lebesgue measure on $\mathbb{R}^d$ with its density function satisfying $p^*(x) = \exp\{-U(x)+C\}$, where C is the normalizing constant and $x \in \mathbb{R}^d$. Assuming that we are able to generate an arbitrary number of independent standard Gaussian random variables $\xi_1, \xi_2, ..., \xi_k, k \in \mathbb{N}$. For a given precision level ε and a metric d on the space of probability measure, the goal is to devise a function G_ε such that the distribution ν_k of the random variable $v_k = G_\varepsilon(v_{k-1}, \varepsilon_k)$ is close to π step by step. However, we hope that when $k = K$, the distribution ν_K of v_K will satisfy $d(\nu_K, \pi) < \varepsilon$.

One simple way sampling from π is to consider the first-order Langevin diffusion:

$$dx_t = -\nabla U(x_t)dt + \sqrt{2}dB_t, t \geq 0, \tag{1.1}$$

where B_t is d-dimension Brownian motion. The stationary distribution of the process x_t about SDE(1.1) is π if $\int e^{-U(x)}dx < \infty$. For a step size h, here we have to construct a discrete process of x_t defined as

$$d\tilde{x}_t = -\nabla U(x_{kh})dt + \sqrt{2}dB_t, t \in [kh, (k+1)h), k \in \mathbb{N}. \tag{1.2}$$

For the first-order Langevin diffusion, the update rule associated to this process can be obtained by using the Euler discretization given by the equation as $v_{k+1} = v_k - h\nabla U(v_k) + \sqrt{2h}\xi_k$, where $\xi_k \triangleq (B_{kh} - B_{(k-1)h})/\sqrt{h}$ and $\xi_k \sim \mathcal{N}(0, I_{d\times d})$. In this case, we know $v_{k+1} \sim \mathcal{N}(v_k - h\nabla U(v_k), 2hI_{d\times d})$ which has the same distribution as $\tilde{x}_{(k+1)h}$.

* **Received date:** 2020-12-11 **Accepted date:** 2021-03-09

Biography: Huang Jianhua (1996–), male, postgraduate, born at Dazhou, Sichuan, major in random process.

The approximate sampling algorithm with first-order Langevin diffusion is called overdamped Langevin Monte Carlo (LMC). For the LMC algorithm, there are many results. The first non-asymptotic analysis of the discrete Langevin diffusion (1.2) is due to Dalalyan in [1]. In that paper, the authors use the total variation distance as the metric between v_K and π and prove that the convergence rate of continuous-time process is e^{mt}, and then they get the number of iteration $K \sim O(d/\varepsilon^2)$. Durmus and Moulines also establish convergence under 2-Wasserstein distance in [2] and they get that the number of iteration $K \sim O(d/\varepsilon^2)$, but with the Lipschitz continuity of the Hessian matrix, the number of iteration can be ruduced $K \sim O(d/\varepsilon)$. Then, Cheng, X. and Bartlett establish the first nonasymptotic convergence with Kullback-Leibler divergence in [3]. Based on this, they also unify the proof of convergence in total-variation and 2-Wasserstein distance as simple corollaries. In their research, the number of iteration $K \sim O(d/\varepsilon^2)$ under total-variation and 2-Wasserstein distance, but better bound $K \sim O(d/\varepsilon)$ with KL-divergence.

The problem of sampling from non-logconcave distribution has been studied by Raginsky et al(2017) in [4], and they get a worst-case convergence rate under weaker assumptions, which is exponential in dimension d. Under non-convex situation, Eberle (2016) proves that the convergence rate is exponential at rate $c > 0$ with 1-Waseertsein distance in [5]. Then, under the same assumptions as Eberle, Cheng et al (2019) get that the running step $K \sim O(\frac{d}{\varepsilon^2}e^{cMR^2})$ with overdamped Langevin MCMC and $K \sim O(\frac{\sqrt{d}}{\varepsilon}e^{cMR^2})$ with underdamped Langevin MCMC with 1-Waseertsein distance in [6], where c is a explicit positive constant and MR^2 is a measure of non-convexity.

The main purpose of our work is to pursue the investigation of overdamped Langevin MCMC initiated in Cheng et al (2019) by addressing the following questions. Firstly, what is the convergence rate of continuous-time process if replacing the metric 1-Waseertsein distance by 2-Waseertsein distance? Secondly, After replacing the metric by 2-Waseertsein distance, is it possible to reduce the running step? The rest of this paper is to answer the two questions. For question 1, the answer is true, but we need to do some changes for the auxiliary distance function. For question 2, we find that the convergence rate of continuous-time process can be improved in Proposition 3.1. In Theorem 3.3, we show that the running step is also $O(\frac{d}{\varepsilon^2}e^{cMR^2})$ but with smaller explicit positive constant c.

2 Notations, Definitions and Assumptions

Firstly, we set up the notations to continuous and discrete processes for the overdamped Langevin diffusion. With an initial condition $x_0 \sim p^{(0)}$ for some distribution $p^{(0)}$ on $\mathbb{R}^d$, we let p_t denote the distribution of x_t and let Φ_t denote the operator that maps from $p^{(0)}$ to p_t:

$$\Phi_t p^{(0)} = p_t, \tag{2.1}$$

and we define $\tilde{\Phi}_t$ similarly for the discrete process. When $t = 0$, Φ_0 and $\tilde{\Phi}_0$ are unit operators. Here we follow the same assumptions on the potential function U in Cheng et al (2019).

(A1) The potential function U is continuously differentiable on $\mathbb{R}^d$ and has M-Lipschitz continuous gradients; that is $\|\nabla U(x) - \nabla U(y)\|_2^2 \leq M\|x-y\|_2^2$.

(A2) The potential function U has a stationary point at zero; that is $\nabla U(0) = 0$.

(A3) The potential function U is m-strongly convex outside of a ball with radius R; that is, there exists constants $m, R > 0$ such that for all $x, y \in \mathbb{R}^d$ with $\|x-y\|_2 > R$, we have $\langle \nabla U(x) - \nabla U(y), x-y\rangle \geq m\|x-y\|_2^2$.

We denote by $\mathcal{B}(\mathbb{R}^d)$ the Borel σ-field of $\mathbb{R}^d$. For given probability measures μ and ν on $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$, we define a transference plan ζ between μ and ν on $(\mathbb{R}^d \times \mathbb{R}^d, \mathcal{B}(\mathbb{R}^d \times \mathbb{R}^d))$ such that for all sets $A \in \mathcal{B}(\mathbb{R}^d)$, $\zeta(A \times \mathbb{R}^d) = \mu(A)$ and $\zeta(\mathbb{R}^d \times A) = \nu(A)$. We denote $\Gamma(\mu, \nu)$ as the set of all transference plans. A pair of random variables (X, Y) is called coupling if there exists a $\zeta \in \Gamma(\mu, \nu)$ such that (X, Y) are distributed according to ζ.

Given a function $f : \mathbb{R}^+ \mapsto \mathbb{R}^+$, we define the f-Wasserstein distance between a pair of probability measures as follows:

$$W_f(\mu, \nu) \triangleq \left(\inf_{\zeta \in \Gamma(\mu,\nu)} \int_{\mathbb{R}^d \times \mathbb{R}^d} f(\|x-y\|_2^2) d\zeta(x,y) \right)^{\frac{1}{2}}.$$

Finally we denote by $\Gamma_{opt}(\mu, \nu)$ the set of transference plans that achieve the infimum in the definition of the Wasserstein distance between μ and ν. For any $q \in \mathbb{N}$ we define the q-Wasserstein distance as

$$W_q(\mu, \nu) \triangleq \left(\inf_{\zeta \in \Gamma(\mu,\nu)} \int_{\mathbb{R}^d \times \mathbb{R}^d} \|x-y\|_2^q d\zeta(x,y) \right)^{\frac{1}{q}}.$$

We denote $(W_f(\cdot,\cdot))^q$ by $W_f^q(\cdot,\cdot)$ and $(W_2(\cdot,\cdot))^q$ by $W_2^q(\cdot,\cdot)$.

We begin by defining auxiliary functions $\psi(r), \Psi(r)$ and $g(r)$, all defined $\mathbb{R}^+ \mapsto \mathbb{R}^+$:

$$\psi(r) \triangleq e^{-\alpha_f \min(r, R_f)}, \quad \Psi(r) \triangleq \int_0^r \psi(s) ds, \quad g(r) \triangleq 1 - \frac{1}{2} \frac{\int_0^{\min(r,R_f)} \frac{\Psi(s)}{s} ds}{\int_0^{R_f} \frac{\Psi(s)}{s} ds}.$$

It is worth noting that the function $g(r)$ has no derivative at 0 and R_f, but it has no effect on our analysis. Finally, the distance function f is defined as

$$f(r) \triangleq \int_0^r \psi(s) g(s) ds.$$

Lemma 2.1 The distance function f has the following important properties.

(F1) $f(0) = 0, f'(0) = 1$.

(F2) $\frac{1}{2} e^{-\alpha_f R_f} \leq \frac{1}{2}\psi(r) \leq f'(r) \leq 1$.

(F3) $\frac{1}{2} r e^{-\alpha_f R_f} \leq \frac{1}{2}\Psi(r) \leq f(r) \leq \Psi(r) \leq r$.

(F4) For $0 < r < R_f$, $rf''(r) + \alpha_f r f'(r) \leq -\frac{e^{-\alpha_f R_f}}{R_f} f(r)$.

(F5) For almost everywhere $r > 0$, $f''(r) \leq 0$ and $f''(r) = 0$ when $r > R_f$.

Proof of Lemma 2.1 We refer to the definitions of the functions ψ, Ψ, g, and f.

(F1) It is easily checked by the definition of f, ψ, Ψ and g.

(F2) Noticing that $\frac{1}{2} \leq g(r) \leq 1$ and $\psi_{\max}(r) = \psi(0) = 1$, So (F2) can be easily checked.

(F3) According to the first mean value theorem for integrals, $\Psi(r) = \psi(x)r, x \in [0, r]$, and we know $\psi(x) \geq e^{-\alpha_f R_f}$, so the first inequality holds. The next three inequalities are followed by the notes mentioned in (F2).

(F4) Here we know $g'(r) = -\frac{1}{2}\frac{\Psi(r)}{\int_r^{R_f} \frac{\Psi(s)}{s} ds}$ and $\psi'(r) = -\alpha_f \psi(r)$ if $r \leq R_f$, thus

$$\begin{aligned} rf''(r) + \alpha_f r f'(r) &= r\psi'(r)g(r) + r\psi(r)g'(r) + \alpha_f r f'(r) = r\psi(r)g'(r) \\ &= -\frac{1}{2}\frac{\Psi(r)\psi(r)}{\int_0^{R_f} \frac{\Psi(s)}{s} ds} \overset{(1)}{\leq} -\frac{f(r)\psi(r)}{\int_0^{R_f} \frac{\Psi(s)}{s} ds} \overset{(2)}{\leq} -\frac{e^{-\alpha_f R_f}}{R_f} f(r), \end{aligned}$$

where (1) follows by $f(r) \leq \Psi(r)$, (2) is by $e^{-\alpha_f R_f} \leq \psi(r)$ and $\Psi(s) \leq s$.

(F5) It can be easily checked that $\psi'(r) \leq 0$ and $g'(r) \leq 0$ for almost everywhere $0 < r \leq R_f$, so we have $f''(r) \leq 0$. For $r > R_f, \psi'(r) = g'(r) = 0$, so in that case $f''(r) = 0$.

3 Main Result

The first concerned issue is the convergence rate of the first-order continuous-time process to the invariant distribution. We define the second process as:

$$dy_t = -\nabla U(y_t)dt + \sqrt{2}(I_{d\times d} - 2\gamma_t \gamma_t^\top) dB_t, \tag{3.1}$$

with $y_0 \sim p^*$, where $\gamma_t \triangleq \frac{x_t - y_t}{\|x_t - y_t\|_2} \mathbb{I}[x_t \neq y_t]$. Here $\mathbb{I}[x_t \neq y_t]$ is the indicator function, which is 1 if $x_t \neq y_t$ and 0 otherwise and $\gamma_t^\top$ is the transpose of γ_t. In this case, we couple the processes such that the initial joint distribution of x_0 and y_0 correspond to the optimal coupling between the processes under W_2. To simplify notations, we define $z_t \triangleq x_t - y_t$ and

$$dz_t = -[\nabla U(x_t) - \nabla U(y_t)]dt + 2\sqrt{2}\gamma_t \gamma_t^\top dB_t. \tag{3.2}$$

In (3.2), we define $\nabla_t \triangleq \nabla U(x_t) - \nabla U(y_t)$ and $dB_t^1 \triangleq \gamma_t^\top dB_t$, where dB_t^1 is a one-dimensional Brownian motion by Lévy's characterization.

Proposition 3.1 Let f and W_f be as defined in Section 2 with $\alpha_f = \frac{M}{8}$ and $R_f = R^2$.Then for any $t > 0$, and any probability measure p_0, we have

$$W_f(\Phi_t p_0, p^*) \leq \exp\left\{-e^{-MR^2/8} \min\left\{\frac{8}{R^2}, m\right\} t\right\} W_f(p_0, p^*).$$

Furthermore, we have

$$W_2(\Phi_t p_0, p^*) \leq \sqrt{2}\exp\left\{MR^2/16 - e^{-MR^2/8} \min\left\{\frac{8}{R^2}, m\right\} t\right\} W_2(p_0, p^*),$$

where Φ_t is defined in (2.1) and p^* is the invariant density of x_t in (1.1).

Proof of Proposition 3.1 Firstly, we define $v_t \triangleq \|z_t\|_2^2$ and then by Itô's Formula we have

$$dv_t = d\|z_t\|_2^2 = 2z_t^\top dz_t = -2z_t^\top \nabla_t dt + 4\sqrt{2}z_t^\top \gamma_t dB_t^1.$$

Hence, v_t is a continuous semi-martingale. Applying Itô's formula to $f(v_t)$ again, we have

$$df(v_t) = f'(v_t)dv_t + \frac{1}{2}f''(v_t)d[v,v]_t = f'(v_t)(-2z_t^\top \nabla_t dt + 4\sqrt{2}z_t^\top r_t dB_t^1) + 16v_t f''(v_t)dt,$$

where $[v,v]_t$ denotes the quadratic variation process of v_t. After taking an expectation, then

$$d\mathbb{E}f(v_t) = -2\mathbb{E}[z_t^\top \nabla_t f'(v_t)]dt + 16\mathbb{E}[v_t f''(v_t)]dt. \tag{3.3}$$

Here we have to consider two cases:

Case 1: $v_t < R^2$. In this case, with the smoothness assumption (A3) on $U(x)$ and combining with (3.3), we have

$$d\mathbb{E}f(v_t) \le 2M\mathbb{E}[v_t f'(v_t)]dt + 16\mathbb{E}[v_t f''(v_t)]dt = 16\mathbb{E}\left(\frac{M}{8}v_t f'(v_t) + v_t f''(v_t)\right)dt.$$

By using (F4) of Lemma 2.1 and let $\alpha_f = \frac{M}{8}$ and $R_f = R^2$, we can conclude that

$$d\mathbb{E}f(v_t) \le -\frac{16e^{-MR^2/8}}{R^2}\mathbb{E}f(v_t)dt.$$

Case 2: $v_t \ge R^2$. In this case, we know that for points that are outside of the ball, the potential function satisfies the strongly convex condition. Also, by using (F2) and (F5) of Lemma 2.1, $f''(v_t) = 0$, $f'(v_t) \ge e^{-MR^2/8}$ and $f(v_t) \le v_t$, we get

$$d\mathbb{E}f(v_t) = -2\mathbb{E}(z_t^\top \nabla_t f'(v_t))dt \le -2me^{-MR^2/8}\mathbb{E}(v_t)dt \le -2me^{-MR^2/8}\mathbb{E}f(v_t)dt.$$

Combining the two cases we get that, for any $v_t > 0$,

$$d\mathbb{E}f(v_t) \le -e^{-MR^2/8}\min\left\{\frac{16}{R^2}, 2m\right\}\mathbb{E}f(v_t)dt.$$

So by using Grönwall's inequality, we can get that

$$W_f^2(\Phi_t p^{(0)}, p^*) \le \mathbb{E}f(v_t) \le \exp\left\{-e^{-MR^2/8}\min\left\{\frac{16}{R^2}, 2m\right\}t\right\}\mathbb{E}f(v_0).$$

The first claim is proved by assuming that the initial distributions are optimally coupled under W_f as $\mathbb{E}f(v_0) = W_f^2(p^{(0)}, p^*)$. Then following by (F3) of Lemma 2.1 with the given conditions, for any two measures p and q, we can get

$$\frac{1}{2}e^{-MR^2/8}W_2^2(p,q) \le W_f^2(p,q) \le W_2^2(p,q)$$

as $\frac{1}{2}e^{-MR^2/8}r \le f(r) \le r$. Here let $r = v_t$,Thus,

$$W_2(\Phi_t p^{(0)}, p^*) \le \sqrt{2}\exp\left\{MR^2/16 - e^{-MR^2/8}\min\left\{\frac{8}{R^2}, m\right\}t\right\}W_2(p^{(0)}, p^*).$$

So we can get the two desired claims.

Here we also need to control the discretization error between the continuous and discrete processes. Luckily, this work has been done by Cheng et al (2019) in Proposition 2.3. Now, we quote the conclusion from Cheng et al (2019) for the need of next subsection.

Proposition 3.2 Let the initial distribution $p^{(0)}$ be a Dirac-delta distribution at $x^{(0)}$ with $\|x^{(0)}\|_2 \le R$. Let $p^{(k)}$ be the distribution of $x^{(k)}$. Then, for all $k \in \mathbb{N}$, and step size $h \in [0, \frac{m}{512M^2}]$,

$$\mathbb{E}_{(\tilde{x},x)\sim(\tilde{\Phi}_h p^{(k)}, \Phi_h p^{(k)})}\|\tilde{x} - x\|_2^2 \le M^4h^4\left(80R^2 + \frac{8d}{m}\right) + 2M^2h^3d.$$

The specific proof of Proposition 3.2 can be seen at appendix B in Cheng et al (2019). Based on Proposition 3.1 and Proposition 3.2, we will present the main result as Theorem 3.3.

Theorem 3.3 Let $p^{(0)}$ be the Dirac delta distribution at $x^{(0)}$ with $\|x^{(0)}\|_2 \le R$. Let $p^{(n)}$ denote the distribution of n^{th} iteration of the overdamped Langevin MCMC Algorithm. Let the step size

$$h \le \min\left\{\frac{\varepsilon e^{-3MR^2/16}}{32\max\left\{\frac{R^2}{8}, \frac{1}{m}\right\}M^2\sqrt{\left(160R^2 + \frac{16d}{m}\right)}}, \frac{\varepsilon^2 e^{-3MR^2/8}}{128\max\left\{\frac{R^2}{8}, \frac{1}{m}\right\}^2 M^2 d}\right\}$$

and the number of iteration

$$n \ge \frac{e^{\frac{MR^2}{8}}\log(\frac{\sqrt{320R^2 + \frac{32d}{m}}}{\varepsilon}e^{\frac{MR^2}{16}})\max\left\{\frac{R^2}{8}, \frac{1}{m}\right\}}{\min\left\{\frac{\varepsilon e^{-3MR^2/16}}{32\max\left\{\frac{R^2}{8}, \frac{1}{m}\right\}M^2\sqrt{\left(160R^2 + \frac{16d}{m}\right)}}, \frac{\varepsilon^2 e^{-3MR^2/8}}{128\max\left\{\frac{R^2}{8}, \frac{1}{m}\right\}^2 M^2 d}\right\}}.$$

Then $W_2(p^{(n)}, p^*) \le \varepsilon$.

Proof of Theorem 3.3 From (F3) of Lemma 2.1, we know that for any measures p and q, $W_f^2(p,q) \le W_2^2(p,q)$ as $f(r) \le r$. We have $p^{(0)}(S) = \mathbb{I}(x^{(0)} \in S)$, it is easy to show that for any $i \in \mathbb{N}$ for $h \in [0, \frac{m}{512M^2}]$,

$$W_f^2(\tilde{\Phi}_h p^{(i)}, \Phi_h p^{(i)}) \le W_2^2(\tilde{\Phi}_h p^{(i)}, \Phi_h p^{(i)}) \le M^4h^4\left(80R^2 + \frac{8d}{m}\right) + 2M^2h^3d.$$

By the concavity of f, we have

$$\sqrt{f(\|x - z\|_2^2)} \le \sqrt{2f(\|x - y\|_2^2)} + \sqrt{2f(\|y - z\|_2^2)} \quad x, y, z \in \mathbb{R}^d.$$

Thus, we know

$$W_f(\tilde{\Phi}_h p^{(i)}, p^*) \le \sqrt{2}W_f(\tilde{\Phi}_h p^{(i)}, \Phi_h p^{(i)}) + \sqrt{2}W_f(\Phi_h p^{(i)}, p^*).$$

By using Proposition 3.1 and Proposition 3.2, we can get that

$$W_f(\tilde{\Phi}_h p^{(i)}, p^*) \leq \sqrt{2}\exp\left\{-e^{-MR^2/8}\min\left\{\frac{8}{R^2}, m\right\}h\right\} W_f(p^{(i)}, p^*) + \left(M^4h^4\left(160R^2 + \frac{16d}{m}\right) + 4M^2h^3d\right)^{\frac{1}{2}}.$$

Noticing that the sum of geometric series $1+x+x^2+... = \frac{1}{1-x}$ and $\frac{x}{2} \leq 1-e^{-x}$ for $x \in [0,1]$, so by unrolling last inequality for k steps, we get

$$W_f((\tilde{\Phi}_h)^k p^{(0)}, p^*) \leq \sqrt{2}\exp\left\{-e^{-MR^2/8}\min\left\{\frac{8}{R^2}, m\right\}kh\right\} W_f(p^{(0)}, p^*) + 2e^{MR^2/8}\max\left\{\frac{R^2}{8}, \frac{1}{m}\right\}\left(M^4h^4\left(160R^2 + \frac{16d}{m}\right) + 4M^2h^3d\right)^{\frac{1}{2}}.$$

By (F3) of Lemma 2.1, we know

$$W_2((\tilde{\Phi}_h)^k p^{(0)}, p^*) \leq \sqrt{2}\exp\left\{\frac{MR^2}{16} - e^{-MR^2/8}\min\left\{\frac{8}{R^2}, m\right\}kh\right\} W_2(p^{(0)}, p^*) + 2e^{3MR^2/16}\max\left\{\frac{R^2}{8}, \frac{1}{m}\right\}\left(M^4h^2\left(160R^2 + \frac{16d}{m}\right) + 4M^2hd\right)^{\frac{1}{2}}. \tag{3.4}$$

Here we want the second term in (3.4) smaller than $\frac{\varepsilon}{2}$, so we choose

$$h \leq \min\left\{\frac{\varepsilon e^{-3MR^2/16}}{32\max\left\{\frac{R^2}{8}, \frac{1}{m}\right\}M^2\sqrt{\left(160R^2 + \frac{16d}{m}\right)}}, \frac{\varepsilon^2 e^{-3MR^2/8}}{128\max\left\{\frac{R^2}{8}, \frac{1}{m}\right\}^2 M^2 d}\right\},$$

Then we pick suitable

$$n \geq \frac{e^{\frac{MR^2}{8}}\log(\frac{2\sqrt{2}W_2(p^{(0)}, p^*)}{\varepsilon}e^{\frac{MR^2}{16}})\max\left\{\frac{R^2}{8}, \frac{1}{m}\right\}}{h} \tag{3.5}$$

to make the first term in (3.4) is smaller than $\frac{\varepsilon}{2}$ and that is

$$\sqrt{2}\exp\left\{\frac{MR^2}{16} - e^{-MR^2/8}\min\left\{\frac{8}{R^2}, m\right\}nh\right\} W_2(p^{(0)}, p^*) \leq \frac{\varepsilon}{2}.$$

Finally, we have to control the distance between $p^{(0)}$ and p^*. Here we can upper bound $W_2(p^{(0)}, p^*)$ by using triangle inequality, and that is

$$W_2^2(p^{(0)}, p^*) = \inf_{\zeta \in \Gamma(p^{(0)}, p^*)} \int_{\mathbb{R}^d \times \mathbb{R}^d} \|x-y\|_2^2 d\zeta(x,y) \leq 2\mathbb{E}[\|x\|_2^2] + 2\mathbb{E}[\|y\|_2^2],$$

where $x \sim p^{(0)}$ and $y \sim p^*$. So $\mathbb{E}_{p^{(0)}}\|x\|_2^2 \leq R^2$ and $\mathbb{E}_{p^*}[\|y\|_2^2]$ can be bounded by $(\frac{2d}{m} + 18R^2)$ from Lemma E.3 in Cheng et al (2018), thus,

$$W_2(p^{(0)}, p^*) \leq \left(\frac{4d}{m} + 38R^2\right)^{\frac{1}{2}}.$$

So plugging this into (3.5) can get the desired result.

References

[1] Dalalyan A S. Theoretical guarantees for approximate sampling from smooth and log-concave densities[J]. J. R. Stat. Soc. Ser. B. Stat. Methodol. 2017, 79: 651–676.

[2] Durmus A, Moulines E. High-dimensional Bayesian inference via the unadjusted Langevin algorithm[J]. Bernoulli. 2019, 25: 2854–2882.

[3] Cheng X, Bartlett P L. Convergence of Langevin MCMC in KL-divergence[J]. Proceedings of ALT2018, 2018, 83: 186–211.

[4] Raginsky M, Rakhlin A, Telgarsky M. Non-convex learning via Stochastic Gradient Langevin Dynamics: a nonasymptotic analysis[J]. Conference on Learning Theory, 2017, 1: 1674–1703.

[5] Eberle A. Reflection couplings and contraction rates for diffusions[J]. Probability Theory and Related Fields, 2016, 166: 851–886.

[6] Cheng X, Chatterji N S, Abbasi-Yadkori Y, Bartlett P L, Jordan M I. Sharp convergence rates for langevin dynamics in the nonconvex setting[J]. arXiv:1805.01648v3, 2018.

基于Langevin扩散过程在非凸情况下的抽样

黄建华

(武汉大学数学与统计学院, 湖北 武汉 430072)

摘要: 本文研究了在目标分布非凸情况下, 利用Langevin扩散过程进行抽样的问题, 和文献[6]相对比, 本文推广了其抽样算法在二阶Wasserstein度量下的收敛性, 并且在其指数收敛的情形下找到了一个更好的常数c.

关键词: Langevin扩散过程; 抽样; Wasserstein距离

MR(2010)**主题分类号**: 60H10; 65C30; 65C05 **中图分类号**: O211.63