Ann Inst Stat Math (2009) 61:949–967
DOI 10.1007/s10463-008-0180-6

Metropolis–Hastings Algorithms with acceptance ratios of nearly 1

Kengo Kamatani

Received: 25 August 2005 / Revised: 18 December 2007 / Published online: 24 April 2008

Abstract We develop the results on polynomial ergodicity of Markov chains and apply to the Metropolis–Hastings algorithms based on a Langevin diffusion. When a prescribed distribution p has heavy tails, the Metropolis–Hastings algorithms based on a Langevin diffusion do not converge to p at any geometric rate. However, those Langevin based algorithms behave like the diffusion itself in the tail area, and using this fact, we provide sufficient conditions of a polynomial rate convergence. By the feature in the tail area, our results can be applied to a large class of distributions to which p belongs. Then, we show that the convergence rate can be improved by a transformation. We also prove central limit theorems for those algorithms.

Keywords Metropolis–Hastings algorithm · Polynomial ergodicity · Langevin diffusion · Metropolis adjusted Langevin algorithm

1 Introduction

Various forms of Markov chain Monte Carlo methods are widely used for simulation of a probability density $p(x)dx$ on $(\mathbf{R}^d, \mathcal{B}(\mathbf{R}^d))$, and the Metropolis–Hastings algorithms form a popular sub-class of those.

In order to describe the Metropolis–Hastings algorithms for the *target distribution* p, we first consider a *candidate transition kernel* Q which generates potential transition for a discrete time Markov chain. In this paper, we will assume that there exists a measurable (in both variables) function $q(x, y)$ such that $Q(x, dy) = q(x, y)dy$.

K. Kamatani (✉)
Graduate School of Mathematical Sciences, The University of Tokyo,
3-8-1 Komaba, Meguro-ku, Tokyo 153-0041, Japan
e-mail: kengok@ms.u-tokyo.ac.jp

In the Metropolis–Hastings algorithm, a candidate transition is accepted with probability $\alpha(x, y) = \min\{1, p(y)q(y, x)/(p(x)q(x, y))\}$, otherwise, the jump is rejected and the chain remains its original state. Thus the actual Metropolis–Hastings chain $(M_n^x; n \in \mathbf{N}_0)$ starting from $M_0^x = x$ is defined as follows:

$$\begin{cases} Y_n^x \sim q(M_{n-1}^x, y)\mathrm{d}y \ (n \in \mathbf{N}) \\ M_n^x = \begin{cases} Y_n^x & \text{with probability } \alpha(M_{n-1}^x, Y_n^x) \\ M_{n-1}^x & \text{with probability } 1 - \alpha(M_{n-1}^x, Y_n^x). \end{cases} \end{cases} \tag{1}$$

In this paper, we mainly consider two classes of the Metropolis–Hastings algorithms. One is called "random-walk based", in which

$$q(x, y) = q^*(x - y), \tag{2}$$

where the q^* is a probability density on $\mathbf{R}^d$. The other is called *Metropolis adjusted Langevin algorithm* or simply, *Langevin algorithm* whose candidate transition kernel is

$$Q(x, \mathrm{d}y) \sim \mathrm{N}\left(x + \frac{1}{2}h\nabla \log p(x), h\right), \tag{3}$$

where h is a positive constant, and ∇ denotes the gradient operator. This class is motivated by the Langevin diffusion satisfied by

$$\mathrm{d}X_t = \mathrm{d}B_t + \frac{1}{2}\nabla \log p(X_t)\mathrm{d}t; \quad X_0 = x, \tag{4}$$

for a Brownian motion $(B_t; t \in \mathbf{R}^+)$. The Langevin algorithm and other Langevin diffusion based algorithms are studied in, for example, Grenander and Miller (1994), Roberts and Tweedie (1996b), Stramer and Tweedie (1999a,b) and Roberts and Stramer (2002).

We are concerned with the rate of convergence of these algorithms for a probability density $p(x)\mathrm{d}x$. It is known that the rate of convergence depends on the tail of the distribution $p(x)\mathrm{d}x$ (cf. Mengersen and Tweedie 1996; Roberts and Tweedie 1996a). For example, the tail of p needs to be uniformly exponential for geometric ergodicity for the Metropolis–Hastings algorithms based on random-walk candidate distributions (Theorem 3.3 of Mengersen and Tweedie 1996). The similar statement was proved in Roberts and Tweedie (1996b) for the Langevin algorithm.

In this paper, we assume that p has heavy tails. The Metropolis–Hastings algorithms when p has heavy tails were studied in, for example, Douc et al. (2004) and Fort and Roberts (2005). A significant step in this direction was made by Jarner and Roberts (2002a), which served as a basis of the present study. They showed that the random-walk with Gaussian increment based algorithm and the Langevin algorithm converge at the same polynomial rate to p with heavy tails. Moreover, they showed that the convergence rate of a random-walk based algorithm is improved by using a distribution

with heavier tails. Their results can be validated for a certain class of probability distribution p. The class of functions they considered consists of p that satisfying

$$p(x) = \frac{l(|x|)}{|x|^{\eta}} \quad (|x| \to \infty), \tag{5}$$

with $\eta > d$ where $|\cdot|$ denotes the Euclidean norm and l is a normalized slowly varying function such that $l(x) \to a > 0$ $(x \to \infty)$. Therefore, p should be a symmetric function in the limit. It is not easy to relax the condition. The difficulty comes from the fact that if the target is not symmetric, then the acceptance ratio is difficult to treat.

We show that when p has heavy tails, the behavior of the Langevin algorithm in the tail area in $\mathbf{R}^d$ is similar to that of the Langevin diffusion itself. For this fact, almost all proposal is accepted in the tail area and polynomial rate of convergence of the Langevin algorithm follows from ergodicity of the Langevin diffusion. We do not have to assume technical conditions for p. We only assume that the probability density p is C^2 and

$$\lim_{|x|\to\infty} |\nabla \log p(x)| = 0, \quad \lim_{|x|\to\infty} \|\nabla^T \nabla \log p(x)\| = 0,$$

where $\|(a_{i,j})_{i,j=1,\ldots,d}\| = \left(\sum_{i,j=1}^{d} a_{i,j}^2\right)^{1/2}$ and $\nabla^T f(x)$ denotes the Jacobi matrix for the vector $f(x)$. Then, we propose an algorithm with transformation, which transform heavy tails of p into lighter tails, and by using that we can improve the convergence rate. The convergence rate is the same for the random-walk based algorithm with heavier increment distribution, which is proposed in Jarner and Roberts (2002a), though this convergence for the new algorithm is validated for a wider class of target distributions.

In Sect. 2, we formulate central limit theorems for Markov chains with polynomial ergodicity. Those results are used for concrete examples in Sect. 4. In Sect. 3, which is the main part of this paper, we prove generalized version of a polynomial rate of convergence for the Langevin algorithm. Then we propose an improved algorithm and prove its convergence. In Sect. 4, we demonstrate the efficiency of our methods by numerical calculations.

2 Markov chain and its polynomial ergodicity

Let $(\Omega, \mathcal{F}, \mathbf{P})$ be a probability space and $(E, \mathcal{E})$ a measurable space where $\mathcal{E}$ is a countably generated σ-algebra. Let $(X_n; n \in \mathbf{N}_0)$ be a discrete time Markov chain having state space $(E, \mathcal{E})$. The transition kernel of $(X_n; n \in \mathbf{N}_0)$ is denoted by P:

$$\mathbf{P}(X_n \in A | X_{n-1}) = P(X_{n-1}, A) \text{ a.s.}$$

This transition kernel P can be interpreted as a linear operator on a function space by defining $Pf(x) = \int P(x, \mathrm{d}y) f(y)$. If P_1, P_2 are two kernels, their product $P_1 P_2$

is defined by $(P_1P_2)(x, A) = \int P_1(x, \mathrm{d}y)P_2(y, A)$. The iterates P^n is defined by $P^1 = P$ and $P^n = P^{n-1}P$.

Markov chain will be assumed to be irreducible, aperiodic and positive Harris recurrent; for definitions, see Meyn and Tweedie (1993). Note that for the Metropolis–Hastings algorithms (1), if $p(x)$ and $q(x, y) > 0$ are continuous in both variables, then the Markov chain is $p(x)\mathrm{d}x$-irreducible, aperiodic and any compact set of positive Lebesgue measure is a small set [Lemma 1.2 of Mengersen and Tweedie (1996)]. Hitting time τ_A of a set $A \in \mathcal{E}$ is defined by $\tau_A = \inf\{n \geq 1; X_n \in A\}$. Hitting times of a petite set play an important role in the ergodicity of Markov chain. A subset $\mathcal{E}^+$ of $\mathcal{E}$ is defined by $\mathcal{E}^+ = \{A \in \mathcal{E}; A \text{ has a positive measure by an irreducibility measure}\}$.

Let $V : E \to \mathbf{R}^+$ be an $\mathcal{E}$-measurable function. Let $\|\cdot\|_V$ be a norm over the space of signed measures on $(E, \mathcal{E})$ to $\mathbf{R}$ defined by

$$\|\nu\|_V := \sup_{|f| \leq V} |\nu(f)| \ (\nu : \text{ signed measure}).$$

When $V \equiv 1$, the norm corresponds to the total variation.

Sub-geometric rate of convergence is studied in, for example, by Tuominen and Tweedie (1994), Fort and Moulines (2000), Jarner and Roberts (2002a,b) and Douc et al. (2004). In Jarner and Roberts (2002b), they proved the following theorem.

Theorem 1 (Jarner and Roberts) *Suppose a Markov chain $(X_n; n \in \mathbf{N}_0)$ with transition kernel P is irreducible and aperiodic. Suppose that there exist an $\mathcal{E}$-measurable function $V : E \to [1, \infty)$, constants $c, b > 0$, $0 \leq \gamma < 1$, and a small set C, such that*

$$PV(x) \leq V(x) - cV(x)^\gamma + b1_C(x). \tag{6}$$

Then there exists a probability measure Π and the following polynomial convergence property holds for any $x \in E$ where $1 \leq \beta \leq 1/(1-\gamma)$ and $V_\beta(x) = V(x)^{1-\beta(1-\gamma)}$:

$$(n+1)^{\beta-1}\|P^n(x, \cdot) - \Pi\|_{V_\beta} \to 0. \tag{7}$$

In particular, $\gamma/(1-\gamma)$ is the polynomial order of convergence in total variation norm.

A central limit theorem is said to hold for f if $\Pi(|f|) < \infty$ and there exists $0 < \sigma^2 < \infty$ such that

$$\frac{S_n(\bar{f})}{\sqrt{n}} \xrightarrow{d} N(0, \sigma^2) \ (n \to \infty),$$

where $\bar{f} = f - \Pi(f)$ and $S_n(f) = \sum_{i=1}^n f(X_i)$. We need some lemmas to prove central limit theorems. These lemmas are closely related to Theorem 11.3.9 of Meyn and Tweedie (1993), Proposition 3.1 of Tuominen and Tweedie (1994) and Theorem 3.2 of Jarner and Roberts (2002b). First, lemma is merely a modification of Theorem 3.2 of Jarner and Roberts (2002b).

Lemma 1 *Let $A, C \in \mathcal{E}$, $W_i : E \to [1, \infty)$ $(i = 0, 1, \ldots)$ and $PW_i - W_i \leq W_{i+1} + \beta_i 1_C$, $i = 0, 1, 2, \ldots, k$. Then for any $l = 0, 1, 2, \ldots, k$,*

$$\mathbf{E}_x \left[\sum_{n=0}^{\tau_A - 1} \frac{(n+l)!}{n!} W_{l+1}(X_n) \right] \leq l! W_0(x) + \sum_{m=0}^{l} \frac{(n+m)!}{n!} \beta_m \mathbf{E}_x \left[\sum_{n=0}^{\tau_A - 1} 1_C(X_n) \right]. \tag{8}$$

In particular, if $A, C \in \mathcal{E}^+$ and C is a petite set, then there exists a constant $c < \infty$ such that for any $l = 0, 1, 2, \ldots, k$,

$$\mathbf{E}_x \left[\sum_{n=0}^{\tau_A - 1} \frac{(n+l)!}{n!} W_{l+1}(X_n) \right] \leq l! W_0(x) + c \sum_{m=0}^{l} \frac{(n+m)!}{n!} \beta_m. \tag{9}$$

Proof At the first step, from the assumption $PW_0 - W_i \leq W_1 + \beta_0 1_C$ and using Theorem 11.3.2 of Meyn and Tweedie (1993), we obtain

$$\mathbf{E}_x \left[\sum_{n=0}^{\tau_A - 1} W_1(X_n) \right] \leq W_0(x) + \beta_0 \mathbf{E}_x \left[\sum_{n=0}^{\tau_A - 1} 1_C(X_n) \right]. \tag{10}$$

At the lth step, we have

$$\begin{aligned} &\frac{(n+l)!}{n!} PW^l - \frac{(n+l-1)!}{(n-1)!} W^l \\ &\leq -\frac{(n+l)!}{n!} W^{l+1} + l \frac{(n+l-1)!}{(n-1)!} W^l + \frac{(n+l)!}{n!} \beta_l 1_C. \end{aligned}$$

Then using Theorem 11.3.2 of Meyn and Tweedie (1993), we obtain

$$\begin{aligned} \mathbf{E}_x \left[\sum_{n=0}^{\tau_A - 1} \frac{(n+l)!}{n!} W_{l+1}(X_n) \right] &\leq l \mathbf{E}_x \left[\sum_{n=0}^{\tau_A - 1} \frac{(n+l-1)!}{(n-1)!} W_l(X_n) \right] \\ &\quad + \frac{(n+l)!}{n!} \beta_l \mathbf{E}_x \left[\sum_{n=0}^{\tau_A - 1} 1_C(X_n) \right]. \end{aligned} \tag{11}$$

From this fact, the first claim of the lemma can be obtained easily by using induction. The second claim is $\sup_x \mathbf{E}_x \left[\sum_{n=0}^{\tau_A - 1} 1_C(X_n) \right] < \infty$, which is stated in Theorem 11.3.11 of Meyn and Tweedie (1993). □

Lemma 2 *Let (P, V, γ, C, b, c) satisfy the drift condition (6), $A, C \in \mathcal{E}^+$ and C be a petite set. Then for any $\eta \in (0, 1]$, there exist constants c_1, c_2 such that for any (not necessarily integer) $l \in [0, \eta/(1-\gamma) - 1]$,*

$$\mathbf{E}_x \left[\sum_{n=0}^{\tau_A - 1} (n+1)^l V^{\eta - (l+1)(1-\gamma)}(X_n) \right] \leq c_1 V^{\eta}(x) + c_2. \tag{12}$$

In particular, if we take $l = \eta/(1-\gamma) - 1$, *then we have*

$$\mathbf{E}_x\left[\tau_A^{l+1}\right] \le (l+1)(c_1 V^{\eta}(x) + c_2). \tag{13}$$

Proof From Lemma 3.5 of Jarner and Roberts (2002b), for any integer $k \in [0, \eta/(1-\gamma))$, there exist constants c_k, b_k such that $PV^{\eta-k(1-\gamma)} - V^{\eta-k(1-\gamma)} \le -c_k V^{\eta-(k+1)(1-\gamma)} + b_k 1_C$. Then from the previous lemma, we have for any integer $l \in [0, \eta/(1-\gamma))$,

$$\begin{aligned}\mathbf{E}_x\left[\sum_{n=0}^{\tau_A-1} \frac{(n+l)!}{n!} V^{\eta-(l+1)(1-\gamma)}(X_n) \prod_{k=0}^{l} c_k\right] \\ \le l! V^{\eta}(x) + c\sum_{m=0}^{l} \frac{(n+m)!}{n!} b_{m-1} \prod_{k=0}^{m-1} c_k.\end{aligned}$$

Since $(n+1)^l \le (n+l)!/n!$ we obtain (12) for any integer $l \in [0, \eta/(1-\gamma))$. Next, we consider the equation for any real number $l \in [0, \eta/(1-\gamma)-1]$. For any $t \in [l-1, l)$, we know

$$\left(\frac{n+1}{V(x)^{(1-\gamma)}}\right)^{l-1} + \left(\frac{n+1}{V(x)^{(1-\gamma)}}\right)^{l} \ge \left(\frac{n+1}{V(x)^{(1-\gamma)}}\right)^{t}, \tag{14}$$

hence the claim follows. □

In the following theorem, $L^p = L^p(E, \mathcal{E}, \Pi)$ denotes the space of p-power integrable functions f, $\int |f(x)|^p \Pi(dx) < \infty$.

Theorem 2 *Let* (P, V, γ, C, b, c) *satisfy the drift condition (6), and* $A, C \in \mathcal{E}^+$ *and* C *is a petite set. Then for any* $\eta \ge 1/2$ *such that* $\Pi(V^{\gamma+2\eta-1}) < \infty$, *for any* $\epsilon > (1-\gamma)/(\eta-(1-\gamma))$, *a central limit theorem for the Markov chain holds for any* f *which is in* $L^{2+\epsilon}$ *or* $|f| \le d\, V^{\gamma+\eta-1}$ *where* d *is a positive constant.*

Proof First, we show the measure $\lambda(\mathrm{d}x) = (\Pi I_{|f|})(\mathrm{d}x) = |f(x)|\Pi(\mathrm{d}x)$ is $|f|$-regular for any $f \in L^{2+\epsilon}$ where ϵ is in the above range. If the claim holds, then using Theorem 7.6 of Nummelin (1984), this Markov chain has a central limit theorem.

Consider $f \in L^{2+\epsilon}$. For any $A \in \mathcal{E}^+$, using Hölder's inequality, and for any $p, q > 1$ such that $p^{-1} + q^{-1} = 1$,

$$\begin{aligned}\mathbf{E}_\lambda\left[\sum_{n=0}^{\tau_A-1} |f|(X_n)\right] &= \sum_{n=0}^{\infty} \mathbf{E}_\Pi[|f|(X_0)|f|(X_n)1_{\{n<\tau_A\}}] \\ &\le \sum_{n=0}^{\infty} \|f\|_{L^p} (\mathbf{E}_\Pi[|f|(X_0)^q 1_{\{n<\tau_A\}}])^{\frac{1}{q}}.\end{aligned}$$

Since $1_{\{n<\tau_A\}} \le (\tau_A/n)^r$ for any $r > 1$ and $\beta \in (0, 1]$, we have

$$\begin{aligned}\mathbf{E}_\Pi[|f|(X_0)^q 1_{\{n<\tau_A\}}] &\le \mathbf{E}_\Pi[|f|(X_0)^q \mathbf{E}_\Pi[1_{\{n<\tau_A\}}|\mathcal{F}_0]] \\ &\le \mathbf{E}_\Pi\left[|f|(X_0)^q \mathbf{E}_\Pi\left[\left(\frac{\tau_A}{n}\right)^{\frac{\beta}{1-\gamma}}|\mathcal{F}_0\right]\right] \\ &\le \beta n^{-\frac{\beta}{1-\gamma}} \mathbf{E}_\Pi[|f|(X_0)^q (c_1 V^\beta(X_0) + c_2)].\end{aligned}$$

Then for any $p', q' > 1$ such that $p'^{-1} + q'^{-1} = 1$,

$$\mathbf{E}_\Pi[|f|(X_0)^q V^\beta(X_0)] \le \|f\|_{L^{p'q}} (\mathbf{E}_\Pi[V^{\beta q'}(X_0)])^{\frac{1}{qq'}}.$$

Sufficient conditions for $\mathbf{E}_\lambda[\sum_{n=0}^{\tau_A-1} |f|(X_n)] < \infty$ are $f \in L^p = L^{p'q}$, $\beta q' = \gamma + 2\eta - 1$ and $q(1-\gamma) < \beta$. We can find β and p, q, p', q' which satisfy the above sufficient conditions for any $f \in L^{2+\epsilon}$. Hence the claim follows.

The case of $|f| \le d\ V^{\gamma+\eta-1}$ is quite similar. The only difference is the last inequality. In this case, we do not have to use Hölder's inequality but the inequality $|f| \le d\ V^{\gamma+\eta-1}$. □

If a Markov chain is geometrically ergodic, then integrability condition of the drift function like the above is not necessary. However, in sub-geometric case, we need it. Since, we know $\Pi(V^\gamma) < \infty$ from the drift condition, $\eta = 1/2$ requires no assumption for the integrability of V. In the case of $\eta = 1/2$, central limit theorems for the Markov chain are already showed in Theorem 9 of Jones (2004), which uses a mixing theory. Recently, while editing the galley proof, we found that a result similar to the latter part of Theorem 2 was noted in Corollary 1 of Jarner and Roberts (2007).

Markov chain is said to be reversible when $\Pi(\mathrm{d}x)P(x, \mathrm{d}y) = \Pi(\mathrm{d}y)P(y, \mathrm{d}x)$. Metropolis–Hastings chain is reversible. We can show a slight extension of the above result when the Markov chain is reversible.

Theorem 3 *Let (P, V, γ, C, b, c) satisfy the drift condition (6), and $A, C \in \mathcal{E}^+$ and C be a petite set. Further, we assume that the Markov chain is reversible. Then for any $\eta \ge 1/2$ such that $\Pi(V^{\gamma+2\eta-1}) < \infty$, and for $\epsilon = (1-\gamma)/(\eta - (1-\gamma))$, the Markov chain has a central limit theorem for any $f \in L^{2+\epsilon}$.*

Proof The proof of the theorem uses the same argument as above. Since the Markov chain is reversible, we have

$$\begin{aligned}\mathbf{E}_\lambda\left[\sum_{n=0}^{\tau_A-1} |f|(X_n)\right] &\le \sum_{n=0}^{\infty} \mathbf{E}_\Pi[1_{\{n<\tau_A\}}|f|(X_0)^2]^{\frac{1}{2}} \mathbf{E}_\Pi[1_{\{n<\tau_A\}}|f|(X_n)^2]^{\frac{1}{2}} \\ &= \sum_{n=0}^{\infty} \mathbf{E}_\Pi[1_{\{n<\tau_A\}}|f|(X_0)^2] = \mathbf{E}_\Pi[\tau_A|f|(X_0)^2].\end{aligned}$$

Using Lemma 2, and the Schwarzs inequality, the claim follows. □

3 Algorithm and main theorems

3.1 Langevin algorithms

Let $(\Omega, \mathcal{F}, (\mathcal{F}_t)_{t\geq 0}, \mathbf{P})$ be a filtered probability space. Let $p : \mathbf{R}^d \to \mathbf{R}$ be a strictly positive C^1 function and consider the stochastic differential equation (4). Under certain conditions, there exists a unique solution to the stochastic differential equation and the solution has an invariant measure $p(x)\mathrm{d}x$. Let $(Y_n^x; n \in \mathbf{N}_0)$ be an Euler–Maruyama discretization of $(X_t^x; t \in \mathbf{R}^+)$, that is

$$Y_n^x = \sqrt{h}W_n + hb(Y_{n-1}^x); \quad Y_0^x = x, \tag{15}$$

where $W_n := h^{-1/2}(B_{hn} - B_{h(n-1)})$. Roberts and Tweedie (1996b) proved that if $|\nabla \log p(x)| \to 0$ $(|x| \to \infty)$ then the Langevin algorithm does not converge at geometric rate (Theorem 4.2). We are going to prove its polynomial rate of convergence. First, we show polynomial ergodicity for this Markov chain, the candidate chain of the Langevin algorithm.

Theorem 4 *Let $p : \mathbf{R}^d \to \mathbf{R}$ be a C^1 function. Suppose there exists $\eta > d$,*

$$\limsup_{|x|\to\infty} x^T \cdot \nabla \log p(x) \leq -\eta, \quad \lim_{|x|\to\infty} |\nabla \log p(x)| = 0. \tag{16}$$

Then the Euler–Maruyama discretization $(Y_n^x; n \in \mathbf{N}_0)$ satisfies the drift condition (6) for any $h > 0$, $2 < s < 2 + \eta - d$, $V(x) = (|x|^2 + 1)^{s/2}$, $\gamma = (s-2)/s$ and a compact set C of positive Lebesgue measure. In particular, the upper bound of the polynomial convergence rate of the total variation norm is $(\eta - d)/2$.

Proof It is enough to show

$$\limsup_{|x|\to\infty} \frac{PV(x) - V(x)}{V(x)^\gamma} < 0, \tag{17}$$

since $C = \{|x| \leq N\}$ is a small set for any $N > 0$. Let $(X_t^x, t \in [0,1])$ be a stochastic process satisfying $dX_t^x = dB_t + b(x)\mathrm{d}t$, where B_t is a standard Brownian motion. Then $\mathcal{L}(X_h) = \mathcal{L}(Y_1^x)$ and

$$\begin{aligned} PV(x) - V(x) &= \mathbf{E}[V(X_h^x) - V(x)] \\ &= \mathbf{E}\left[\int_0^h \sum_{i=1}^d \frac{\partial V}{\partial x_i}(X_t^x)\mathrm{d}X_t^{x,i} + \frac{1}{2}\frac{\partial^2 V}{\partial x_i \partial x_j}(X_t^x)\mathrm{d}\langle X^{x,i}, X^{x,j}\rangle_t\right] \\ &= \frac{sh}{2}\mathbf{E}\left[\int_0^h (|X_t^x|^2+1)^{\frac{s}{2}-1}\left(2\sum_{i=1}^d X_t^{x,i} b^i(x) + s - 2 + d\right)\right. \\ &\quad \left. - (|X_t^x|^2+1)^{\frac{s}{2}-2}\mathrm{d}t\right]. \end{aligned}$$

Since $X_t^x = x + B_t + tb(x)$, after some calculations such as $\limsup \mathbf{E}[(|X_t|^2+1)^n] \cdot |x|^{-2n} \le 1$, we have

$$\limsup_{|x|\to\infty} \frac{PV(x) - V(x)}{V(x)^\gamma} = \limsup_{|x|\to\infty} \frac{sh}{2}\left(2\sum_{i=1}^d x_i b^i(x) + s - 2 + d\right) \le \frac{sh}{2}(-\eta + s - 2 + d).$$

When $2 < s < 2 + \eta - d$, $\limsup_{|x|\to\infty}(PV(x) - V(x))/V(x)^\gamma < 0$ by the above inequality. □

Fort and Roberts (2005) already showed polynomial ergodicity of a tempered Langevin diffusion. Their results are more general than our results though they consider continuous stochastic processes but Markov chains. Roughly speaking, our theorem corresponds to the discretization of Theorem 16 of Fort and Roberts (2005) when a parameter $d = 0$ in a sense of the rate of convergence in $\|\|_f$-norm.

Next, we show the convergence of the Langevin algorithm. Let $(M_n^x; n \in \mathbf{N}_0)$ be the Metropolis–Hastings chain of the Langevin algorithm starting from $M_0^x = x$, that is:

$$\begin{cases} Y_n^x &= M_{n-1}^x + \sqrt{h}W_n + hb(M_{n-1}^x) \\ M_n^x &= \begin{cases} Y_n^x & \text{with probability } \alpha(M_{n-1}^x, Y_n^x) \\ M_{n-1}^x & \text{with probability } 1 - \alpha(M_{n-1}^x, Y_n^x). \end{cases} \end{cases} \tag{18}$$

where $b = \nabla \log p(x)/2$ and $q(x, y)$ is the density of the transition kernel (3), that is:

$$q(x, y) = \frac{1}{(2h\pi)^{\frac{d}{2}}} \exp\left(-\frac{|y - x - hb(x)|^2}{2h}\right). \tag{19}$$

This Langevin algorithm does not have geometrical ergodicity but polynomial ergodicity.

Theorem 5 *Let $p : \mathbf{R}^d \to \mathbf{R}$ be a C^2 function satisfying (16) and*

$$\lim_{|x|\to\infty} \|\nabla^T \nabla \log p(x)\| = 0. \tag{20}$$

Then the Metropolis–Hastings chain of the Langevin algorithm $(M_n^x; n \in \mathbf{N}_0)$ satisfies the drift condition (6) for $2 < s < 2 + \eta - d$, $V(x) = (|x|^2+1)^{s/2}$, $\gamma = (s-2)/s$ and a compact set C. In particular, the upper bound of the polynomial convergence rate for the total variation norm is $(\eta - d)/2$.

Proof We know by Theorem 4, there exist constants $c < 1$, $b > 0$ and a compact set C of positive Lebesgue measure satisfying

$$\begin{aligned} PV(x) &= \mathbf{E}[V(X_h^x)\alpha(x, X_h^x)] + \mathbf{E}[V(x)(1 - \alpha(x, X_h^x))] \\ &= \mathbf{E}[V(X_h^x)] - \mathbf{E}[(V(X_h^x) - V(x))(1 - \alpha(x, X_h^x))] \\ &\le V(x) - cV(x)^\gamma + b1_C(x) - \mathbf{E}[(V(X_h^x) - V(x))(1 - \alpha(x, X_h^x))], \end{aligned}$$

where $\mathrm{d}X_t^x = \mathrm{d}B_t + b(x)\mathrm{d}t$. Hence, it is enough to show $\lim_{|x|\to\infty} |\mathbf{E}[(V(Y_1^x) - V(x))(1-\alpha(x, X_h^x))]|/V(x)^\gamma = 0$ when $\gamma = (s-2)/s$. By the Schwarzs inequality,

$$\begin{aligned}&\mathbf{E}[(V(X_h^x) - V(x))(1-\alpha(x, X_h^x))]| \\ &\quad \le \mathbf{E}[(V(X_h^x) - V(x))^2]^{\frac{1}{2}}\mathbf{E}[(1-\alpha(x, X_h^x))^2]^{\frac{1}{2}} \end{aligned} \tag{21}$$

Since the first term, $\limsup \mathbf{E}[(V(X_h^x) - V(x))^2]^{\frac{1}{2}} V(x)^{-\gamma} \le 1$, we will check $\lim \mathbf{E}[(1-\alpha(x, X_h^x))^2]^{\frac{1}{2}} = 0$. Let $\beta(x, y) = p(y)q(y, x)/(p(x)q(x, y))$, then

$$\begin{aligned}\mathbf{E}[(1-\alpha(x, X_h^x))^2] &\le \mathbf{E}[(1-\beta(x, X_h^x))^2] \le \mathbf{E}[\log \beta(x, X_h^x)^2] \\ &= \mathbf{E}[(\log p(X_h^x) - \log p(x) + \log q(X_h^x, x) - \log q(x, X_h^x))^2] \\ &= \mathbf{E}[(\log p(X_h^x) - \log p(x) - (b(x) + b(X_h^x))^T (X_h^x - x) \\ &\quad -\frac{h}{2}(|b(X_h^x)|^2 - |b(x)|^2))^2].\end{aligned}$$

It is easy to check $\lim \mathbf{E}[((b(x) - b(X_h^x))^T (X_h^x - x))^2] = 0$ and $\lim \mathbf{E}[(|b(X_h^x)|^2 - |b(x)|^2)^2] = 0$. The remainder of the above is

$$\begin{aligned}&\mathbf{E}[(\log p(X_h^x) - \log p(x) - 2b(x)^T (X_h^x - x))^2] \\ &= \mathbf{E}\left[\left(\int_0^h \sum_{i=1}^d \left(\frac{\partial \log p}{\partial x_i}(X_t^x) - \frac{\partial \log p}{\partial x_i}(x)\right)\mathrm{d}X_t^{x,i} + \frac{1}{2}\frac{\partial^2 \log p}{\partial x_i^2}(X_t^x)\mathrm{d}t\right)^2\right],\end{aligned}$$

and the main part of the above equation is

$$\begin{aligned}&\mathbf{E}\left[\left(\int_0^h \sum_{i=1}^d \left(\frac{\partial \log p}{\partial x_i}(X_t^x) - \frac{\partial \log p}{\partial x_i}(x)\right)\mathrm{d}W_t^i\right)^2\right] \\ &= \mathbf{E}\left[\int_0^h \sum_{i=1}^d \left(\frac{\partial \log p}{\partial x_i}(X_t^x) - \frac{\partial \log p}{\partial x_i}(x)\right)^2 \mathrm{d}t\right] \\ &\le \mathbf{E}\left[\sup_{0\le t\le h} \sum_{i=1}^d \left(\frac{\partial \log p}{\partial x_i}(X_t^x) - \frac{\partial \log p}{\partial x_i}(x)\right)^2\right].\end{aligned}$$

We want to show the last term of the above goes to 0 if $|x| \to \infty$. For any $\epsilon > 0$, there exist $\delta_1, \delta_2, \delta_3 > 0$ such that

$$\begin{aligned}&\|\nabla^T \nabla \log p(x)\|^2 < \frac{\epsilon}{2d^3 C_h} \ (|x| \ge \delta_1) \\ &h|b(x)| \le \delta_1 \wedge 1 \ (|x| \ge \delta_2) \\ &\sup_{\xi\in\mathbf{R}^d} |\nabla \log p(\xi)|^2 4d\mathbf{P}(\sup_{0\le t\le h} |B_t| > \delta_3) < \frac{\epsilon}{2},\end{aligned}$$

where $C_h = \mathbf{E}[\sup_{0\le t\le h}(|W_t|+1)^2]$ which is a bounded constant by Doob's inequality. Let $|x| > \delta_1 + \delta_2 + \delta_3$, then we divide the term into two parts,

$$\mathbf{E}\left[\sup_{0\le t\le h}\sum_{i=1}^{d}\left(\frac{\partial \log p}{\partial x_i}(X_t^x) - \frac{\partial \log p}{\partial x_i}(x)\right)^2\left(1_{\{\sup_{0\le t\le h}|B_t|>\delta_3\}}+1_{\{\sup_{0\le t\le h}|B_t|\le\delta_3\}}\right)\right].$$

The first term is bounded above by $4d\sup_{\xi\in\mathbf{R}^d}\|\nabla \log p(\xi)\|^2\mathbf{P}(\sup_{0\le t\le h}|B_t| > \delta_3) \le \epsilon/2$. By Taylor's expansion, the second term is bounded above by

$$\mathbf{E}\left[\sup_{0\le t\le h}\sum_{i=1}^{d}\left(\sum_{j=1}^{d}\sup_{|\xi|>\delta_1}\left|\frac{\partial^2 \log p}{\partial x_i \partial x_j}(\xi)\right||X_t^x - x|\right)^2\right] \le \frac{\epsilon}{2C_h}\mathbf{E}[\sup_{0\le t\le h}|X_t^x - x|^2]$$
$$\le \frac{\epsilon}{2C_h}\mathbf{E}[\sup_{0\le t\le h}(|W_t| + t|b(x)|)^2] \le \frac{\epsilon}{2}.$$

Hence $\mathbf{E}[\sup_{0\le t\le h}\sum_{i=1}^{d}(\frac{\partial \log p}{\partial x_i}(X_t^x) - \frac{\partial \log p}{\partial x_i}(x))^2]$ goes to 0. □

When $d = 1$ and the target distribution can be written in the form $p(x) = C|x|^{-\eta}$ when $|x|$ is large enough, Jarner and Roberts (2002a) have already proved the same result. Moreover, the proof of Jarner and Roberts (2002a) is the basis of the proof of Theorem 5, though the assumptions of Theorem 5 is more general.

When $d > 1$, Jarner and Roberts (2002a) have also proved that the random-walk based Metropolis–Hastings algorithms have the same order of convergence as the Langevin algorithm when the increment distributions of the random-walks have light tails (Proposition 3.5 in Jarner and Roberts 2002a). The random-walk based algorithms are simpler than the Langevin algorithm in the sense of computer calculation, it is better to use the former algorithms if the convergence theorem can be validated for a wide class of target distributions p. However their results for random walk based algorithms are validated for a smaller class of target distributions. They assumed our assumptions and a roundness property about $A(x) = \{y; p(x) \le p(y)\}$ and $A(x)$ should be a convex set when $|x|$ is large enough in their paper. For example, the two-dimensional probability distribution function $p(x, y) \propto (x^4 + y^2 + 1)^{-1}$ does not satisfy the extra properties. This distribution function satisfies (16), (20) and $\eta = 2$, but $A(x)$ is not a convex set. In fact, the distribution satisfies $\limsup |x| \cdot |\nabla \log \pi(x)| < \infty$ and $\limsup |x|^2 \cdot \|\nabla^T \nabla \log \pi(x)\| < \infty$.

Many probability distributions which have heavy tails satisfy property (16), (20). For example, Student's t distribution satisfies the properties.

Example 1 (*Multivariate Student's t distribution*) Consider following d-dimensional Student's t distribution with $m > 0$ degrees of freedom,

$$p(x) = \frac{\Gamma\left(\frac{m+d}{2}\right)}{\Gamma(\frac{m}{2})(m\pi)^{\frac{d}{2}}}(\det \Sigma)^{-\frac{1}{2}}\left(1 + \frac{(x-\mu)^T\Sigma^{-1}(x-\mu)}{m}\right)^{-\frac{m+d}{2}}. \quad (22)$$

It satisfies $\limsup_{|x|\to\infty} |x| \cdot |\nabla \log p(x)| < \infty$, $\limsup_{|x|\to\infty} x^T \cdot \nabla \log p(x) \leq -(m+d)$ and $\limsup_{|x|\to\infty} |x|^2 \|\nabla^T \nabla \log p(x)\| < \infty$. The proof uses the fact that is a positive definite symmetric matrix, there exists $\lambda > 0$ such that $\lambda|x|^2 \leq x^T \Sigma^{-1} x$. By Theorem 3, the Langevin algorithm with proposal p has a central limit theorem for $L^{2+\epsilon}$ with $\epsilon > 4/(m-2)$.

Example 2 (An example which does not satisfy (16)) Consider the following probability distribution function:

$$p(x) \propto \prod_{i=1}^{d} \frac{1}{1+x_i^2} \quad (x = (x_1, \ldots, x_d) \in \mathbf{R}^d).$$

This function satisfy the left hand side of (16) but right hand side of it . Since

$$|\nabla \log p(x)| = \left(\sum_{i=1}^{d} \left(\frac{2x_i}{1+x_i^2} \right)^2 \right)^{\frac{1}{2}},$$

if we take $x = (0, t, \ldots, t)$ and $t \to \infty$, then $|\nabla \log p(x)| \to 2$.

3.2 Transformed Langevin algorithm

We introduce a transformation of a Markov chain $(M_n^x; n \in \mathbf{N}_0)$ by a function $F : \mathbf{R}^d \to \mathbf{R}^d$. Suppose there is a C^2 function $f : \mathbf{R} \to \mathbf{R}$ which holds $f'(x) > 0$, $f(0) = 0$ and $\lim_{x\to 0} f(x)/x \neq 0$ such that

$$F(x) = \begin{cases} f(|x|)\frac{x}{|x|} & \text{if } x \neq 0 \\ 0 & \text{if } x = 0. \end{cases}$$

Then F is a C^2 function with $\det \nabla^T F(x) > 0$. Under certain conditions, if a Markov chain $(M_n^x; n \in \mathbf{N}_0)$ with an invariant measure $p^*(x)dx := p(F(x)) \det \nabla^T F(x) \mathrm{d}x$ satisfies (6), then $(F(M_n^x); n \in \mathbf{N}_0)$ has an invariant measure $p(x)\mathrm{d}x$ and satisfies (6) (Proposition 1).

Let p be a d-dimensional probability distribution function and and $V : \mathbf{R}^d \to [0, \infty)$ be a norm-like function, that is for any $r > 0$, $\{x; V(x) \leq r\}$ is a relatively compact set.

Proposition 1 *Let $|F^{-1}(x)|$ be a norm-like function. Let $p(x) > 0$ be a C^1 function and $Q^*(x, \mathrm{d}y) = q^*(x, y)\mathrm{d}y$ be a transition kernel where $q^*(x, y) > 0$ is continuous in both variables. Let $(M_n^*; n \in \mathbf{N}_0)$ be a Metropolis–Hastings chain with a candidate kernel Q^* and an invariant measure $p^*(x)\mathrm{d}x$. Suppose there exist a compact set C^* with positive Lebesgue measure, a function $V^* : \mathbf{R}^d \to [1, \infty)$ and constants $0 \leq \gamma \leq 1$, $b, c > 0$ such that the drift condition (6) holds. Then for $(M_n = F(M_n^*); n \in \mathbf{N}_0)$, there exist constants γ, b, c, a compact set $C \supset C^*$ with positive Lebesgue measure such that the drift condition (6) for C, $V = V^* \circ F^{-1}$.*

Proof Denote the transition kernel of $(M_n^*; n \in \mathbf{N}_0)$ by P^* and that of $(M_n; n \in \mathbf{N}_0)$ by P. First, we show that $(M_n; n \in \mathbf{N}_0)$ is a Metropolis–Hastings chain with the candidate kernel

$$Q(x, \mathrm{d}y) = q(x, y)\mathrm{d}y := q^*(F^{-1}(x), \quad F^{-1}(y)) \det \nabla^T F^{-1}(y)\mathrm{d}y,$$

and the invariant probability measure $p(x)\mathrm{d}x$. Let $(Y_n^*; n \in \mathbf{N}_0)$ be a candidate chain of $(M_n^*; n \in \mathbf{N}_0)$ and denote the acceptance ratio for the Metropolis–Hastings chain by α^*. Let $Y_n := F(Y_n^*)$, then

$$\mathbf{P}(Y_n \in A | Y_{n-1}) = \int_{F^{-1}(A)} q^*(Y_{n-1}^*, y)\mathrm{d}y = \int_A q(Y_{n-1}, y)\mathrm{d}y,$$

hence, Q is its transition kernel. Let $\alpha(x, y) = 1 \wedge p(y)q(y, x)/(p(x)q(x, y))$ then

$$\alpha^*(x, y) = 1 \wedge \frac{p^*(y)q^*(y, x)}{p^*(x)q^*(x, y)} = 1 \wedge \frac{p(F(y))q(F(y), F(x))}{p(F(x))q(F(x), F(y))} = \alpha(F(x), F(y)),$$

and it proves the first claim. Because q is strictly positive and continuous in both variables by its definition, $(M_n; n \in \mathbf{N}_0)$ is irreducible and any compact set with positive Lebesgue measure is a small set by Lemma 1.2 of Mengersen and Tweedie (1996). By the conditions,

$$\begin{aligned} P^*V^* - V^* \le cV^{*\gamma} + b1_{C^*} &\Rightarrow P(V \circ F) - V \circ F \le c(V \circ F)^\gamma + b1_{C^*} \\ &\Rightarrow PV - V \le cV^\gamma + b1_{C^*}(F^{-1}(x)) \ (x \in \mathbf{R}^d). \end{aligned}$$

Since C^* is a compact set, there is $r > 0$ such that $C^* \subset \{|x| \le r\}$, then $\{F^{-1}(x) \in C^*\} \subset \{|F^{-1}(x)| \le r\}$ and if we take $C = \overline{\{|F^{-1}(x)| \le r\}}$, then C is a compact set since $|F^{-1}|$ is a norm-like function. We can take C large enough to have positive Lebesgue measure, hence C is a small set. Then for C, V, γ, b, c, the drift condition (6) holds. □

We take $f(x) = x^{2/(2-r)}$ $(x > 1)$ and set properly to satisfy above conditions when $x \le 1$. When $|x| > 1$, $\nabla^T F(x) = (I_d + r/(2-r)x \cdot x^T/|x|^2)|x|^{r/(2-r)}$ and $\det \nabla^T F(x) = (2/(2-r))|x|^{dr/(2-r)}$. When $(M_n^x; n \in \mathbf{N}_0)$ is from the Langevin algorithm we call this transform algorithm, the transformed Langevin algorithm.

For practical purpose, it is convenient to take $f(x) \equiv x (x \le 1)$ and it is enough to establish the following conclusion, though it does not a C^2 function. We restrict f to be a C^2 function in our proof since it simplifies our proof.

Theorem 6 *Let p be a C^2 function that satisfies*

$$\limsup_{|x| \to \infty} x^T \cdot \nabla \log p(x) \le -\eta, \tag{23}$$

$$\lim_{|x| \to \infty} |x|^{\frac{r}{2}} \cdot |\nabla \log p(x)| = 0, \tag{24}$$

$$\lim_{|x| \to \infty} |x|^r \cdot \|\nabla^T \nabla \log p(x)\| = 0. \tag{25}$$

Consider the Transformed Langevin algorithm by F when $0 \le r < 2$. Then the drift condition (6) holds for $2 < s < 2+(\eta-d)(2/(2-r))$, $V(x) = (|F^{-1}(x)|^2+1)^{s/2}$ and $\gamma = (s-2)/s$. In particular, the upper bound of the polynomial order of convergence in total variation norm is $(\eta - d)/(2 - r)$.

Proof If p^* satisfies the properties (16) and (20), by using Theorem 5 for $(M_n^x; n \in \mathbf{N}_0)$, the claim follows by Proposition 1. Through the proof, we assume $|x| > 1$. By the definition of p^*, $\nabla \log p^*(x) = (\nabla^T F(x))^T \cdot (\nabla \log p)(F(x)) + \nabla \log \det \nabla^T F(x)$. Because $x^T \cdot \nabla^T F(x) = (2/(2-r))F(x)^T$, $\nabla \log \det \nabla^T F(x) = (\mathrm{d}r/(2-r))x/|x|^2$ and $\|\nabla^T F(x)\| \cdot |x| \le (d^{1/2} + r/(2-r))|F(x)|$, we obtain

$$x^T \cdot \nabla \log p^*(x) = \frac{2}{2-r} F(x)^T \cdot \nabla \log p(F(x)) + \frac{\mathrm{d}r}{2-r},$$

$$|x| \cdot |\nabla \log p^*(x)| \le \left(d^{\frac{1}{2}} + \frac{r}{2-r} \right) |F(x)| \cdot |\nabla \log p(F(x))| + \frac{\mathrm{d}r}{2-r}.$$

Let $\eta^* = (1-r/2)^{-1}(\eta - rd/2)$, then the following properties hold since $|F(x)|^{r/2} = |F(x)|/|x|$:

$$\limsup_{|x|\to\infty} x^T \cdot \nabla \log p^*(x) \le -\eta^*, \quad \lim_{|x|\to\infty} |x|^{\frac{r}{2}} |\nabla \log p^*(x)| = 0. \tag{26}$$

Next, we show that $\|\nabla^T \nabla \log p^*(x)\|$ goes to 0 in the limit. We take some steps to calculate it. In the following calculations, we sometimes drop the operator "$\cdot$" and the state x to simplify the inequalities and equations. First, divide $\|\nabla^T \nabla \log p^*(x)\|$ into two parts, $\|\nabla^T (\nabla^T F(x))^T \nabla \log p(F(x))\|$ and $\|\nabla^T \nabla \log \det \nabla^T F(x)\|$. About the second term, it is easy to see

$$|x|^2 \|\nabla^T \nabla \log \det \nabla^T F(x)\| \le (\mathrm{d}r(2-r))(d^{1/2} + 2).$$

Now consider the first term. We have

$$\begin{aligned} \nabla^T ((\nabla^T F(x)^T) \cdot \nabla \log p(F(x))) = {} & \nabla^T (|x|^{\frac{r}{2-r}} \nabla \log p(F(x))) \\ & + \frac{r}{2-r} \nabla^T (|x|^{\frac{r}{2-r}-2} x \cdot x^T \cdot \nabla \log p(F(x))). \end{aligned} \tag{27}$$

Then the first term in the above is $\nabla \log p(F) \nabla^T |x|^{r/(2-r)} + |x|^{r/(2-r)} \nabla^T (\nabla \log p(F))$. The norm of the first term of it is smaller than $(r/(2-r))|F(x)| \cdot |x|^{-2} \cdot |\nabla \log p(F(x))|$ and the second term is

$$\begin{aligned} \| |x|^{\frac{r}{2-r}} \nabla^T (\nabla \log p(F(x))) \| &\le |x|^{\frac{r}{2-r}} \|\nabla^T \nabla \log p(F(x)) \nabla^T F(x)\| \\ &\le \left(d^{\frac{1}{2}} + \frac{r}{2-r} \right) |F(x)|^{\frac{r}{2}} |\nabla^T \nabla \log p(F(x))|, \end{aligned}$$

hence both of them converge to 0. Finally, we show that the norm of the second term of (27) goes to 0 in the limit. We write

$$A := \nabla^T(|x|^{\frac{r}{2-r}-2}xx^T\nabla\log p(F(x))) = xx^T\nabla\log p(F(x))\nabla^T|x|^{\frac{r}{2-r}-2} + |x|^{\frac{r}{2-r}-2}\nabla^T(xx^T\nabla\log p(F(x))).$$

Since

$$\nabla^T(xx^T\nabla\log p(F(x))) = xx^T(\nabla^T(\nabla\log p(F(x))) + x^T\nabla\log p(F(x))I_d + \nabla\log p(F(x))^T x,$$

we obtain

$$\begin{aligned}|x|^2\|A\| &\le |x|^4|\nabla\log p(F)|\cdot|\nabla^T|x|^{\frac{r}{2-r}-2}| \\ &\quad + |x|^{\frac{r}{2-r}-2}(|x|^4\|\nabla^T\nabla\log p(F))\|\cdot\|\nabla^T F\| + |x|^3|\nabla\log p(F)|(d^{1/2}+1)) \\ &\le \left(\frac{r}{2-r}+d^{\frac{1}{2}}-1\right)|F|\cdot|\nabla\log p(F)| + \left(d^{\frac{1}{2}}+\frac{r}{2-r}\right)|F|^2\|\nabla^T\nabla\log p(F)\|.\end{aligned}$$

Then by (24) and (25), $\|A\|$ converges to 0. □

As we showed in Example 1, the Langevin algorithm with m degree of freedom Student's t proposal distribution has a central limit theorem for $L^{2+\epsilon}$ with $\epsilon > 4/(m-2)$.
On the other hand, transformed chain has a central limit theorem for $L^{2+\epsilon}$ with $\epsilon > 2(2-r)/(m-(2-r))$.

Jarner and Roberts (2002a) proved the same kind of improvements of the rate of convergence in another way. We transformed the chain to gain the heaviness of the tail. On the other hand, they weighted q^* of the transition kernel $Q(x, \mathrm{d}y) = q^*(|x-y|)\mathrm{d}y$. They took q^* as a probability distribution function of Student's t distributions instead of normal distributions. However they supposed stronger conditions, which is described in (5).

We can transform the random-walks based Metropolis–Hastings algorithm instead of the Langevin algorithm as Theorem 6. However, we cannot prove the improvements like this theorem without some extra conditions, for example, $A(x) = \{p(y) \ge p(x)\}$ should be a convex set. I cannot make out whether these difficulties are avoidable or are essential problems for the schemes.

4 Calculation

We now check the performance of the Metropolis–Hastings algorithms. In practice, we should choose good parameters. As stated in the previous section, we use f as $f(x) \equiv x$ $(x \le 1)$.

Table 1 Example 3, Random-walk with Gaussian increment distribution based algorithm

	$h = 30$	$h = 40$	$h = 50$
$N = 500$	111.70	95.69	90.83
$N = 1{,}000$	100.23	101.83	105.98
$N = 2{,}500$	159.91	267.78	127.22

Table 2 Example 3, Langevin algorithm

	$h = 30$	$h = 40$	$h = 50$
$N = 500$	113.62	121.6	123.18
$N = 1{,}000$	157.27	120.72	137.58
$N = 2{,}500$	165.95	143.11	172.00

Table 3 Example 3, Random-walk with Student's t increment distribution (degree of freedom is 1) based algorithm

	$h = 8$	$h = 10$	$h = 12$
$N = 500$	138.55	134.16	129.60
$N = 1{,}000$	139.53	143.13	144.25
$N = 2{,}500$	147.98.	144.21	145.02

Example 3 (*Multivariate t distribution*) Consider the multivariate t distribution (22) with the degree of freedom $m = 3$, mean $\mu = (2, 2)^T$ and

$$\Sigma = \begin{pmatrix} 2 & 1 \\ 1 & 3 \end{pmatrix}.$$

Start point $X_0 = (2, 3)$. We produced $M = 100,000$ parallel Markov chains $(X_n^m = (X_{1,n}^m, X_{2,n}^m)^T; n \in \mathbf{N}_0)$ $(m = 1, \ldots, M)$ by four algorithms below for each and calculated mean squared error

$$\mathrm{MSE}_{N,M} = \sum_{m=1}^{M} \frac{\left(\sum_{n=1}^{N} g(X_n^m) - \Pi(g)\right)^2}{N * M}. \tag{28}$$

We took $N = 500, 1{,}000, 2{,}500$ for each and $g(x, y) = x$. We consider the following algorithms:

- Random-walk with Gaussian increment distribution based algorithm (Table 1).
- Langevin algorithm (Table 2).
- Random-walk with Student's t increment distribution (degree of freedom is 1) based algorithm (Table 3).
- Transformed Langevin algorithm ($r = 1$) (Table 4).
- Transformed Langevin algorithm ($r = 1.2$) (Table 5).

These algorithms have central limit theorems for $L^{2+\epsilon}$ by Theorem 3, where the value of ϵ differs as follows: $\epsilon > 4$ for the first and second algorithms, $\epsilon > 1$ for

Table 4 Example 3, Transformed Langevin algorithm by $\gamma = 1$

	$h = 1$	$h = 2$	$h = 3$	$h = 10$
$N = 500$	45.65	33.59	45.93	457.5
$N = 1{,}000$	48.24	35.48	48.67	903.19
$N = 2{,}500$	49.14	40.15	45.83	2149.07

Table 5 Example 3, Transformed Langevin algorithm by $\gamma = 1.2$

	$h = 1$	$h = 1.2$	$h = 1.4$
$N = 500$	41.71	41.02	43.61
$N = 1{,}000$	42.59	41.87	43.28
$N = 2{,}500$	42.91	44.08	42.54

Table 6 Example 4, Random-walk with Gaussian increment distribution based algorithm

	$h = 0.5$	$h = 1$	$h = 1.5$
$N = 500$	2.121	2.024	2.246
$N = 1{,}000$	2.109	2.038	2.256
$N = 2{,}500$	2.124	2.025	2.258

the third and fourth, and $\epsilon > 8/11$ for the last one. Therefore in this case, Markov chain produced by the last algorithm have a central limit theorem, but we cannot say anything about the others using Theorem 3.

In Tables 1, 2, 3, 4 and 5, transformed algorithm $\gamma = 1.2$ works well in this case. However you should choose good parameters to obtain such an improvement. When $\gamma = 1.2$, the algorithm behaves badly for $h = 10$.

Example 4 The following example is anti-convex probability distribution:

$$p(x, y) \propto \frac{1}{(x^4 + y^2 + 1)^3}. \tag{29}$$

In this example, $\eta = 6$.

We consider the following algorithms:

- Random-walk with Gaussian increment distribution based algorithm (Table 6).
- Langevin algorithm (Table 7).
- Random-walk with Student's t increment distribution (degree of freedom is 1) based algorithm (Table 8).
- Transformed Langevin algorithm ($r = 1$) (Table 9).

Some of these algorithms have central limit theorems for $L^{2+\epsilon}$ where the value of ϵ differs as follows: $\epsilon > 8/5$ for the second algorithm, $\epsilon > 4/7$ for the last one. Since this probability distribution is not symmetric, we do not know whether other algorithms have a central limit theorem.

Table 7 Example 4, Langevin algorithm

	$h = 0.25$	$h = 0.50$	$h = 0.75$
$N = 500$	1.064	0.569	0.688
$N = 1{,}000$	1.070	0.566	0.689
$N = 2{,}500$	1.074	0.570	0.688

Table 8 Example 4, Random-walk with Student's t increment distribution (degree of freedom is 1) based algorithm

	$h = 0.05$	$h = 0.1$	$h = 0.2$
$N = 500$	4.906	4.782	4.817
$N = 1{,}000$	4.943	4.821	4.875
$N = 2{,}500$	5.018	4.851	4.931

Table 9 Example 4, Transformed Langevin algorithm by $\gamma = 1$

	$h = 0.06$	$h = 0.08$	$h = 0.10$
$N = 500$	1.309	1.222	1.324
$N = 1{,}000$	1.312	1.222	1.325
$N = 2{,}500$	1.328	1.226	1.328

In Tables 6, 7, 8 and 9, we used the same starting point X_0 and the same number of parallel Markov chains M as the previous example. The first algorithm is not so bad and the second algorithm is better than the last one. Transformation does not always show improvements.

5 Conclusion

The purpose of this paper is to introduce the Metropolis–Hastings algorithms that can deal with a wide class of heavy-tailed target distributions. We proved the convergence rate and sufficient conditions for convergence for these algorithms. The transformed algorithm is of the same rate of convergence as the heavy-tailed proposal random-walk algorithm, though the latter algorithm needs strong assumptions for the target.

Next, we want to prove the differences between the random-walk with Gaussian increment distribution based algorithm and the Langevin algorithm. Numerical calculation suggests that the asymptotic variance of the estimator $\widehat{\Pi(f)}_N = N^{-1}\sum_{n=1}^{N} f(M_n^x)$ of the Langevin algorithm is smaller than that of the random-walk based algorithm when the target distribution is not symmetric. Therefore, symmetricity seems to be an important condition for the latter algorithm. However, we could not prove it in theoretically.

Axknowledgments I am very grateful to Professor Nakahiro Yoshida for his valuable suggestions and fruitful discussions, and the referee and the associate editor for their valuable comments.

References

Douc, R., Fort, G., Moulines, E. (2004). Practical drift conditions for subgeometric rates of convergence. *The Annals of Applied Probability*, *14*, 1353–1377.

Fort, G., Moulines, E. (2000). V-subgeometric ergodicity for a Hastings–Metropolis algorithm. *Statistics Probability Letters*, *49*, 401–410.

Fort, G., Roberts, G. O. (2005). Subgeometric ergodicity of strong Markov processes. *The Annals of Applied Probability*, *15*, 1565–1589.

Grenander, U., Miller, M. I. (1994). Representations of knowledge in complex systems. *Journal of the Royal Statistical Society, Series B (Statistical Methodology)*, *56*, 549–603.

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, *57*, 97–109.

Jarner, S. F., Roberts, G. O. (2002a). Convergence of heavy tailed MCMC algorithms. http://www.statslab.cam.ac.uk/~mcmc/.

Jarner, S. F., Roberts, G. O. (2002b). Polynomial convergence rates of Markov chains. *The Annals of Applied Probability*, *12*, 224–247.

Jarner, S. F., Roberts, G. O. (2007). Convergence of Heavy-tailed Monte Carlo Markov Chain Algorithms. *Scandinavian Journal of Statistics*, *34*, 781–815.

Jones, G. L. (2004). On the Markov chain central limit theorem. *Probability Surveys*, *1*, 299–320.

Mengersen, K. L., Tweedie, R. L. (1996). Rates of convergence of the Hastings and Metropolis algorithms. *The Annals of Statistics*, *24*, 101–121.

Metropolis, N., Rosenbluth, W. A., Rosenbluth, M. N., Teller, A. H., Teller, E. (1953). Equations of state calculations by fast computing machines. *Journal of Chemical Physics*, *21*, 1087–1092.

Meyn, S. P., Tweedie, R. L. (1993). *Markov chains and stochastic stability*. London: Springer.

Nummelin, E. (1984). *General irreducible Markov chains and nonnegative operators*. Cambridge: Cambridge University Press.

Roberts, G. O., Stramer, O. (2002). Langevin diffusions and Metropolis-Hastings algorithms. International workshop in applied probability. *Methodology and Computing in Applied Probability*, *4*, 337–357.

Roberts, G. O., Tweedie, R. L. (1996a). Geometric convergence and central limit theorems for multidimensional Hastings and Metropolis algorithms. *Biometrika*, *83*, 95–110.

Roberts, G. O., Tweedie, R. L. (1996b). Exponential convergence of Langevin diffusions and their discrete approximations. *Bernoulli*, *2*, 341–363.

Stramer, O., Tweedie, R. L. (1999a). Langevin-type models. I. Diffusions with given stationary distributions and their discretizations. *Methodology and Computing in Applied Probability*, *1*, 283–306.

Stramer, O., Tweedie, R. L. (1999b). Langevin-type models. II. Self-targeting candidates for MCMC algorithms. *Methodology and Computing in Applied Probability*, *1*, 307–328.

Tuominen, P., Tweedie, R. L. (1994). Subgeometric rates of convergence of f-ergodic Markov chains. *Advances in Applied Probability*, *26*, 775–798.